Published in final edited form as:

Lang Speech Hear Serv Sch. 2008 January ; 39(1): 44–53. doi:10.1044/0161-1461(2008/005).

Empirically Derived Combinations of Tools and Clinical Cutoffs: An Illustrative Case With a Sample of Culturally/Linguistically Diverse Children

Janna B. Oetting and Lesli H. Cleveland

Louisiana State University, Baton Rouge

Robert F. Cope III

Southeastern Louisiana University, Hammond

Abstract

Purpose—Using a sample of culturally/linguistically diverse children, we present data to illustrate the value of empirically derived combinations of tools and cutoffs for determining eligibility in child language impairment.

Method—Data were from 95 4- and 6-year-olds (40 African American, 55 White; 18 with language impairment, 77 without) who lived in the rural South; they involved primarily scores from the Comprehension subtest of the Stanford-Binet Intelligence Scale—Fourth Edition (CSSB; R. Thorndike, E. Hagen, & J. Sattler, 1986), but scores from an experimental nonword repetition task (NRT; C. Dollaghan & T. Campbell, 1998) were also included as supplements to these scores.

Results—Although the CSSB led to low fail rates in children without impairment and a statistically reliable group difference as a function of the children's clinical status but not their race, only 56% of children with impairment were accurately classified when -1 SD was employed as the cutoff. Diagnostic accuracy improved to 81% when an empirically derived cutoff of $-.5$ SD was used. When scores from the NRT were added to those from the CSSB, diagnostic accuracy increased to 90%.

Implications—This illustrative case adds to the growing number of studies that call for empirically derived combinations of tools and cutoffs as one option within an evidence-based practice framework.

Keywords

assessment; child language disorders; cultural/linguistic diversity

Norm-referenced tests, although not the only tools used within assessment, play an important role in determining who is eligible for clinical services in the field of speech-language pathology. This is especially true in public school settings where established eligibility guidelines in the form of -1 SD, -1.5 SD, or -2 SD are set at the district, parish, or state level. In a recent article, however, Spaulding, Plante, and Farinella (2006) examined the diagnostic accuracy rates of 43 assessment tools, and from their review argued against “wholesale” use of a single low cut-off score to determine eligibility. In this article, we

present data from a sample of culturally/linguistically diverse children to further support and expand on the clinical dialogue that Spaulding et al. initiated.

Our illustrative case focuses primarily on scores from the Comprehension subtest of the Stanford-Binet Intelligence Scale— Fourth Edition (CSSB; Thorndike, Hagen, & Sattler, 1986), although scores from an experimental nonword repetition task (NRT; Dollaghan & Campbell, 1998) are also included as supplements to these scores. Our data for both of these tools were collected in the late 1990s because research from other labs supported their use for children from culturally/linguistically diverse backgrounds. Unfortunately, our initial evaluation of these tools using some very basic methods led to mixed findings that were difficult to reconcile. After reading Spaulding et al.'s (2006) article and learning more about evidence-based practice methods, we re-examined our data using discriminant analyses. These analyses allowed for the identification of an empirically derived cut-off score that improved the classification accuracy of the CSSB. These analyses also showed that we could further improve the classification accuracy of the CSSB by combining it with the children's NRT scores. In the current study, we present these data. Given that the Stanford Binet Intelligence Scale recently underwent a major revision, and there are now multiple versions of the NRT available to clinicians, the goal of this article is not to endorse the CSSB or the specific NRT version we used. Instead, our analysis is offered as an analog to future considerations of tools that show promise for children from a wide range of cultural and linguistic backgrounds, but in isolation lack high levels of diagnostic accuracy with a predetermined $-1\ SD$, $-1.5\ SD$, or $-2\ SD$ cut-off score.¹

DIAGNOSTIC ACCURACY INDICES

Before reviewing Spaulding et al.'s (2006) study and our data, it is important to describe the role of diagnostic accuracy rates for test evaluation within an evidence-based practice framework. Any time a test is given, there are four possible outcomes.

