Dichotic Listening Test Performance In Children

Kairn Stetler Kelley
University of Vermont

Follow this and additional works at: http://scholarworks.uvm.edu/graddis
Part of the Speech Pathology and Audiology Commons

Recommended Citation
http://scholarworks.uvm.edu/graddis/707

This Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks @ UVM. It has been accepted for inclusion in Graduate College Dissertations and Theses by an authorized administrator of ScholarWorks @ UVM. For more information, please contact donna.omalley@uvm.edu.
DICHTOTIC LISTENING TEST PERFORMANCE IN CHILDREN

A Dissertation Presented

by

Kairn Stetler Kelley

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Clinical and Translational Science

May, 2017

Defense Date: January 27, 2017
Dissertation Examination Committee:

Benjamin Littenberg, M.D., Advisor
Donna Rizzo, Ph.D., Chairperson
Rodger Kessler, Ph.D.
Constance van Eeghen, Dr.P.H.
Cynthia J. Forehand, Ph.D., Dean of the Graduate College
ABSTRACT

Dichotic tests evaluate binaural integration through simultaneous presentation of different stimuli to each ear of a listener who has normal hearing sensitivity in both ears. Dichotic listening deficits may lead to problems with language, communication, reading, or academic performance. If accurately identified, dichotic deficits may be treatable with listening training or managed with accommodation. However, it is not clear which of several commercially-available dichotic test recordings are best for audiologists to use when assessing binaural integration in children.

Literature review revealed limited evidence of reliability, accuracy, usefulness, or value for dichotic tests applied to children. Of 11 dichotic tests identified, five reported some evidence of test-retest reliability. Correlation between results on repeated administration was moderate to good (r=0.59 to 0.92). Evidence of accuracy was identified for 5 tests but was not generalizable due to significant limitations in study design. No evidence was found to either support or dispute claims of usefulness or value.

Since reliability is a necessary prerequisite for good test performance, we sought to directly compare test-retest reliability for three dichotic measures: SCAN-3 Competing Words (CW), Musiek’s Double Dichotic Digits (DD-M), and Bergen Dichotic Listening Test with Consonant-Vowel Syllables (CV-B). Sixty English-speaking children, 7-14 years old with normal hearing, had a single study-visit during which each test was administered twice. Changes on retest were compared to binomial model predictions, summarized by within-subject standard deviation (Sw), and compared among tests. Correlates of variance were explored. All 3 tests had reliability within bounds predicted by binomial model. Forty-item scores were more reliable (Sw=5%) than those based on 20-30 items (Sw=6-8%). No associations between participant characteristics and reliability were found.

CW and DD-M were evaluated for evidence of agreement and decision consistency. Although participants were rank ordered similarly by right ear (ρ = 0.58), left ear (ρ = 0.51) and total (ρ = 0.73) scores, the tests did not agree on ranking by interaural asymmetry (ρ =0.18). CW and DD-M did not agree on direction of ear advantage (κ= 0.01, ρ = 0.93) and had poor agreement on which children displayed dichotic deficits (κ = 0.22, p < 0.01). DD identified significantly more participants with deficits (n=18) than CW (n=3) (p < 0.001).

Although dichotic procedures show moderate reliability, their precision is limited. Assessment of their accuracy is limited by the absence of a widely-accepted gold standard reference test, but two commonly used tests failed to agree on which children had deficits. The data do not yet support routine clinical use of dichotic tests of binaural integration with children. Additional research is needed to determine if there are any conditions under which dichotic procedures demonstrate usefulness or value.
ACKNOWLEDGEMENTS

This work would not have been possible without the mentorship of Dr. Benjamin Littenberg. I am deeply grateful for his generosity in sharing his knowledge, his time, and his love of science.

I am also indebted to the families that volunteered their time to my research.

Words aren’t enough to express my love and gratitude for Sean, Erin, and Ben Kelley. There would be no dissertation if it weren’t for Sean’s help.

CTS faculty, staff, and fellows have taught and encouraged me for years. Thank you, Connie van Eeghen, Rodger Kessler, Alan Rubin, Amanda Kennedy, Charles MacLean, Richard Pinckney, Peter Callas, Debi Henry, Sylvie Frisbie, Marianne Burke, Nancy Gell, Adam Sprouse-Blum, and Abby Crocker.

Becky Webber kept me company and told me it would be OK.

Thank you, Mollie Schmidt.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Acknowledgements</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>vii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>viii</td>
</tr>
</tbody>
</table>

**CHAPTER 1: DICHOTIC TESTS OF BINAURAL INTEGRATION IN CHILDREN: EVIDENCE OF CLINICAL PERFORMANCE** ........................................ 1

1.1. Abstract ............................................................................................................. 1

1.2. Introduction ......................................................................................................... 2

1.3. Materials and Methods ....................................................................................... 4

1.4. Identify Tests of Interest .................................................................................... 5

1.5. Identify Evidence for Evaluation ....................................................................... 6

1.6. Evaluate Literature for Evidence of Test Performance .................................... 7

1.6.1. Reliability ........................................................................................................ 7

1.6.2. Accuracy .......................................................................................................... 7

1.6.3. Usefulness ....................................................................................................... 8

1.6.4. Value ............................................................................................................... 8

1.7. Results ................................................................................................................. 8

1.7.1. Evidence of Test-Retest Reliability ................................................................. 10

1.7.2. Evidence of Accuracy ..................................................................................... 12
CHAPTER 1: EVALUATION OF THE DICHOTIC LISTENING TEST

1.8. Evidence of Usefulness ........................................................................................................ 16

1.9. Evidence of Value .................................................................................................................. 18

1.10. Discussion ............................................................................................................................ 19

1.10.1. Reliability Discussion .................................................................................................... 19

1.10.2. Accuracy Discussion ..................................................................................................... 20

1.10.3. Usefulness Discussion ................................................................................................... 21

1.10.4. Value Discussion ........................................................................................................... 21

1.10.5. Limitations ..................................................................................................................... 21

1.10.6. Next Steps ..................................................................................................................... 22

1.11. Tables .................................................................................................................................. 23

1.12. References .......................................................................................................................... 26

CHAPTER 2: COMPARISON OF DICHOTIC LISTENING TEST-RETEST
RELIABILITY IN CHILDREN ........................................................................................................ 32

2.1. Abstract ................................................................................................................................. 32

2.2. Introduction ........................................................................................................................ 33

2.2.1. Characteristics of Selected Dichotic Tests .................................................................... 35

2.3. Methods ................................................................................................................................ 37

2.3.1. Subjects ............................................................................................................................ 38

2.3.2. Setting ............................................................................................................................ 38

2.3.3. Measures ........................................................................................................................ 39
CHAPTER 3: COMPARISON OF CHILDREN’S DOUBLE DICHOTIC DIGITS AND SCAN-3 COMPETING WORDS FREE-RECALL SCORES .....................55

3.1. Abstract ........................................................................................................55

3.2. Introduction ..................................................................................................56

3.3. Methods .........................................................................................................58

  3.3.1. Subjects ....................................................................................................58

  3.3.2. Test Environment and Equipment ..........................................................59

  3.3.3. SCAN-3 Competing Words (CW) .............................................................59

  3.3.4. Double Dichotic Digits Test (Digits) .........................................................60

  3.3.5. Hearing Sensitivity ....................................................................................60
### LIST OF TABLES

Table 1-1 Dichotic tests for evaluating binaural integration of native English-speaking children

Table 1-2 Evidence of Test-Retest Reliability for Dichotic Speech Tests in Children

Table 1-3 Evidence of Accuracy for Dichotic Speech Tests

Table 2-1 Characteristics of 60 Participants

Table 2-2 Summary of First and Second Scores (% Correct) on Dichotic Digits, Dichotic CV Syllables, and Competing Words tests (n=60)

Table 2-3 Dichotic Test Scores Change on Retest for 60 participants

Table 2-4 Characteristics of 60 Participants

Table 2-5 Mean, (SD), and [Range] of 60 Participants’ Competing Words (CW) and Dichotic Digits (DD) Scores

Table 2-6 Right Ear Advantage (REA) Classification by Competing Words and Dichotic Digits tests

Table 3-4 Participants Normal/Abnormal Classification of Dichotic Performance by Competing Words and Dichotic Digits Tests
LIST OF FIGURES

Figure 2-1 Difference between first and second scores of 60 participants with binomial 95% CI for differences .................................................................51

Figure 3-1 Association between right and left ear scores on Competing Words and Dichotic Digits tests (n=60) ........................................................................................................68

Figure 3-2 Associations between Competing Words and Dichotic Digits total scores and Interaural asymmetry .........................................................................................69
CHAPTER 1: DICHOTIC TESTS OF BINAURAL INTEGRATION IN CHILDREN: EVIDENCE OF CLINICAL PERFORMANCE

1.1. Abstract

Purpose: To evaluate dichotic speech tests of binaural integration used to assess auditory processing in English speaking children age 6-14.

Method: Dichotic speech test recordings and pertinent research studies were identified from iterative searches of the internet and bibliographic databases, as well as communication with colleagues and test publishers. Test documentation and peer-reviewed literature were evaluated for evidence of reliability, accuracy, usefulness, and value.

Results: Eleven dichotic tests of binaural integration were identified. Evidence of test-retest reliability was found for 5 tests and demonstrated moderate to good correlation between results on repeated administration (r=0.59 to 0.92). Evidence of accuracy was identified for 5 tests, but was either inconsistent with accurate performance or was not generalizable due to significant limitations in study design. No evidence was found to either support or dispute claims of usefulness or value.

Conclusions: A medical diagnostic framework is useful for evaluating dichotic tests. Although dichotic procedures show moderate reliability, the absence of a widely-accepted gold standard reference test limits our ability to assess their value. Overall, the data do not support the routine use of dichotic tests of binaural integration for clinical evaluation of children.
1.2. Introduction

Dichotic tasks are often part of auditory processing assessment batteries (Chermak, Silva, Nye, Hasbrouck, & Musiek, 2007; Emanuel, 2002; Emanuel, Ficca, & Korczak, 2011). Dichotic speech tests present recorded speech stimuli through stereo headphones so that different signals reach the listener’s right and left ears simultaneously. Binaural integration is assessed when listeners are instructed to attend to both ears (Bellis, 2002). A number of dichotic speech recordings are available to assess binaural integration.

Dichotic procedures are recommended for assessment of possible auditory processing disorder (APD) (American Academy of Audiology, 2010; American Speech-Language-Hearing Association, 2004, 2005a) because of their clinical feasibility; the availability of a variety of dichotic tests using speech materials including digits, words, and sentences; and the “long and proven record of sensitivity” of dichotic procedures to detect lesions of the central auditory nervous system (Musiek, 1983b). Physiologic models of the central auditory nervous system are importantly informed by studies of how individuals with specific brain lesions have performed on dichotic listening tasks (Broadbent, 1955; Kimura, 1961; Musiek, 1983b; Satz, Strauss, Wada, & Orsini, 1988; Strauss, Gaddes, & Wada, 1987). Children with a range of academic and language deficits have lower performance on these tests when compared with their typically developing peers (de Wit et al., 2016; Vermiglio, 2016). Dichotic scores also serve as important determinants of diagnostic sub-category for both the Bellis-Ferre and Buffalo models of auditory processing disorders (Arnst & Katz, 1982; Bellis, 2002).
Audiologists are obligated to select for each patient an individualized test battery of appropriately normed “tests with good reliability and validity that also demonstrate high sensitivity, specificity, and efficiency” (American Academy of Audiology, 2010; American Speech-Language-Hearing Association, 2004, 2005a), but evaluating tests is a complex task. Friberg and McNamera (2010) discussed challenges audiologists face when considering auditory processing tests including: selecting which tests to evaluate, prioritizing evaluation criteria, finding evidence of clinical performance, and managing the absence of gold-standard for evaluating accuracy. Although interest in formal evaluation of diagnostic procedures in communication sciences including audiology has been growing (Robey, 2004; Vermiglio, 2014), the current literature contains a paucity of rigorous evaluations (Robey, 2014; Vermiglio, 2016). Diagnostic procedures in other fields of health care have usefully been evaluated using a general framework of diagnostic technology assessment (Littenberg, 1992). Criteria are continually being refined to improve the quality and ease of test evaluation (American Educational Research Association, 2014; Friberg, 2010; Gwet, 2012) and reporting of those evaluations (Bossuyt et al., 2015; Deeks, Wisniewski, & Davenport, 2013; Kottner et al., 2011).