- The test can correctly detect impairment that is present (a true positive result).
- The test can detect impairment when it is really absent (a false positive result).
- The test can correctly identify someone as being free of impairment (a true negative result).
- The test can identify someone as being free of impairment when it is really present (a false negative result).

Table 1 presents a contingency table to illustrate the four possible test outcomes. Also provided are formulas for calculating indices that are often used to evaluate a test's classification accuracy. Two of these indices are sensitivity (Se) and specificity (Sp). The former reflects a test's ability to accurately identify true positives, and the latter reflects a test's ability to accurately identify true negatives. Values of 1.0 on both indices indicate a test with 100% classification accuracy; values of .00 indicate a test with 0% accuracy. As discussed by Plante and Vance (1994), opinion differs as to how high on these indices a

¹The Stanford-Binet Intelligence Scale—Fifth Edition was published by Roid in 2003. Within this newest edition, the CSSB and other subtests of the fourth edition were not retained. As with previous revisions of this tool, the fifth edition involved major changes in both content and scoring when compared to the fourth edition. Content changes included moving to a five-factor as opposed to four-factor model and changing items and subtests to (a) place higher priority on vocabulary as opposed to reasoning skills, (b) balance out the number of receptive versus expressive items, and (c) expand the number of items for testing high giftedness and very low functioning. As reported in the technical manual, the fourth and fifth editions of the tool are highly correlated at .90, with the fourth and fifth editions showing an average 4-point difference in composite scores. That the CSSB is no longer readily available to clinicians does not lessen our enthusiasm for a future language assessment tool that has features similar to those of the CSSB. Nevertheless, the goal of the current work is to use the CSSB data to illustrate the value of empirically derived combinations of tools and cutoffs rather than argue for its use within current clinical practice.

test's values should be to be clinically useful, but most agree that values greater than .90 are good and values between .80 and .90 are fair. When values fall below .80, opinions begin to vary.

Other useful indices of classification accuracy are likelihood ratios. These indices reflect the likelihood that a particular test result would be expected for a child with impairment compared to the likelihood that the same test result would be expected for a child without impairment (Mant, 1999). More details about likelihood ratios are included within our analysis and discussion of the CSSB, but for now, what is important to know is that these ratios help clinicians apply test accuracy data to individual children. Specifically, these indices allow a clinician to ask "Does a positive test result mean that this child is language impaired?" and "Does a negative test result mean that this child is typically developing?" Knowing only the sensitivity and specificity of a test does not allow a clinician to examine the probability that a particular child's score is being interpreted correctly.

Likelihood ratios are interpreted in the following way. Positive likelihood ratios greater than 1 indicate that a test result is more likely to come from a child with impairment, and negative likelihood ratios less than 1 indicate that a test result is more likely to come from a child without impairment. According to Mant (1999), the ideal situation would be for a test to yield a high positive likelihood ratio (e.g., ~10 or more) to "rule in" a diagnosis of impairment and a low negative likelihood ratio (e.g., close to zero) to "rule out" a diagnosis. Positive and negative likelihood ratios such as 10 and .1, for example, would mean that a positive test result would be ten times more likely to come from a child with impairment than from a child without impairment, and a negative test result would be ten times more likely to come from a child without impairment than from a child with impairment.

We now turn to Spaulding et al.'s (2006) test review. Their focus was on sensitivity and specificity, the first two indices just described. From their analyses, they found that only 14 of the 43 tools could be evaluated for these indices, and of these, only 5 yielded sensitivity and specificity rates greater than .80. These included the Clinical Evaluation of Language Fundamentals—Fourth Edition (Semel, Wiig, & Secord, 2003), Preschool Language Scales—Fourth Edition (Zimmerman, Steiner, & Pond, 2002), Test of Early Grammatical Impairment (Rice & Wexler, 2001), Test of Language Competence—Expanded Edition (Wiig & Secord, 1989), and Test of Narrative Language (Gillam & Pearson, 2005). Interestingly, even across these 5 tests, the optimal test cutoffs that led to these levels of classification accuracy were not the same. Some used a cutoff that was less than -1 *SD*, and others used a cutoff that was between -1 *SD* and -2 *SD*.