For this review of dichotic test performance, we adapted a framework developed by Newman and Kohn (2009) for clinicians to assess the performance of diagnostic tests based on four increasingly stringent criteria: reliability, accuracy, usefulness, and value. **Reliability** is the ability of a test to provide the same result regardless of who administers it or when it is given. **Accuracy** is the ability of a test to distinguish between
those with and without disease. **Usefulness** is the ability of the test to inform decision making. **Value** is achieved if the benefits of using a test outweigh the costs of its use. A test must be reliable to be accurate, must be accurate to be useful, and must be useful to be valuable. This simple framework is consistent with ASHA recommendations for determining evidence-based practice (American Speech-Language-Hearing Association, 2005b) and is a natural progression from the priorities of our profession (Friberg, 2010; R J McCauley & Swisher, 1984; Rebecca J McCauley, 1996).

The purposes of this review are to: (1) identify the dichotic speech tests used to evaluate binaural integration in American English speaking children; (2) evaluate existing evidence of test performance in children to identify the strengths and weaknesses of each test with regard to reliability, accuracy, usefulness, and value; and (3) identify where further evidence is needed to validate the proposed uses of the dichotic test scores. The results are expected to be of use both to audiologists striving to provide evidence-based clinical care for children with auditory processing problems and for researchers studying auditory processing disorders.

### 1.3. Materials and Methods

This review consisted of identifying currently available dichotic speech test recordings appropriate for assessing binaural integration of English-speaking children, identifying research describing the performance of the tests in samples of children, and evaluating each research report for evidence of reliability, accuracy, usefulness, and value.
1.4. Identify Tests of Interest

We sought to identify dichotic tests meeting the following criteria:

- Available to any audiologist for clinical use by purchase or request
- Stimuli are recorded speech (e.g., syllables or words, including digits)
- Stimuli are dichotically aligned for simultaneous onset to right and left ears
- Stimuli are appropriate for use with native English-speakers
- Listeners are instructed to repeat all or part of what was heard without being told to direct attention to one ear over the other (i.e., assessing binaural integration)
- Interpretation is guided by published norms or cut scores

These criteria exclude “out of print” recordings (e.g., Dichotic Rhyme (Musiek et al., 1989)), experimental recordings not endorsed for clinical auditory evaluation (e.g., Fused Rhymed Words Test (Wexler & Halwes, 1983)), tests using non-English language words (e.g., Spanish versions of dichotic tests), and tests for which the listener is instructed to direct attention to one or the other ear at the exclusion of the other (e.g., SCAN-3 Competing Words Directed Ear (Keith, 2009b, 2009a)). Recordings for which we were unable to identify normative data to inform interpretation of children’s scores are also excluded (i.e., VA CD 1-pair Dichotic Digits; VA CD 3-pair Dichotic Digits, VA CD Dichotic CVs; (Noffsinger, Martinez, & Wilson, 1994; Noffsinger et al., 1994) AUDiTEC Dichotic Digits, Part 1 single pairs; Musiek Dichotic Digits, Version 1 single pairs).
Tests were identified through review of clinical practice guidelines’ reference lists (American Academy of Audiology, 2010; American Speech-Language-Hearing Association, 2005a), auditory processing disorders texts (Bellis, 2002; Chermak & Musiek, 2013; Musiek & Chermak, 2013), published literature including survey and review articles (Emanuel et al., 2011; Friberg, 2010), online resources (American Speech-Language-Hearing Association, n.d.; Auditec, 2014; Pearson Assessment Products, n.d.; Wilson, 1993), the author’s personal files and queries to professional colleagues.

1.5. Identify Evidence for Evaluation

We sought to identify research that reported results of the dichotic procedures of interest administered to children in any portion of the age range 6-14 years old. We reviewed the test documentation for each recording. Peer-reviewed literature was identified using the strategies described above for identifying recordings, as well searches in indexed bibliographic databases (i.e., Ovid Medline, Pubmed, PsycINFO, ERIC, LLBA, ComDisDome, CINAHL, ASHAwire). Searches were iterative, first using general terms (e.g., dichotic) and later exact terms (e.g., “staggered spondaic” and “SSW”) as specific dichotic tests were identified. Peer-reviewed sources were first screened for relevance. If they presented original research in which eligible dichotic tests were administered to groups of English speaking children with normal hearing and in any portion of the age range 6-14, the studies were then evaluated according to the framework described below.
1.6. Evaluate Literature for Evidence of Test Performance

We adapted the hierarchical framework proposed by Newman and Kohn (2009) to organize our assessment of the evidence for dichotic procedures.

1.6.1. Reliability

A source was considered to contain evidence of test-retest reliability if it described repeated administration of tests to the same subjects. The following were considered acceptable evidence to evaluate reliability: within-subject standard deviation, within-subject coefficient of variation, percent agreement, kappa statistic, correlation coefficient, or Bland-Altman plot.

Studies that reported only central tendency (e.g., mean score on first test and second test) were not included because that level of detail is not sufficient to draw conclusions about individual changes in scores. Studies that reported split-list performance were excluded since number of test items is a known contributor to reliability (Thornton & Raffin, 1978).

1.6.2. Accuracy

Any papers that described performance of any dichotic test of interest (the index test) in comparison to an available reference standard different from the index test were evaluated as evidence of accuracy. Studies were included in this section of the paper if they presented at least one of the following (or data sufficient to calculate) were presented: sensitivity/specificity, positive- and negative- predictive value, likelihood ratios, accuracy in the sample, or ROC curve.
Reports of performance on APD test batteries, even those including dichotic test(s), were excluded unless we could determine individual performance on the dichotic measure. Also excluded were studies that report only differences in central tendency between groups, because comparisons of group means are insufficient to draw conclusions about test accuracy. Studies that only partially reported classification of subjects (i.e., insufficient to complete a 2x2 table) were excluded. Finally, studies that used a reference standard that is no longer available (e.g., Willeford recordings) were excluded.

1.6.3. Usefulness

To be included in this section of the review, papers must have described a relationship between scores on the dichotic tests under study and either clinical decisions or patient outcomes. Papers that described both test performance and a clinical outcome (e.g., response to therapy) were evaluated for evidence of usefulness.

1.6.4. Value

The value of a diagnostic test is evaluated by comparing the costs of using it against the benefits of its use. To be included in this section of the review, papers must specifically address costs and benefits.

1.7. Results

Eleven different recordings of dichotically aligned digits, syllables, or words were identified as available for evaluating binaural integration of native English-speaking children (see Table 1-1).
Five recordings using digits for stimuli were identified:

- Random Dichotic Digits Test (RDDT)
- VA 2-pair Dichotic Digits (DD-VA)
- AUDiTEC Dichotic Digits, Part 2 double pairs (DD-A)
- Musiek Dichotic Digits, Version 2 double pairs (DD-M)
- Dichotic Digits subtest of Differential Screening Test for Processing (DD-DSTP)

Two recordings using dichotically aligned consonant-vowel syllables were identified:

- Dichotic Consonant Vowel Test (CV-A)
- Bergen Dichotic Listening Tests with Consonant Vowel Syllables (CV-B)

Dichotically aligned English words were the stimuli for the remaining for recordings that met inclusion criteria:

- Staggered Spondaic Word Test – List EC (SSW)
- Competing Words Free-Recall Subtest of SCAN-3:C and SCAN-3:A (CW)
- Dichotic Words Test (DWT)
- Dichotic Word Listening Test (DWLT)

All stimuli are recordings of a male voice. All of the recordings are available on compact disc; DWT and RDDT are also available as licensed software. Interpreted scores vary among the tests; tests can more than one interpreted score. DD-A and DD-DSTP are interpreted based on the total number of correct responses without regard to ear. Separate right- and left-ear cuts scores exist for SSW, DD-M, DWLT, CV-B, DD-VA, CV-A, and
RDDT. RDDT is further divided with separate cut scores for trials with 1, 2, and 3 digits presented in sequence to each ear. Interaural asymmetry is evaluated by simple right-minus-left calculation (CW, DD-M, and RDDT) or more complex indices (CV-B, DWLT). Only DWT is interpreted based on dominant and non-dominant ear performance rather than right and left ear.

1.7.1. Evidence of Test-Retest Reliability

Test-retest reliability is listed in Table 1-2 with the source of evidence and number of children contributing data. Reliability evidence for DD-M was found in (Musiek, Gollegly, Kibbe, & Verkest-Lenz, 1991); evidence for DD-DSTP, CV-B, SSW, and CW was found in the test documentation provided by the publisher. All reports of reliability were presented as correlation. No data about test-retest reliability in children was identified for RDDT, DD-VA, DD-A, CV-A, DWT, or DWLT.

Musiek et al (1991) described test performance of DD-M in participants 13-73 years old. Evidence of test-retest reliability for percent correct in each ear was reported as r=0.77 for retest of 4 patients (8 ears) with stable CANS lesions. It is not clear if any of these patients were children.

DD-DSTP test retest reliability was described for 78 students in the manual (Richard & Ferre, 2006). The authors did not report how the children were selected for repeat evaluation or specify the interval between first and second test administration. Reliability coefficient (r) ranged 0.39 to 0.96 depending on the age group, with 7-14 students of each age contributing data. Reliability coefficient for the whole sample was reported as 0.74.
CV-B manual reported reliability of “between .70-.80” Although the authors have published details of test-retest reliability in adults (K Hugdahl & Hammar, 1997), the source of the test-retest reliability data for children is not described. The manual states “Children are found to display a less consistent REA [right ear advantage] compared to adults, something which shows itself in terms of both a smaller proportion of individuals with REA, and lower test-retest reliability.”

Reliability of the SSW (full EC list) was reported in the manual for two groups of subjects: 15 children age 7-11 retested 2-4 months after their first test and an unspecified number of children age 6-14 diagnosed with ADHD and CAPD. For both samples, correlation of first and second C-SSW score (total number of errors corrected for word recognition in quiet) was strong (r=0.92, p ≤ 0.01). Correlation of reversals (responses in which all four targets are repeated in unexpected order) was reported for the sample of 15 children and was not significant (r=0.49). The mean C-SSW scores (possible range 0-40) in one sample improved by 3.2 and in the other sample improved by 4.6. Since variance was not described, we cannot determine if this is a significant change.

Reliability for CW was reported in the test manuals (Keith, 2009b, 2009a) for 48 children age 5-11 and an unspecified number of adolescents tested in a mixed group of 58 adolescents and adults. All were retested 1-29 days following their first test as part of the normative samples. Gender, ethnicity, and educational level of the primary caregiver were described for the sample of participants used to study reliability. Pearson’s product-moment correlation coefficient between first and second scores in the sample of children was r= 0.59. The test-retest correlation for adolescents and adults was r=0.69. The mean
scaled-score (possible range 0-19) increased on retest by 1.86 among children and 1.75 among adults, a small but significant improvement on retest in both groups.

1.7.2. Evidence of Accuracy

Six sources were identified that presented data linking individuals’ performance on one of the dichotic tests of interest relative to a reference standard. Four were traditional accuracy studies evaluating performance of SSW, DD-M, and DWLT against a pre-determined reference standard (See Table 1-3); one described accuracy of SSW against a variety of reference criteria found in retrospective chart review (Gustafson & Keith, 2005); and one was a report of prevalence (Moncrieff, Keith, Abramson, & Swann, 2016) included because it reported individual performance on RDDT and DWT sufficient to calculate a 2x2 table. None of the sources prospectively evaluated performance of individual dichotic procedures against an independent, replicable reference standard that was applied to all subjects in the study sample by blinded reviewers. No data about accuracy was found for DD-VA, DD-A, DD-DSTP, CV-A, CV-B, or CW. Although diagnostic performance of SCAN-3 and DSTP test batteries were discussed in the respective manuals, performance of the dichotic subtests was not described independently.

The SSW manual (Katz, 1998) described performance of SSW Traditional Analysis for classifying 311 subjects aged 5-23. In Traditional Analysis, failure to meet threshold performance for any one of four different scores (right- and left-ear scores in the competing- and non-competing conditions) results in a classification as “outside normal limits on SSW.” Of 171 subjects referred for auditory processing evaluation, 145
fell outside normal limits on SSW (sensitivity = 0.85). Scores for the 119 of the 140 control subjects were within normal limits (specificity = 0.85). The reference standard in this study was whether subjects presented to clinics for auditory processing evaluation or were recruited as typically developing controls. This reference standard is not replicable because problems that motivated evaluation were not described. The separate recruitment of cases and controls is a known cause of spectrum bias which causes artificially high estimates of sensitivity and specificity compared to clinical a sample of children with comorbid

Berrick et al (1984) evaluated the performance of SSW left- and right-competing scores for discriminating between children with classroom learning difficulties (n= 97) and children without any academic concerns (n=93). Cut scores at 2 SD below the mean of the study sample resulted in correct classification of 74% of the 190 subjects. Berrick et al demonstrated specificity of 0.95 (defined at the time of setting cut scores) and sensitivity of 0.56. The authors recommended continued evaluation of norms to determine the best cut scores for clinical use. The study suffered from use of a reference standard (classroom learning difficulties) not specified in a way that can be replicated and not specific to dichotic deficit. The generalizability was further limited by use of control subjects with average- to above-average academic performance (spectrum bias) and use of non-standard cut-scores that do not generalize to clinical practice.

SSW and DD-M were among the auditory tests evaluated by Singer, Hurley, and Preece (1998) for their ability to discriminate between 7-13 year old children with classroom learning disability (n=147) and children with no known learning dysfunction
(n=91). They did not evaluate the cut scores recommended by the test authors but rather summarized each child’s dichotic performance as total percent correct in both ears combined for all dichotic stimuli and set cut scores to achieve a 10% false positive rate in the sample. With specificity for each test set at 0.90, the authors calculated sensitivity (hit rate). The SSW total correct score for competing conditions achieved sensitivity of 0.31. DD-M total percent correct score identified 54% of children with classroom learning difficulty (sensitivity = 0.54). Singer concluded that neither DD-M nor SSW would be included in an optimally efficient approach to diagnosing APD in children. This study suffered from a reference standard (classroom learning disability) that was not specified to allow replication and not specific to dichotic deficit. Comparison to control subjects with no known learning problems introduces spectrum bias. Analysis of non-standard scores limits the generalizability of the results.

Roberts et al (1994) administered the 30 item version of DWLT to 163 children: 142 normal controls and 21 with confirmed history of traumatic brain injury. Cut scores were set at 10th percentile of typically developing children with no reported history of brain injury (specificity 0.91). Nine of the 21 brain-injured children in the study had #B30 scores (number of trials for which both right and left targets were repeated from a list with 30 trials) worse than the 10th percentile cut scores (sensitivity= 0.43). The report is somewhat weakened by the authors’ failure to adequately describe the severity of brain injury of the cases studied (i.e., severity of injury, interval between injury and dichotic assessment). As with the other studies described, the analysis suffers from a reference
standard that is neither replicable nor specific to dichotic deficit and from a non-clinical comparison group (spectrum bias).

Gustafson and Keith (2005) reviewed the charts of 159 children aged 5-17 who had been tested with SSW during joint speech-language and audiology auditory processing evaluation. The authors detected no clear relationship between subjects’ Buffalo Model categories assigned based on SSW scores and diagnosis with any of 5 speech language disorders. However, the study was not optimally designed or powered to detect such a relationship. SSW scores were used to assign subjects to one of four Buffalo Model categories (i.e., decoding, tolerance-fading memory, integration, organization) or normal. In addition to SSW, clinicians administered and other auditory and speech-language measures selected from a set of 19 tests or subtests based on case history. The study was not adequately powered to address the 12 hypotheses explicitly stated for likely patterns of relationship between SSW classifications and other tests’ scores. Accuracy analysis focused on agreement between SSW categories and diagnosis with any of 5 speech language disorders (i.e., abnormal receptive- and expressive-language as measured by CELF-R or -3, abnormal articulation detected by the testing clinician reported, or abnormal oral reading or reading comprehension as measured by the Independent Reading Inventory). The study suffered from partial verification bias -- none of the speech-language measures were administered to the whole sample; the proportion of children with testing for problems with expressive language, receptive language, articulation, oral reading, or reading comprehension varied from 31 to 46 subjects (19-29% of the sample). In addition, there was no attempt to protect against review bias using
blinding. SSW results may have influenced which other tests were administered and the results of speech-language tests may have influenced whether SSW was administered.

Moncrieff et al (2016) sought to determine prevalence of three patterns of dichotic deficit (dysaudia, amblyaudia, and amblyaudia plus dysaudia) among 141 children referred for auditory processing evaluation. Children were diagnosed if they demonstrated the same pattern of scores on both RDDT and DWT. The authors reported 56% of could be diagnosed with a specific dichotic deficit based on agreement between RDDT and DWT. Although not designed as an accuracy study, the paper met criteria for inclusion in this review by presenting individual subjects’ performance on both tests. RDDT and DWT performance cannot be compared to ultimate diagnosis because of incorporation bias, but agreement of RDDT and DWT is still of interest. The two tests agreed on presence or absence of deficit 70% of the time, but our analysis suggests this level of agreement is indistinguishable from random chance ($p = 0.43$; $\kappa = 0.07$) given the high rate of disorder in the sample (approximately 80% of children using either test). However, among the 40 children with dysaudia (bilaterally abnormal scores) on both tests, RDDT and DWT agreed 78% of the time on whether the abnormality was symmetrical ($n=17$) or accompanied by larger than expected interaural asymmetry ($n=14$) ($p = 0.023$ for $\chi^2$ of 4x4 table).

1.8. Evidence of Usefulness

We identified four studies that presented individual performance on dichotic tests (DD-M and SSW) pre- and post-intervention. All were small and none contained a control group. None of the studies constitute evidence of usefulness of for DD-M or SSW
to identify appropriate recipients of intervention. No studies was found linking intervention with performance on RDDT, DD-VA, DD-A, DD-DSTP, CV-A, CV-B, CW, DWT, or DWLT. The studies we found are described briefly below.

A small study of personal amplification use among children diagnosed with APD by Kuk et al (2008) was not able to detect improvement on an Auditory Continuous Performance Test (ACPT) after 6 months of hearing aid use. Our analysis of the data presented showed no difference in response between children with normal DD-M scores (n=5) and abnormal DD-M scores (n=3). This small study neither excludes an effect nor provides any evidence to support the use of Dichotic Digits tests to inform which children with APD might benefit from personal amplification.

Dichotic listening training interventions among children diagnosed with dichotic listening deficits were the topic of the remaining identified studies (English, Martonik, & Moir, 2003; Moncrieff & Black, 2008; Stephenson, 2008). In all three studies, the researchers administered their own dichotic training regimen to a group of subject with abnormal dichotic test scores and measured the same dichotic test scores after intervention. Moncrieff and Black (2008) also evaluated oral listening comprehension and reading outcomes.

English, Martonik, and Moir (2003) used DD-M to identify a sample of children with left ear dichotic deficits. Ten of 11 showed improved left ear DD-M scores after 5-11 brief (8 minute) weekly sessions of listening to an audio-book using only their left ear. Moncrieff and Black (2008) used DD-M to identify 20 children with DD-M asymmetry (left-ear worse) for their two phase study. Phase I demonstrated that left ear DD-M scores
improved for most 7 of 8 participants after 11 dichotic training sessions (3 times weekly for four weeks). Phase II was designed to measure effects of more intense dichotic training on a wider range of outcomes. All children who participated in Phase 2 showed increased dichotic scores on post-test. All but three (9/12 children) also showed improvement on listening comprehension, and improvements in listening comprehension were significantly correlated with improvement in left ear dichotic digits scores ($r=0.686$, $p <0.05$). Although both these studies show improved outcomes following dichotic training administered to children with low DD-M scores, neither accounted for regression to the mean, learning, age, or effects of other therapies administered outside of the study intervention.

Stephenson (2008) used SSW and SCAN-C (Keith, 2000) to identify a group of 8 participants with dichotic deficits. After 8 bi-weekly, 45 minute sessions of author developed Dichotic Adaptive Training (DAT), Stephenson observed mean score for his 8 subjects improved for 6 of 19 dichotic scores, including overall and right ear conditions of SSW and DAT. This small study neither proves nor excludes an effect of DAT, nor provides any evidence to support the use of SSW to inform which children might benefit from DAT.

1.9. Evidence of Value

The only study identified that explored the costs and benefits of any dichotic procedures was (Singer et al., 1998). They determined that neither test had sufficient performance to justify being included in the value calculation.
1.10. Discussion

This review revealed significant challenges for audiologists attempting to select dichotic procedures with “good reliability and validity that also demonstrate high sensitivity, specificity, and efficiency” (American Speech-Language-Hearing Association, 2005a) for use with children.

1.10.1. Reliability Discussion

Literature about test-retest reliability of dichotic speech tests was limited. Test-retest reliability evidence was found for only 6 of the dichotic tests, and in all cases was summarized using correlation coefficients. Correlation is a poor measure of reliability since it is sensitive to outliers, is sensitive to the range of measurement, shows a strong relationship in the presence of systematic changes over time, and tests significance against a null-hypothesis of no relation which is not sensible for two measurements of the same individual (Newman & Kohn, 2009). Correlation does not provide clinicians with information about precision, the amount of individual change that can be expected on retest. Reports of test-retest reliability also suffered from small samples (n<4 to 78) that were poorly described (e.g., no discussion of selection criteria). Protocols (e.g., interval between tests, scores evaluated) were poorly specified if at all. None of the reports discussed the range of change observed on retest. Mean change observed on retest was only reported for CW and SSW. Some sources incorrectly reported the standard error of measurement around group mean as an estimate of individual reliability. None of the reports addressed decision consistency (reliability relative to cut-scores).
There is moderate to good correlation between results on repeated administration of dichotic tests to children (r=0.59 to 0.92).