These findings led Spaulding et al. (2006) to argue against eligibility guidelines that mandate an across-the-board low cut-off score, such as -1 *SD*, -1.5 *SD*, or -2 *SD*. Instead, they recommended the use of a single low cut-off score only when the test being administered has been shown to have good classification accuracy at that cutoff. For other tools with good sensitivity and specificity values, they suggested using the empirically derived cutoffs that the test developers generated to achieve the sensitivity and specificity values. For clinicians, this would mean using different eligibility cutoffs for different tools, but key to the selection of these cutoffs would be the availability of empirical data (within the test manual or elsewhere) to support these cutoffs.

Another finding from Spaulding et al.'s (2006) review that is relevant to the current work came from an examination of the average test score difference between children with and without impairments as reported in standard deviation units. Thirty-three of the 43 tools reported this information in the test manuals, but only 10 showed the average score difference between children with and without impairments to be greater than 1.5 *SD*. The 23

others showed average score differences that were less than 1.5 *SD*, with 9 of these showing a difference that was less than 1 *SD*. This latter group of 9 included the Diagnostic Evaluation of Language Variation (Seymour, Roeper, & de Villiers, 2003), a tool that was developed for children who speak a range of different nonstandard dialects of English. As will be seen, our data show the CSSB (and NRT) to align with these 9 tools. Although Spaulding et al. recommended that these 9 tools and others like them not be used to determine eligibility, our approach offers an alternative that allows for continued consideration of these tools within an evidence-based framework.

CSSB as a Promising Assessment Tool

In the late 1990s, the CSSB showed promise as a clinically useful language assessment tool for children from culturally/ linguistically diverse backgrounds. Examples of questions on this tool include “Why do people use umbrellas?” and “Why do doctors and nurses give people shots?” Although this tool is part of a general intelligence test that is usually administered by psychologists, Peña and Quinn (1997) became interested in it for African American and Latino American children because it requires children to use language to describe and explain information to others, and these types of tasks reflect the types of communication interactions that are thought to occur within these children’s homes. According to Peña and Quinn, this type of task offers an alternative to the types of culturally biased picture-naming and labeling tasks that are often used within the field.

Another appealing feature of the CSSB is that the scoring system is dependent on the content, not on the phonological and morphosyntactic characteristics of a child’s response. Therefore, children who produce nonstandard patterns of English are not penalized. Finally, the scoring procedure of the CSSB differs from that of many traditional tests of language because children’s responses are not initially scored as correct or incorrect. Instead, they are scored as satisfactory, ambiguous, or unsatisfactory, and any response that falls into the ambiguous category is queried. Only after the query are ambiguous responses scored as unsatisfactory. In other words, the query procedure provides a clinician with an additional mechanism by which to familiarize a child to the testing format of the tool.

Peña and Quinn’s (1997) initial study of the CSSB involved two waves of data collection. The first included 50 children and the second included 77. In both, the children’s CSSB scores were compared to those of either the original or revised version of the Expressive One Word Picture Vocabulary Test (EOWPVT; Gardner, 1979, 1990). In both cases, children who were classified as typically developing scored higher on the CSSB than on the EOWPVT. In addition, statistically reliable group differences between children with and without impairments were found for the CSSB, but similar group differences were not found for the EOWPVT. From these results, the authors concluded that the CSSB is a better language assessment tool than the EOWPVT for children from culturally/ linguistically diverse backgrounds. As will be shown, our initial evaluation of the CSSB replicated that of Peña and Quinn, but it also showed the diagnostic accuracy of the tool to be low when a -1 *SD*, -1.5 *SD*, or -2 *SD* clinical cutoff was considered.