1.10.2. Accuracy Discussion

Evaluations of dichotic test accuracy were importantly limited by the use of reference standards that have limited association with dichotic processing ability, reference standards that are not replicable, biased sampling strategies that don’t allow any estimate of prevalence and are not representative of clinically important populations, evaluation of non-standard interpretations of the dichotic tests, failure to blind reviewers to reduce bias, and inadequate sample sizes. In the absence of a gold standard, studies of accuracy assessed the ability of dichotic procedures to identify children diagnosed with speech-language disorders (Gustafson & Keith, 2005), learning disability (Berrick et al., 1984; Singer et al., 1998), with a history of head trauma (Roberts et al., 1994), or referred for auditory processing testing (Katz, 1998). Although dichotic listening deficits may be more prevalent in these groups, not all the children would be expected to have dichotic deficit. Investigator judgment about group assignment (i.e., likely to have dichotic problem or not) is not a replicable reference standard unless the criteria for group assignment are clearly specified. Different recruitment procedures for cases and controls are a known contributor to spectrum bias, which results in falsely high estimates of both sensitivity and specificity caused by focusing on the extreme ends of the performance spectrum (clearly disordered and clearly normal or high-performing) (Ransohoff & Feinstein, 1978). Case-control sampling also limits the generalizability of results because it fails to inform about the prevalence of disorder in the population. Without a reasonable
estimate of prevalence, clinicians cannot begin to estimate the probability that a child with an abnormal dichotic score actually has a dichotic deficit. The sensitivity and specificity observed for exploratory cut-scores used in Singer et al (1998) and Berrick et al (1984) will not apply to tests administered according to the publisher’s instructions.

Nonetheless, as studied, dichotic tests don’t perform well. The sensitivity of dichotic procedures for identifying children with classroom learning disability or auditory referral ranged from 0.31 to 0.87. Specificity ranged 0.82 to 0.95 depending on how cut scores were selected.

1.10.3. Usefulness Discussion

We identified four studies that presented individual performance on dichotic tests pre- and post-intervention, but none contribute evidence that dichotic tests are useful or necessary to identify children who can benefit from dichotic treatment.

1.10.4. Value Discussion

Only one study was identified that set out to address the question of value of auditory processing tests in children, and none of the dichotic tests performed sufficiently well to justify inclusion in that study’s value analysis (Singer et al., 1998). Without an understanding of the relative probabilities of making a wrong decision (accuracy) or the benefit that could be reasonably expected from good interventions (usefulness), it’s impossible to meaningfully assert the value of any of these tests.

1.10.5. Limitations

The primary limitation of this review is that we may have missed some relevant sources of evidence. In addition, the results are only generalizable to dichotic tests.
available at the time of the review for assessing binaural integration in English speaking children.

1.10.6. Next Steps

Since we should not perform measurements on our patients for which the benefits cannot be demonstrated to be greater than the costs, we recommend discontinuing dichotic testing of children until evidence is provided to demonstrate reliability, accuracy, usefulness, and value. Reliability studies should be designed and reported in accordance with published guidelines (Kottner et al., 2011). Accuracy studies should adhere to STARD Criteria (Bossuyt et al., 2015) (which are required for publication in ASHA journals). In the absence of a gold standard for diagnosis of APD, accuracy studies are particularly challenging and research may need to simultaneously address questions of accuracy and usefulness by determining whether dichotic test performance can predict outcomes from intervention.
### 1.11. Tables

#### Table 1-1: Dichotic tests for evaluating binaural integration of native English-speaking children

<table>
<thead>
<tr>
<th>Test Name</th>
<th>Abbrev</th>
<th>Publisher</th>
<th>Stimuli</th>
<th>Items</th>
<th>Scores&lt;sup&gt;b&lt;/sup&gt;</th>
<th>References for children:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Dichotic Digits Test aka VA 1-, 2-, &amp; 3-pair Dichotic Digits</td>
<td>RDDT</td>
<td>Arizona State Univ. Foun.</td>
<td>Digits</td>
<td>216</td>
<td>10</td>
<td>Moncrieff et al., 2016&lt;sup&gt;a&lt;/sup&gt;; Moncrieff &amp; Wilson, 2009&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>VA 2-pair Dichotic Digits</td>
<td>DD-VA</td>
<td>Arizona State Univ. Foun.</td>
<td>Digits</td>
<td>80&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2</td>
<td>McDermott et al., 2016</td>
</tr>
<tr>
<td>AUDiTEC Dichotic Digits, Part 2</td>
<td>DD-A</td>
<td>Auditec, Inc.</td>
<td>Digits</td>
<td>200</td>
<td>1</td>
<td>auditecinfo, 2015&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Musiek Dichotic Digits, Version 2</td>
<td>DD-M</td>
<td>Auditec, Inc.</td>
<td>Digits</td>
<td>80</td>
<td>3</td>
<td>auditecinfo, 2015&lt;sup&gt;b&lt;/sup&gt;; Musiek, 1983&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Dichotic Digits subtest of Differential Screening Test for Processing</td>
<td>DD-DSTP</td>
<td>PRO-ED</td>
<td>Digits</td>
<td>12</td>
<td>1</td>
<td>Richard &amp; Ferre, 2006</td>
</tr>
<tr>
<td>Dichotic Consonant Vowel Test (D-CV)</td>
<td>CV-A</td>
<td>Auditec, Inc.</td>
<td>CV</td>
<td>60</td>
<td>3</td>
<td>auditecinfo, 2015&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Bergen Dichotic Listening Tests with Consonant Vowel Syllables</td>
<td>CV-B</td>
<td>University of Bergen</td>
<td>CV</td>
<td>30</td>
<td>3</td>
<td>Hugdahl, 2013</td>
</tr>
<tr>
<td>Competing Words Free-Recall Subtest of SCAN-3:C and A</td>
<td>CW</td>
<td>Auditec, Inc.</td>
<td>Words</td>
<td>40</td>
<td>2</td>
<td>Keith, 2009&lt;sup&gt;b&lt;/sup&gt;, 2009&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Dichotic Words Test</td>
<td>DWT</td>
<td>Dichotics Inc</td>
<td>Words</td>
<td>50</td>
<td>3</td>
<td>Moncrieff, 2015&lt;sup&gt;c&lt;/sup&gt;; Moncrieff et al., 2016</td>
</tr>
<tr>
<td>Dichotic Word Listening Test</td>
<td>DWLT</td>
<td>Auditec, Inc.</td>
<td>Words</td>
<td>120 or 60</td>
<td>4</td>
<td>Roberts et al., 1994</td>
</tr>
</tbody>
</table>

<sup>a</sup> recording contains 2 dichotic pairs in each of 25 trials. Norms in McDermott et al (2016) calculated for scores based on responses to 20 trials.

<sup>b</sup> Number of calculated scores reflects the number of cut scores that can be used to evaluate children’s performance on a test.
### Table 1-2 Evidence of Test-Retest Reliability for Dichotic Speech Tests in Children

<table>
<thead>
<tr>
<th>Test(\text{a})</th>
<th>Score</th>
<th>Number (n) (b)</th>
<th>Correlation (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DD-M</td>
<td>% Correct for Each Ear</td>
<td>&lt;4 (c)</td>
<td>0.77</td>
</tr>
<tr>
<td>DD-DSTP</td>
<td>Trials with all responses correct</td>
<td>78</td>
<td>0.74</td>
</tr>
<tr>
<td>CV-B</td>
<td>Unknown</td>
<td>Unknown</td>
<td>0.70-0.80</td>
</tr>
<tr>
<td>SSW</td>
<td>C-SSW (corrected number of total errors)</td>
<td>15</td>
<td>0.92</td>
</tr>
<tr>
<td>SSW</td>
<td>Reversals</td>
<td>15</td>
<td>0.49</td>
</tr>
<tr>
<td>CW of SCAN-3:C</td>
<td>Scaled Score</td>
<td>48</td>
<td>0.59</td>
</tr>
<tr>
<td>CW of SCAN-3:A</td>
<td>Scaled Score</td>
<td>&lt;58 (c)</td>
<td>0.69</td>
</tr>
</tbody>
</table>

\(a\)DD-M Musiek Dichotic Digits, Version 2 (double pairs); DD-DSTP Dichotic Digits subtest of DSTP; CV-B Bergen Dichotic Listening Tests with Consonant Vowel Syllables; SSW Staggered Spondaic Word Test – List EC; CW Competing Words Free-Recall Subtest;

\(b\)Number of subjects younger than 14;

\(c\)Mixed age group, unclear how many were children
<table>
<thead>
<tr>
<th>Evidence Source</th>
<th>Index Test</th>
<th>Evaluated Score</th>
<th>Cut Score</th>
<th>Reference Standard</th>
<th>n normal</th>
<th>n abnormal</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Singer et al., 1998)</td>
<td>DD-M</td>
<td>Total % correct</td>
<td>Exploratory (10% false positive rate in sample)</td>
<td>Group assignment: classroom learning disability or control</td>
<td>91</td>
<td>147</td>
<td>0.54</td>
<td>0.89</td>
<td>0.68</td>
</tr>
<tr>
<td>(Singer et al., 1998)</td>
<td>SSW</td>
<td>% correct of RC+LC</td>
<td>Exploratory (10% false positive rate in sample)</td>
<td>Group assignment: classroom learning disability or control</td>
<td>91</td>
<td>147</td>
<td>0.33</td>
<td>0.89</td>
<td>0.53</td>
</tr>
<tr>
<td>(Katz, 1998)</td>
<td>SSW</td>
<td>Any one of: C-RNC, C-RC, C-LC, C-LNC</td>
<td>Clinical (2 SD below mean from Combined National Sample)</td>
<td>Group assignment: referred for auditory evaluation because of LD or control group</td>
<td>140</td>
<td>171</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>(Berrick et al., 1984)</td>
<td>SSW</td>
<td>Either of: C-LC or C-RC</td>
<td>Exploratory: 2 SD below mean in study sample</td>
<td>Group assignment: classroom learning disability or control</td>
<td>93</td>
<td>97</td>
<td>0.56</td>
<td>0.95</td>
<td>0.74</td>
</tr>
<tr>
<td>(Roberts et al., 1994)</td>
<td>DWLT</td>
<td>Number of trials with both correct</td>
<td>10th percentile from normative sample</td>
<td>Group assignment: head injury or control</td>
<td>142</td>
<td>21</td>
<td>0.43</td>
<td>0.90</td>
<td>0.84</td>
</tr>
</tbody>
</table>

DD-M Musiek Dichotic Digits, Version 2; SSW Staggered Spondaic Word Test – List EC; DWLT Dichotic Word Listening Test (30 items)
1.12. References


29


CHAPTER 2: COMPARISON OF DICHOTIC LISTENING TEST-RETEST RELIABILITY IN CHILDREN

2.1. Abstract

Purpose: To compare test-retest reliability of three dichotic listening tests: SCAN-3 Competing Words, Double Dichotic Digits, and Bergen Dichotic Listening Test with Consonant-Vowel Syllables.

Method: Sixty English-speaking children, 7-14 years old with normal hearing, had a single study-visit during which each test was administered twice. Changes on retest were compared to binomial model predictions, summarized by within-subject standard deviation (Sw), and compared among tests. Correlates of variance were explored.

Results: All 3 tests had reliability within bounds predicted by binomial model. Scores based on 40 items were more reliable (Sw=5%) than those based on 20-30 items (Sw=6-8%). No associations between participant characteristics and reliability were found.

Conclusions: Digits Right, Digits Left, and Words Total Scores – each based on 40 items-- had the best reliability among the clinically used scores, but smallest detectable difference on retest averaged 14%. Scores based on fewer items were even less precise. No other characteristic of the tests or the participants were associated with reliability. Poor precision may contribute to misdiagnosis in clinic and non-differential misclassification in research. More precise estimates of dichotic listening ability require longer or adaptive tests. Audiologists may need to administer more items if children’s dichotic scores fall close to cut-offs.
2.2. Introduction

Little has been published about test-retest reliability of dichotic listening tests, which are commonly used to identify children with auditory processing disorder (American Academy of Audiology, 2010; Chermak et al., 2007; Emanuel, 2002; Emanuel et al., 2011). All dichotic listening tests present to a listener a different stimulus (e.g., target word) in each ear simultaneously. Listeners repeat some or all of what is heard, and results are believed to give insight into neuro-audiologic function (American Academy of Audiology, 2010; Kimura, 1961; Musiek, 1983b). Although “some of the tests for (C)APD in current clinical use lack rigorous psychometric design, construction, and validation,” “appropriate and substantial evidence” is believed to be available to support the use of dichotic listening tasks (American Academy of Audiology, 2010, p. 15).

A listener’s score on a dichotic listening test is intended to be an estimate of underlying dichotic listening ability, but reliability of the score in practice may be degraded by a number of factors (Psychological Testing, 2004): general characteristics of the listener (e.g., health, fatigue, motivation, emotional strain, and linguistic competence), interaction of the listener with the test (e.g., comprehension of the specific test task, specific tricks for dealing with the particular test materials, and fluctuations of attention), the listener’s experience of the test environment (e.g., freedom from distractions, clarity of instructions, interaction of with the audiologist), the performance of the audiologist administering the test (e.g., hearing acuity, distraction, difficulty interpreting listener’s speech), and chance.
Reliability refers to the repeatability or consistency of measurements. If we assume that a listener’s underlying neurologic structures driving dichotic listening ability don’t change between test administrations, changes in individuals’ scores (within-subject variation) can be attributed to measurement error. Measurement error decreases reliability (the consistency of scores) and precision (the width of the confidence interval around a score). From within-subject variance, a number of useful metrics can be calculated: within-subject standard deviation ($S_w$, average change around subject’s own mean score), precision (width of confidence intervals around a score), and the smallest difference that can be detected between two scores on the same test.

Few studies report details about within-subject variance (Kottner et al., 2011). Instead, many studies report mean change on retest, which can reveal biases (e.g., learning or fatigue) in a group (Amos & Humes, 1998), but do not tell the clinician how much measurement error might be incorporated in individual scores. It is also common to report linear correlation between individuals’ first and second test scores. Correlation is a poor measure of reliability since it is sensitive to outliers, is sensitive to the range of measurement, shows a strong relationship in the presence of systematic changes over time, and tests significance against a null-hypothesis of no relation which is not sensible for two measurements of the same individual (Newman & Kohn, 2009). Finally, few reports of test-retest reliability describe their subjects in enough detail (e.g., age, original score) to know how the results should be generalized (Kottner et al., 2011).

Thornton and Raffin expressed these concerns about reliability studies’ reporting of mean change and linear correlation, rather than within-subject standard-deviation, in
their 1978 study of adults’ speech discrimination scores (Thornton & Raffin, 1978). They demonstrated that within-subjects variance of adults’ speech discrimination scores increased as the number of items used to calculate the speech discrimination score decreased and that smaller changes occurred at extreme scores (i.e., closer to 0% and 100%). They concluded that within-subjects variance in speech audiometry is explained by the amount of chance variation predicted by the binomial distribution, is dependent only on the listener’s proportion of correct responses and number of words in a list, and that “all other characteristics of the population are irrelevant with respect to variability across test forms”. However, others have observed children’s scores on psychoacoustic tests are usually less accurate and reliable than that of adults. (Moore, Cowan, Riley, Edmondson-Jones, & Ferguson, 2011). The increased variance of children’s scores has been attributed to differences in motivation and attention. It is unknown whether non-auditory factors cause greater variance in supra-threshold speech audiometry scores, such as those from dichotic listening tests, than expected by chance.

2.2.1. Characteristics of Selected Dichotic Tests

Dichotic listening test recordings vary in the number and type of stimuli (e.g., syllables, digits, words), the directions to the listener (e.g., repeat everything—free recall; repeat the item that is most clear—non-directed ear; or repeat from one ear—directed ear), interpreted score (e.g., number or percent correct of right- and left-ears separately, total correct in both ears, right minus left, ratio of left to right, etc.), and reference criteria.
The **Dichotic Digits** test (Musiek, 1983a) is among those recommended by the American Academy of Audiology *Guidelines for the Diagnosis, Treatment and Management of Children and Adults with Central Auditory Processing Disorder* (American Academy of Audiology, 2010). Published data on test-retest reliability is limited. In a study of 4 individuals with “relatively stable” lesions of the central auditory nervous system, retest scores were correlated \((r=0.77)\) with results obtained 2-12 months earlier (Musiek et al., 1991). Strouse and Hall (1995) found that 20 adults, 10 with Alzheimer’s disease, retested at 2 months, had ear specific retest correlations of \(r=0.82\) to 0.97 with no changes more than 14% points. No studies have reported test-retest data for Dichotic Digits administered to children.

**Dichotic Consonant-Vowel Syllables** tests are recommended by the guidelines for use when a dichotic test with a low-linguistic load is desired (American Academy of Audiology, 2010). One of these, The Bergen Dichotic Listening Test with Consonant-Vowel Syllables, has been shown to identify the same ear as dominant after one year in 77% of a sample of Swedish children (Andersson & Hugdahl, 1987). Correlation of first and second laterality indices (right ear score minus left ear score divided by total number of items) was reported as 0.608. No similar study has been reported for US children.

**SCAN Competing Words** subtests (directed ear and free recall) are among the most commonly used dichotic tests (Emanuel et al., 2011). Their use appears to be increasing (Chermak et al., 2007; Emanuel, 2002; Emanuel et al., 2011) despite concerns that the scores may not be reliable (Bellis, 2002). Correlation between children’s first and second scores of an early version of the SCAN Competing Words test was low \((r = 0.44\)
to 0.73) (Cacace & McFarland, 1998), and mean score was shown to improve significantly on retest (Amos & Humes, 1998). Although the publisher’s manual for the SCAN-3:C and SCAN-3:A (Keith, 2009a, 2009b) reports test-retest correlation has improved ($r=0.59$ for $n=48$ children; $r=0.69$ for $n=58$ adults/adolescents) from previous versions, no independent evaluations of the its reliability have yet been published.

The present study examines children’s within-subject variation on same-session retest for the Double Dichotic Digits test, the Bergen Dichotic Listening Test with Consonant-Vowel Syllables, and the SCAN-3 Competing Words Free-Recall tests. We sought to determine whether any of the tests showed evidence of systematic changes on retest (i.e., change in mean). We tested Thornton and Raffin’s hypothesis that random binomial variation fully describes within-subject variance regardless of listener characteristics by testing the number of scores falling outside the 95% confidence interval for difference between scores predicted by the binomial theorem. Finally, we sought to determine if reliability was associated with any participant characteristics that would merit further study. We followed the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) in the analysis and reporting of results (Kottner et al., 2011).

2.3. Methods

We studied repeatability of children’s scores for three dichotic listening test recordings: Musiek’s Double Dichotic Digits test, the Bergen Dichotic Listening Test with Consonant-Vowel Syllables, and the SCAN-3 for Children Competing Words. Every participant had a single study-visit during which each test was administered twice.
We analyzed differences between first and second administration of each dichotic test. We looked for associations between variance and participant characteristics.

2.3.1. Subjects

We studied child volunteers between the ages of 7-14 with normal hearing who spoke English and were able to complete a study visit. Participants were recruited from the community using flyers posted in public places (e.g., pediatricians’ offices) or distributed at public events (e.g., sporting events), and through social media postings (e.g., list serves and Facebook). Participants were given a report of test performance and coupons donated by local businesses (e.g., arcade card, free beverage) to thank them for their time and cooperation.

The study was approved by the University of Vermont Committee on Human Research in the Medical Sciences. Guardian consent and child-consent (age 11 and up) or -assent (age 10 and under) were documented at the start of the visit. The Principal Investigator administered all aspects of the protocol, including scheduling, testing, and referral for clinical services as needed.

2.3.2. Setting

Participants each made a single visit to the University of Vermont Medical Center Otolaryngology office in Berlin, VT, April-June 2014. Testing was conducted in an Industrial Acoustics Company double-walled audiology booth meeting ASNI standard S3.1-1999 (ANSI, 1999) inside the audiometric test suite using a GSI-61 clinical audiometer and the participant’s choice of EAR-Tone 3A insert earphones or Telephonics TDH-50P headphones calibrated to ANSI standard S3.6-1996 (ANSI, 1996). Dichotic
stimuli were presented from a commercially available (Denon) CD changer. All dichotic
tests were administered at 50 dB HL on the audiometer dial with the SCAN calibration
tone adjusted to 0 VU.

2.3.3. Measures

The order of administration of the three tests and the routing of the stereo
channels (i.e., right and left) was determined by block randomization.

**Double Dichotic Digits Test (Digits)**

The stimuli for Musiek’s Double Dichotic Digits Test (Musiek, 1983a) are audio-
recordings of single-syllable digits (1-10 excluding 7) spoken in English by a male voice
with two digits presented to each ear (four digits total per trial). Participants were
instructed by the test administrator that they would be hearing different numbers in each
ear at the same time and should repeat all of the numbers heard, regardless of order. After
4 practice trials, the full test comprises 20 trials (40 digits per ear= 80 digits total). The
recording was paused, per test protocol, during inter-stimulus intervals if listeners
required more time to respond. Right- and left-ear percent correct scores were calculated
by dividing the number of correct responses to stimuli presented to each ear by the
number of stimuli (40 per ear) and multiplied by 100. Although not used clinically, we
also calculated Digits Total correct (number correct divided by 80 times 100) for
comparison of total correct scores among tests.

**Bergen Dichotic Listening Test with Consonant-Vowel Syllables (Syllables)**

The stimuli for the Bergen Dichotic Listening Test with Consonant-Vowel
Syllables (K Hugdahl & Hammar, 1997) are audio-recordings of a male voice speaking
the six stop consonants paired with /a/ (i.e., /ka/, /ga/, /ta/, /da/, /pa/, /ba/). The listener was directed by test administrator to look at this list of syllables (posted on the wall at eye level) and to repeat whatever one syllable was heard loudest or most clearly without spending time to think about what was heard. The list comprises 36 trials. Every syllable pairing is presented twice during the test counterbalanced between ears including 6 diotic pairs in which the same syllable is presented to both ears. Only responses to the 30 dichotic presentations were scored. There are no practice items. The number of correct responses to stimuli presented to the right ear were divided by the number of stimuli (30 possible) and multiplied by 100 to get a percent correct score. The same calculation was performed for the number of correct responses to stimuli presented to the left ear (30 possible). Since listeners were instructed to repeat only one word per trial, the maximum number of correct responses possible is 30, the same as the number of trials. Therefore, the Total score for the Syllables test is the number of correct responses regardless of which ear received the stimulus repeated divided by the total number of trials (30) times 100. Note that this results in a Total percent correct score for syllables that is the sum of the Right percent correct and Left percent correct rather than the mean of the two ears, as it is with the other tests.

**SCAN-3:C and SCAN-3:A Competing Words (Words)**

The stimuli for the SCAN Competing Words test are audio-recordings of single-syllable English words spoken by a male voice. One target word was presented to each ear during each of 20 trials (40 words total) after recorded instructions and two practice trials. Participants were instructed to repeat both words for each trial. Words Right
percent correct scores were calculated by summing the number of correct responses to stimuli presented to the right ear, dividing by 20, and multiplying the proportion by 100. Words Left scores were calculated similarly. Total scores were calculated for each participant by summing the total number of correct responses and dividing by 40 before multiplying by 100. For clinical use, there is a right ear and a left ear list. In our study, the routing of the stereo channels was randomized, so approximately half the children had the “right list” directed to their left ear.

**Hearing Sensitivity**

Hearing sensitivity was measured at 500, 1000, 2000, and 4000 Hz using a modified Hughson-Westlake method (Carhart & Jerger, 1959).

**Questionnaire**

A one page questionnaire completed by the adult accompanying the subject at the visit captured the following information about the child participants: demographic information (sex, race, ethnicity, educational status of adults in household, household income); handedness; difficulties with academics, attention, or development; specials services received; medications; general health; musical training; and symptoms of auditory processing disorder. Parents were asked to rate their child for 13 symptoms of auditory processing disorder listed in the AAA Clinical Guidelines (American Academy of Audiology, 2010) using a scale of 0 (Never) to 3 (Most of the time).

**2.3.4. Data**

Study data were collected and managed using the REDCap (Research Electronic Data Capture) system (Harris et al., 2009) hosted at the University of Vermont. For every
dichotic test item, the examiner judged whether the child responded correctly or not. This was immediately entered into forms created in REDCap for this study.

2.3.5. Analysis

Right-(ear), Left-(ear), and Total-Percent Correct scores were analyzed for each test. The authors sought to determine if there were any systematic differences in within-subject variance among dichotic tests under study. Changes in mean score on retest were evaluated using t-tests. Sign-tests were used to determine if the number of participants who improved or declined on retest was larger than predicted by chance. Estimates of mean within-subject standard deviation ($S_w$) were calculated for each score and presented with 95% confidence intervals. Mean Repeatability Coefficient (CR, also called the smallest detectable difference between two scores on the same test) was calculated for each test by multiplying mean $S_w$ by 2.77 ($\sqrt{2}$ times 1.96). (Vaz, Falkmer, Passmore, Parsons, & Andreou, 2013).

Bland-Altman plots were used to display participants’ changes in score on retest vs. the mean of the two scores. Binomial 95% confidence intervals for differences between scores are displayed as ellipses drawn using the formula:

$$2.77\sqrt{(p \times (1 - p))/n)},$$

where $p$ is the proportion of total responses judged to be correct (mean score) and $n$ is the number of items on the test. The formula is a logical combination of the Repeatability Coefficient formula ($2.77*S_w$) and the formula for within subject standard deviation of a binomially distributed variable from Thornton and Raffin (1978). The plots allow readers to see the predicted Repeatability Coefficient for every possible mean score (0-100%). We used the binomial exact test to determine if 5%
of scores fell outside the 95% confidence interval of the Repeatability Coefficients predicted by the binomial theorem.

Variance between groups was compared using t-tests for reported sex, handedness, race/ethnicity (non-Hispanic white vs. any other), ADHD (present vs. absent), school services (receiving vs. not receiving), and phones used (insert vs. TDH). Linear regression was used to explore the relationship between variance and age, adult rating of APD symptoms, parental education, and test order. For all tests, a two-tailed $P$-value < 0.05 was required for statistical significance. Analyses were conducted in Stata version 13 (StataCorp, 2013).

2.4. Results
2.4.1. Subjects

Sixty-five children presented for the study; 5 were excluded because of hearing loss. Characteristics of the 60 eligible participants are summarized in Table 2-1. All had normal hearing sensitivity (thresholds 20 dB HL or better at all tested frequencies) and symmetrical hearing (no more than 10 dB difference in thresholds between ears at each frequency). They ranged in age from 7-14 years old (1st-8th grade) with a mean of 10.0 years (SD 1.9 years). Thirty-one were female (52%) and 56 were right handed (93%). Fifty-five (92%) attended public school; 5 were homeschooled. Ten participants had been diagnosed with ADHD. Fifty-one (85%) identified as non-Hispanic white. Children from 38 families participated in the study (17 single children, 20 sibling pairs, one family brought 3 eligible children). Annual income of participants’ families spanned the range of
questionnaire categories. Thirty-six participants (60%) were from homes with an adult who had post-graduate education.

Adults were asked to rate their child(ren) for 13 symptoms of auditory processing disorder using a scale of 0 (Never) to 3 (Most of the time). Participants’ ratings, averaged across the 13 symptoms, ranged 0 to 2.8 (mean: 0.8, std. dev: 0.7). Five participants’ average symptom ratings exceeded 2.

2.4.2. Test-Retest Reliability

Participant scores ranged 0-90% for Words, 23-100% for Digits, and 10-100% for Syllables. Detailed summaries of all the scores are presented in Table 2-2. Mean score did not change significantly on retest for any of the dichotic tests. The absolute change in Words test mean score was small (1.8%) but approached statistical significance (p=0.06). Sign-test revealed no significant difference in the number of participants whose Words Total scores increased (n=34) vs. decreased (n=20) (p=0.08).

Individual participants’ changes on retest are summarized in Table 2-3. Within-subject standard deviation (Sw) increased as the number of items used to calculate the score decreased. In Figure 2-1, we present individual participants’ change on retest as a function of mean score. The ellipses show the predicted 95% confidence limits of the binomial Repeatability Coefficients for all possible scores and highlight that participants’ scores changed less as they neared the extremes (i.e., closer to 0% and 100%) as predicted by the binomial model. We evaluated whether the binomial distribution described our data by testing the number of scores falling outside the range predicted by the binomial Repeatability Coefficient ellipses. Our hypothesis was that 5% of scores
would fall outside the 95% confidence interval predicted by the binomial theorem. The exact number of participants whose score change fell outside the 95% confidence interval ranged from none (for Words and Syllables left ears) to five (for Digits left)—none significantly different than the 3 expected from a sample of 60.

We found no evidence of a systematic effect on reliability for sex, handedness, race/ethnicity, school services, ADHD diagnosis, APD symptoms, parental education, income, order of enrollment, or phones used by linear regression. Initial analysis showed older children had smaller variance on Words Total Score, even when controlling for score (2.5% smaller $S_w$ per year of age, $p = 0.047$). However, post-hoc analysis revealed that this effect was largely driven by a single participant who had both the largest change in score (23%) and youngest possible age (7 years old). When this participant’s score was excluded from consideration, the estimated association between $S_w$ and age decreased (1.6% smaller $S_w$ per year of age) and no longer achieved significance ($p = 0.28$). No consistent effect of age on reliability was seen for Digits or Syllables. No other associations between participant characteristics and reliability were found.

2.5. Discussion

The purpose of the present study was to examine and compare children’s within-subject variation on retest of the Double Dichotic Digits test, the Bergen Dichotic Listening Test with Consonant-Vowel Syllables, and the SCAN-3 Competing Words Free-Recall tests. We found random binomial variation easily explains all the variation observed in the data. We found no evidence to support any systematic sources of variation. We found no association between subject characteristics and variance. In other
words, there is no inherent difference in the reliability of the three tests, other than the number of items used. This is completely consistent with the conclusions of Thornton and Raffin (1978) that “binomial characteristics of speech-discrimination tests are relatively independent of subject characteristics, listening conditions, and type of stimulus.” It also explains the unexpected finding of no association between age and reliability. Although younger subjects may have less reliable performance on some psychoacoustic tasks (Moore et al., 2011), this does not appear to be the case for speech audiometry using supra-threshold stimuli.

The results of this study emphasize that reliability of audiologists’ estimates of individuals’ dichotic listening ability is largely determined by the length of the test selected. The clinically interpretable scores with the best reliability in this study are those calculated based on 40 items: Dichotic Digits Left, Dichotic Digits Right, and Words Total. Within-subject variance is small enough that the smallest detectable difference is a change of 14-15%. A change as small as 10% is detectable on Digits Total score, calculated using 80 items. The smallest detectable differences on scores calculated using 20-30 items are closer to 20%.

2.6. Limitations

The application of the binomial model, which assumes independence of each measure, to our analysis of repeated measures could be considered a limitation of this study. However, the fact that our data largely conformed to the predictions of the binomial model supports using it in this way. We considered a model that included covariance, but put it aside because the simpler model appears to serve adequately.
The results from this volunteer sample in a single rural area, recruited by word of mouth and passing of flyers, may have limited generalizability. For instance, the children tended to be white and from affluent homes with college educated parents. Nonetheless, participants presented with a wide range of both auditory processing symptom scores (0/3 to 2.8/3) and dichotic test scores (Words 0-90%; Digits 22-100%; Syllables 10-73%).

2.7. Conclusions

Reliability of dichotic listening tests is better for scores based on larger numbers of items (40-80) than smaller numbers of items (20-30). The smallest detectable difference on retest for clinically interpreted scores based on 40 items (each ear of Dichotic Digits and SCAN Words Total) averaged 14%. Scores based on fewer items were even less precise. Poor precision may contribute to misdiagnosis in clinic and non-differential misclassification in research. More precise estimates of dichotic listening ability require longer or adaptive tests.
### 2.8. Tables and Figures

**Table 2-1 Characteristics of 60 Participants**

<table>
<thead>
<tr>
<th>Characteristics</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, mean (SD)</td>
<td>10.0</td>
<td>(1.9)</td>
</tr>
<tr>
<td>Female, No. (%)</td>
<td>31</td>
<td>(52)</td>
</tr>
<tr>
<td>Right Handed, No. (%)</td>
<td>56</td>
<td>(93)</td>
</tr>
<tr>
<td>Non-Hispanic White (%)</td>
<td>51</td>
<td>(85)</td>
</tr>
<tr>
<td>School Services, No. (%)</td>
<td>22</td>
<td>(37)</td>
</tr>
<tr>
<td>ADHD Diagnosis, No. (%)</td>
<td>10</td>
<td>(17)</td>
</tr>
<tr>
<td>Annual Household Income &lt;$25K, No. (%)</td>
<td>4</td>
<td>(7)</td>
</tr>
<tr>
<td>$25K-74K, No. (%)</td>
<td>23</td>
<td>(38)</td>
</tr>
<tr>
<td>$75K-$99K, No. (%)</td>
<td>21</td>
<td>(35)</td>
</tr>
<tr>
<td>$100K-more, No. (%)</td>
<td>12</td>
<td>(20)</td>
</tr>
<tr>
<td>Parent with post-college education, No. (%)</td>
<td>36</td>
<td>(60)</td>
</tr>
<tr>
<td>APD Symptom Rating (0-3), mean(SD)</td>
<td>0.8</td>
<td>(0.7)</td>
</tr>
</tbody>
</table>
# Table 2-2 Summary of First and Second Scores (% Correct) on Dichotic Digits, Dichotic CV Syllables, and Competing Words tests (n=60).

<table>
<thead>
<tr>
<th>Test</th>
<th>/# items</th>
<th>First Time</th>
<th>Second Time</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
<th>Difference</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digits(^a)</td>
<td>Left/40</td>
<td>84% (15%)</td>
<td>28,100</td>
<td>84%</td>
<td>15%</td>
<td>22,100</td>
<td>&lt;1%</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Right/40</td>
<td>90% (9%)</td>
<td>65,100</td>
<td>89%</td>
<td>11%</td>
<td>63,100</td>
<td>&lt;1%</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>Total/80(^a)</td>
<td>87% (11%)</td>
<td>47,100</td>
<td>87%</td>
<td>12%</td>
<td>44,100</td>
<td>&lt;1%</td>
<td>0.90</td>
</tr>
<tr>
<td>Syllables(^b)</td>
<td>Left/30</td>
<td>36% (10%)</td>
<td>17,63</td>
<td>36%</td>
<td>11%</td>
<td>10,63</td>
<td>1%</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>Right/30</td>
<td>50% (11%)</td>
<td>20,73</td>
<td>50%</td>
<td>10%</td>
<td>33,73</td>
<td>1%</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>Total/30(^b)</td>
<td>85% (8%)</td>
<td>67,100</td>
<td>87%</td>
<td>8%</td>
<td>67,100</td>
<td>1%</td>
<td>0.19</td>
</tr>
<tr>
<td>Words(^a)</td>
<td>Left/20</td>
<td>55% (19%)</td>
<td>0,90</td>
<td>57%</td>
<td>20%</td>
<td>5,90</td>
<td>2%</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>Right/20</td>
<td>64% (14%)</td>
<td>30,90</td>
<td>66%</td>
<td>13%</td>
<td>30,90</td>
<td>2%</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>Total/40(^c)</td>
<td>59% (14%)</td>
<td>23,90</td>
<td>61%</td>
<td>14%</td>
<td>30,88</td>
<td>2%</td>
<td>0.06</td>
</tr>
</tbody>
</table>

\(^a\) Participants instructed to repeat all stimuli (total is proportion of correct responses to all stimuli presented to either ear)

\(^b\) Participants instructed to repeat one stimulus per trial (total is proportion of correct responses to all trials, either left or right ear correct response)
<table>
<thead>
<tr>
<th>Test</th>
<th>Score</th>
<th># Items&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Sw&lt;sup&gt;b&lt;/sup&gt; (95% CI of Sw)</th>
<th>Range of Change</th>
<th>CR&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digits</td>
<td>Total</td>
<td>80</td>
<td>3.7% (2.1, 5.4)</td>
<td>-14%, 11%</td>
<td>10%</td>
</tr>
<tr>
<td>Digits</td>
<td>Right</td>
<td>40</td>
<td>5.0% (2.8, 7.2)</td>
<td>-20%, 13%</td>
<td>14%</td>
</tr>
<tr>
<td>Digits</td>
<td>Left</td>
<td>40</td>
<td>5.4% (2.8, 8.0)</td>
<td>-23%, 23%</td>
<td>15%</td>
</tr>
<tr>
<td>Syllables</td>
<td>Total</td>
<td>30</td>
<td>6.0% (3.3, 8.7)</td>
<td>-17%, 23%</td>
<td>17%</td>
</tr>
<tr>
<td>Syllables</td>
<td>Right</td>
<td>30</td>
<td>7.2% (4.1, 10.3)</td>
<td>-27%, 20%</td>
<td>20%</td>
</tr>
<tr>
<td>Syllables</td>
<td>Left</td>
<td>30</td>
<td>7.0% (4.1, 9.9)</td>
<td>-23%, 23%</td>
<td>19%</td>
</tr>
<tr>
<td>Words</td>
<td>Total</td>
<td>40</td>
<td>5.3% (3.0, 7.6)</td>
<td>-17%, 23%</td>
<td>15%</td>
</tr>
<tr>
<td>Words</td>
<td>Right</td>
<td>20</td>
<td>7.9% (4.5, 11.3)</td>
<td>-30%, 25%</td>
<td>22%</td>
</tr>
<tr>
<td>Words</td>
<td>Left</td>
<td>20</td>
<td>7.2% (4.3, 10.2)</td>
<td>-25%, 20%</td>
<td>20%</td>
</tr>
</tbody>
</table>

<sup>a</sup> Number of items contributing to score
<sup>b</sup> Sw: Within-Subject Standard Deviation
<sup>c</sup> CR: Coefficient of Repeatability, also called Smallest Detectable Difference
Figure 2-1 Difference between first and second scores of 60 participants with binomial 95% CI for differences
2.9. References


StataCorp. (2013). Stata Statistical Software: Release 13. College Station, TX: StataCorp LP.


CHAPTER 3: COMPARISON OF CHILDREN’S DOUBLE DICHOTIC DIGITS AND SCAN-3 COMPETING WORDS FREE-RECALL SCORES

3.1. Abstract

Purpose: We sought to compare SCAN-3 Competing Words Free Recall (CW) and Musiek’ Dichotic Digits (DD) tests to determine if they could be used interchangeably to identify ear advantage or dichotic deficit in children.

Methods: CW and DD were administered to 60 children (aged 7-14). We used Spearman’s rho (ρ) to assess associations between rank-ordering of participants and kappa statistic (κ) to assess decision consistency.

Results: Participants were rank ordered similarly by CW and DD using right ear (ρ = 0.58), left ear (ρ = 0.51) and total (ρ = 0.73) scores, but not by inter-aural asymmetry (ρ =0.18). The tests agreed no better than chance on direction of ear advantage (κ= 0.01, p = 0.93) and had poor agreement on which children displayed dichotic deficits (κ = 0.22, p < 0.01). DD identified significantly more participants with deficits (n=18) than CW (n=3) (p < 0.001).

Conclusions: Although children with high scores on one test tend to have high scores on the other, CW and DD do not agree on ear advantage or presence of deficit. They are not interchangeable for clinical use. Additional research is needed to determine if either is appropriate for identifying children who would benefit from treatment for dichotic listening deficits.
3.2. Introduction

Free-Recall dichotic listening tests, in which a listener is asked to repeat everything heard when different stimuli are presented simultaneously to each ear, are a staple of auditory processing evaluation of children in the United States (American Academy of Audiology, 2010; Chermak et al., 2007; Emanuel et al., 2011). Dichotic tests may give insight into the organization and capacity of the auditory central nervous system (American Academy of Audiology, 2010; Kenneth Hugdahl, 2011). In people with documented lesions, bilaterally low scores on dichotic listening tests are associated with damage to the auditory cortex, and asymmetric right and left ear scores are consistent with damage to the corpus callosum or unilateral lesions (Musiek, 1983a; Musiek et al., 1991). Low and asymmetric dichotic listening scores in children have been associated with reading and language disorders (Abigail & Johnson, 1976; Agnew, 2004; Dlouha, Novak, & Vokral, 2007; Moncrieff & Musiek, 2002). Auditory training may help remediate impairment associated with dichotic listening deficits (Moncrieff & Wertz, 2008; Musiek, 2012).

The American Academy of Audiology Clinical Practice Guidelines for Diagnosis, Treatment and Management of Children and Adults with Central Auditory Processing Disorder (American Academy of Audiology, 2010) recommends including dichotic listening tests as part of the auditory processing test battery. Audiologists are advised to interpret both interaural asymmetry (i.e., difference between right ear and left ear performance by a listener) and listener performance relative to normative cut-off criteria (i.e., two standard deviations below normal-listeners’ mean). The guidelines state “a child
with a typically developing auditory system should [...] have greater right-ear than left-ear scores on dichotic speech tasks. This right-ear advantage diminishes and left-ear performance improves as the child matures. Findings other than these, such as an exaggerated right-ear advantage or a left-ear advantage have implications for the diagnosis of [(central) auditory processing disorder] (C)APD.” They recommend avoiding diagnosis of APD in the face of conflicting test findings such as “right-ear deficit on one task combined with a left-ear deficit on another similar task within the same individual.”

The SCAN-3 Competing Words Free Recall (CW) (Keith, 2009b) and Double Dichotic Digits (DD) (Musiek, 1983a) tests are among the most commonly used dichotic listening tests (Chermak et al., 2007; Emanuel, 2002). Most recently, 73% of audiologists who responded to a survey about auditory processing assessment reported using CW and 88% reported using some version of DD (Emanuel et al., 2011). Both CW and DD are free-recall dichotic speech tests and present the listener with a similar task: to repeat back to the examiner what was heard in both ears. The stimuli differ between tests; CW stimuli are single syllable words and DD stimuli are single-syllable digits (1-10 excluding 7).

Cut-off criteria for CW total scores and ear difference are published in the SCAN-3 manuals (Keith, 2009a, 2009b) and are derived from a normative sample stratified by age, sex, race/ethnicity, geographic region, and educational level of the primary caregiver. Audiologists are advised to collect local norms for DD (Musiek, 1983a) but sample cut-off criteria for DD right ear and left ear scores are available (Bellis, 2002).
The scores and classification of individual children by free-recall CW and DD have not previously been directly compared in a single sample. The purpose of this study was to compare the performance of free-recall CW and DD in a single group of children. We sought to determine if children who had high scores on one test also scored well on the other test. Since both tests purportedly measure capacity of the right- and left-ear auditory pathways, we’d expect a strong association between scores. We also wanted to determine whether CW and DD could be used interchangeably to describe ear advantage and to identify children with abnormal dichotic listening. Both tests are currently used to contribute to APD diagnosis. In order for the tests to be interchangeable, they must agree on direction of ear advantage and presence of deficit.

3.3. Methods

This is a cross-sectional study of children’s performance on CW and DD tests measured at a single study visit in the spring of 2014. The data were collected as part of a study of test-retest reliability. In the current study, we compare scores from only the first administration of each test.

3.3.1. Subjects

English-speaking volunteers between the ages of 7 and 14, with normal hearing sensitivity and able to complete the study visit, were eligible to participate. Participants were recruited using flyers, word of mouth, and social media postings. Participants were provided with a report of test performance and coupons donated by local businesses (e.g., arcade card, free beverage) to thank them for their time and cooperation.
The study was approved by the University of Vermont Committee on Human Research. Child consent (age 11 and up) or assent (age 10 and under) and guardian consent were documented at the start of the visit. The Principal Investigator administered all aspects of the protocol, including scheduling, testing, and follow-up as needed.

3.3.2. Test Environment and Equipment

Testing was conducted between April and June of 2014 in an International Acoustics Company double-walled audiology booth using a GSI-61 clinical audiometer and the participant’s choice of EAR-Tone 3A insert earphones or Telephonics TDH-50P headphones. All were calibrated to ANSI standard (ANSI, 1999). Dichotic stimuli were presented from a commercially available (Denon) CD changer. All dichotic tests were administered at 50 dB HL on the audiometer dial. Test order and routing of stereo channels was randomized among participants.

3.3.3. SCAN-3 Competing Words (CW)

The Competing Words (CW) subtest of the SCAN-3:C (for Children) was administered to participants 7-11 years old and the SCAN-3:A (for adolescents and adults) was used with participants 12-14 years old (Keith, 2009b). The SCAN-3:C and SCAN-3:A stimuli are identical. The CW stimuli are audio-recordings of single-syllable English words spoken by a male voice. One target word was presented to each ear during each of 20 trials (40 words total). Each target word is scored as correct or incorrect (i.e., child did or did not repeat the target word). Each word pair was presented once during the test. For clinical use there is a right ear and a left ear list. In this study, the routing of the
stereo channels was randomized, so approximately half the children had the “right list” directed to their left ear.

Participants were classified using the age-specific norms published in the SCAN-3:C and SCAN-3:A test manuals. Dichotic deficit on CW was defined as having total score (right plus left) more than 2 standard deviations below the mean for age OR interaural asymmetry (right minus left) more extreme than 96% of the normative sample for age (2% at each tail of the distribution).

3.3.4. Double Dichotic Digits Test (Digits)

Musiek’s Dichotic Digits Test (Hurley & Musiek, 1997) double pairs (DD) was administered to the participants. The stimuli are audio-recordings of single-syllable digits (1-10 excluding 7) spoken in English by a male voice with two consecutive pairs of digits presented to each ear (four digits total per trial). The test comprises 20 trials (40 digits per ear = 80 digits total). The listener was instructed by the test administrator that s/he would be hearing different numbers in each ear at the same time and should repeat all of the numbers heard, regardless of order. As per Musiek’s recommended protocol (Musiek, 1983a), the participants were given as much time as they needed to respond to each trial (the recording was paused if necessary).

Dichotic deficit on DD was defined as having either right ear score or left ear score below the age specific cut-offs published in Bellis (2002).

3.3.5. Hearing Sensitivity

Hearing sensitivity was measured at 500, 1000, 2000, and 4000 Hz using a modified Hughson-Westlake method (Carhart & Jerger, 1959).
3.3.6. Parent Questionnaire

Demographic information (sex, race, ethnicity, education of adults in household and household income) and participants’ characteristics (handedness; difficulties with academics, attention, or development; specials services received; medications; musical training) were captured on a questionnaire completed by the adult(s) accompanying the participant to the study visit.

3.3.7. Analytic Plan

CW and DD right ear, left ear, and total scores are presented as proportion correct (number of correct responses divided by number of stimuli presented). Interaural asymmetry is presented as the number of items different (right ear number correct minus left ear number correct). We used Wilcoxon’s sign-rank test to compare scores between tests and between ears.

We sought to determine if CW and DD were measuring the same underlying phenomenon (efficiency of right- and left- auditory pathways) by comparing the association between each test’s raw (right and left) and calculated scores (total and difference). Because the distribution of scores violated the assumptions required to interpret linear correlation, we used the Spearman correlation coefficient (\( \rho \)) to quantify the association between rank ordering by each score. The null hypothesis was “no association” for each of the four comparisons (right, left, total, and asymmetry).

Decision consistency of CW and DD classification of interaural asymmetry (right ear advantage present or absent) was evaluated using the kappa coefficient (\( \kappa \)), which compares observed agreement to the amount of chance agreement expected given the
distribution of the two variables compared (Viera & Garrett, 2005). We classified each participant as having right ear advantage on a test if right ear score minus left ear score was greater than zero.

Participants were classified as having dichotic deficits if their scores fell below the published two standard deviation cut-off for age. For CW, the cut-offs were relative to total score and interaural asymmetry. For DD, cut-offs are relative to right ear score and left ear score. Kappa was used to evaluate the agreement between dichotic deficit classifications by CW and DD.

3.4. Results

We enrolled 60 volunteer participants aged 7-14 with normal hearing sensitivity. Characteristics of participants are summarized in Table 3-1. About one-third (n=22) were receiving support in school for developmental, educational, or emotional difficulties; 10 were reported as having Attention Deficit Hyperactivity Disorder. Adult-completed questionnaires of 13 participants endorsed “concerns about hearing, listening, or ability to understand.”

Dichotic test scores are summarized in Table 3-2. Participants had higher proportions of correct responses on DD than CW (p <0.001) and higher right ear than left ear scores on both CW (p < 0.001) and DD (p < 0.001). Mean interaural asymmetry was 1.6 words on CW and 2.2 digits on DD. There was no systematic difference in direction of interaural asymmetry between the two tests (p = 0.40). Having a higher score on the right ear was associated with having a higher left ear score on both CW (ρ = 0.43, p < 0.001) and DD (ρ = 0.52, p < 0.001).
3.4.1. Association between CW and DD scores

Participants were rank ordered similarly (see Figure 3-1) by DD and CW right ear scores ($\rho = 0.58$, $p < 0.0001$) and left ear scores ($\rho = 0.51; p < 0.0001$). The association was even stronger between CW and DD total correct scores. ($\rho = 0.73$, $p < 0.0001$) (see Figure 3-2). The association of rankings by CW and DD inter-aural asymmetry scores (right-left) was weak ($\rho = 0.18$) and not statistically significant ($p = 0.18$) (see Figure 3-2).

3.4.2. Agreement between Right Ear Advantage on CW and DD

Thirty-eight participants had right ear advantage (REA) on CW. DD identified 34 participants with REA. The two tests agreed on the presence of REA for only 31/60 participants (52%), a result easily explained by chance (see Table 3-3; $\kappa= 0.01; p = 0.93$).

3.4.3. Agreement between Normal/Abnormal Classification by CW and DD

DD identified significantly more participants with dichotic deficits (n=18) than CW (n=3) ($p < 0.001$) (see Table 3-4). The two tests agreed on classification of 45/60 participants (75%, $\kappa = 0.22$, $p < 0.01$). Based on the proportion of subjects classified as passing by each test, agreement of 68% is expected by chance. The observed agreement of 75% is only 7% higher than the agreement expected by chance and is therefore classified as “poor agreement.” When there was disagreement between the tests (n=15), participants were always classified as normal by CW but abnormal by DD. Among the children who were classified as having abnormal dichotic listening by both tests (n=3), one had bilaterally low scores on DD, low CW Total, and normal interaural asymmetry; one had bilaterally low scores on DD, low CW Total, and abnormal CW interaural
asymmetry; and one low DD left ear score with abnormally large CW right ear advantage and abnormal CW total score.

3.5. Discussion

Although CW and DD present the listener with similar tasks, these data demonstrate that they cannot be used interchangeably to describe direction of ear advantage or to identify children with abnormal dichotic listening.

Both tests showed an average interaural asymmetry that favored the right ear as expected, but CW and DD had only chance agreement on which individuals had right ear advantage. The association between ranking of participants in the study sample by size and direction of CW and DD asymmetry (right-left) was also indistinguishable from chance. This lack of association may reflect that the two tests are measuring fundamentally different phenomenon, or could be due to homogeneity in the study sample or poor precision of interaural asymmetry estimates. Few of the children in our sample have significant asymmetry and noise in the measurements could obscure small differences. However, if poor precision (i.e., noise) prevents the detection of association between interaural asymmetry measures by CW and DD, it is unclear how audiologists could use direction of ear advantage to cross-check individuals’ results.

In our sample, DD identified six times more children as abnormal than did CW (30% vs 5%). Using published criteria, 25% of participants (n=15) were diagnosed as disordered by DD but not by CW. Because the cut-off criteria reference different scores (DD right and left vs. CW total and difference), it’s not clear whether poor agreement is caused by different sensitivity and specificity or whether the scores are describing
different constructs. However, whatever the source of disagreement, the tests were clearly not interchangeable in this sample.

3.5.1. Limitations

This study is limited to a sample of mostly typically developing volunteers with few participants identified by either test as abnormal. The very narrow range of interaural asymmetry scores (about -5 to+15) observed in this sample may not reveal an association that could be apparent in a clinical sample. Since there is no gold-standard for dichotic listening deficit, we cannot address questions about which test is more accurate or why the sensitivity and specificity of the two tests appears to be different.

3.5.2. Conclusions

Participants scores (right ear, left ear, and total) on Dichotic Digits and Competing Words free-recall tests are associated, but the two tests do not agree on interaural asymmetry estimates or on which children should be identified with dichotic listening deficits. The tests are not interchangeable for clinical use. Additional research is urgently needed to determine if either of the tests are appropriate for identifying children who would benefit from treatment for dichotic listening deficits. Clinicians should be transparent about this uncertainty to patients, families, and referring providers before administering dichotic listening tests.
### 3.6. Tables and Figures

#### Table 3-1 Characteristics of 60 Participants

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Mean or n</th>
<th>(SD) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, mean (SD)</td>
<td>10.0</td>
<td>(1.9)</td>
</tr>
<tr>
<td>Female, No. (%)</td>
<td>31</td>
<td>(52%)</td>
</tr>
<tr>
<td>Right Handed, No. (%)</td>
<td>56</td>
<td>(93%)</td>
</tr>
<tr>
<td>Homeschool, No. (%)</td>
<td>5</td>
<td>(8%)</td>
</tr>
<tr>
<td>Non-Hispanic White (%)</td>
<td>51</td>
<td>(85%)</td>
</tr>
<tr>
<td>Highest Parent Education:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High School, No (%)</td>
<td>2</td>
<td>(3%)</td>
</tr>
<tr>
<td>College, No (%)</td>
<td>22</td>
<td>(37%)</td>
</tr>
<tr>
<td>Graduate, No (%)</td>
<td>36</td>
<td>(60%)</td>
</tr>
<tr>
<td>Annual Household Income,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;$25K, No (%)</td>
<td>4</td>
<td>(7%)</td>
</tr>
<tr>
<td>$25K-$74K, No (%)</td>
<td>23</td>
<td>(38%)</td>
</tr>
<tr>
<td>$75-$99K, No (%)</td>
<td>21</td>
<td>(35%)</td>
</tr>
<tr>
<td>&gt;$99K, No (%)</td>
<td>12</td>
<td>(20%)</td>
</tr>
</tbody>
</table>
Table 3-2 Mean, (SD), and [Range] of 60 Participants’ Competing Words (CW) and Dichotic Digits (DD) Scores

<table>
<thead>
<tr>
<th>Test</th>
<th>Right Ear</th>
<th>Left Ear</th>
<th>Total Score</th>
<th>Interaural Asymmetry</th>
</tr>
</thead>
<tbody>
<tr>
<td>CW</td>
<td>64% (14) [30,90]</td>
<td>55% (19) [0,90]</td>
<td>59% (14) [22,90]</td>
<td>1.7 words (3.7) [-5,10]</td>
</tr>
<tr>
<td>DD</td>
<td>90% (9) [65,100]</td>
<td>84% (15) [28,100]</td>
<td>87% (11) [47,100]</td>
<td>2.2 digits (4.1) [-5,16]</td>
</tr>
</tbody>
</table>

Table 3-3 Right Ear Advantage (REA) Classification by Competing Words and Dichotic Digits tests.

<table>
<thead>
<tr>
<th>Digits REA Absent</th>
<th>Digits REA Present</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CW REA Absent</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>CW REA Present</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>34</td>
</tr>
</tbody>
</table>

Table 3-4 Participants Normal/Abnormal Classification of Dichotic Performance by Competing Words and Dichotic Digits Tests.

<table>
<thead>
<tr>
<th>Words Abnormal</th>
<th>Digits Abnormal</th>
<th>Digits Normal</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words Abnormal</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Words Normal</td>
<td>15</td>
<td>42</td>
<td>57</td>
</tr>
<tr>
<td>Total</td>
<td>18</td>
<td>42</td>
<td>60</td>
</tr>
</tbody>
</table>
Figure 3-1 Association between right and left ear scores on Competing Words and Dichotic Digits tests (n=60)
CW and DD Ear Total and Interaural Asymmetry

**Right + Left Proportion Correct**
\[ \rho = 0.73 \ (p < 0.001) \]

**Right - Left Items Difference**
\[ \rho = 0.18 \ (p = 0.18) \]

Figure 3-2 Associations between Competing Words and Dichotic Digits total scores and Interaural asymmetry.
3.7. References


CONCLUSIONS

This work demonstrated that clinical use of dichotic tests, although recommended in practice guidelines, is not yet well supported by evidence. Although dichotic procedures show moderate reliability, their precision is limited. Improving dichotic test precision will require increasing the number of test items or employing a completely different measurement strategy (e.g., computerized adaptive testing). However, we should establish whether precise knowledge of dichotic listening score could have clinical utility before investing in efforts to refine the tests.

Assessment of dichotic test accuracy is limited by the absence of a widely-accepted gold standard reference test. Prospective studies evaluating whether dichotic listening scores can predict benefit from dichotic intervention are recommended for future research. If benefits are confirmed, results could be used to inform development of dichotic tests sensitive to detecting the characteristics that predict response to treatment. If patients receive no benefit from dichotic intervention, there is no need to further refine dichotic tests.


StataCorp. (2013). *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP.