NRT as a Promising Assessment Tool

In the late 1990s, the NRT was also being advocated for children from culturally/ linguistically diverse backgrounds. The format of these tasks requires children to listen to nonce words and repeat them. The nonce words do not contain semantic or grammatical information. This is an appealing feature of the tool because variation frequently exists across different dialects of a language in the areas of vocabulary and grammatical surface structure. Another appealing feature of NRTs is that the phonemes used to make up the nonsense words can be selected from a child’s existing phonological repertoire.

Before the data collection phase of the current project, and since that time, findings from a number of studies support the use of NRTs for a wide range of diverse language learners, including children from minority backgrounds and multilingual backgrounds (Bishop, North, & Donlan, 1996; Campbell, Dollaghan, Needleman, & Janosky, 1997; Conti-Ramsden, 2003; Conti-Ramsden, Botting, & Faragher, 2001; Dollaghan & Campbell, 1998; Ellis Weismer et al., 2000; Gathercole & Baddeley, 1990; Laing & Kamhi, 2003; Oetting & Cleveland, 2006; Rodekohr & Haynes, 2001; Thal, Miller, Carlson, & Vega, 2005; but see also Crago & Goswami, 2006; Stokes, Wong, Fletcher, & Leonard, 2006). For example, Ellis Weismer et al.'s study included five-hundred-eighty-one 7- to 8-year-olds (164 with language impairment; 417 without), of which 15% were from a minority racial and/or ethnic group (i.e., African American, Hispanic, Asian, American Indian). Results showed a significant group difference on the NRT between children with and without impairment and a nonsignificant effect for the children's race and/or ethnicity. Interestingly, though, when an empirically derived cutoff of 75% correct was applied to the NRT data, only 40% of the children with impairment and 85% of the children without impairment were classified correctly.

Other studies have shown that the diagnostic accuracy of NRTs also varies as a function of a child's age. For example, using The Children's Test of Nonword Repetition (Gathercole & Baddeley, 1996), Conti-Ramsden and colleagues (2001, 2003) showed NRTs to have relatively fair rates of diagnostic accuracy for 11-year-olds ($Se = .78$; $Sp = .87$) using the 16th percentile (~ -1 SD) from a normative sample. Using this same cutoff with 5-year-olds, however, led to lower classification rates ($Se = .59$; $Sp = 1.00$), although these values increased slightly ($Se = .66$ and $Sp = 1.00$) when the 25th percentile ($\sim -.67$ SD) was employed as the cut-off. In a study examining the NRT data presented in the current article, Oetting and Cleveland (2006) reported similar findings for 6-year-olds. Using the experimental NRT task from Dollaghan and Campbell (1998) and an empirically derived cutoff of 70% correct, this tool yielded sensitivity and specificity rates of .56 and .92, respectively. From the perspective of Spaulding et al.'s test review, all of these levels of classification accuracy are too low for NRTs to be used for eligibility decisions.

PURPOSE OF THE STUDY

The purpose of the current illustrative case study was two-fold. First, we wanted to use data from the CSSB to support Spaulding et al.'s (2006) recommendations to use diagnostic accuracy rates and empirically derived cut-off scores when evaluating and using assessment tools within the field of child language. Second, we wanted to use data from both the CSSB and the NRT to illustrate the added benefit of also considering empirically derived combinations of tools within assessment. The specific questions guiding the study were:

- Does an empirically derived cutoff score lead to a higher level of diagnostic classification accuracy for the CSSB than when an arbitrary cutoff of -1 SD , -1.5 SD , or -2 SD is used?
- Does consideration of empirically derived cut-off scores from the CSSB and NRT in combination lead to a higher level of diagnostic classification accuracy than when scores from the CSSB are considered in isolation?

METHOD

Participants

The participants were ninety-five 4- and 6-year-olds who were solicited for two word learning studies (Horohov & Oetting, 2004; Oetting, 2003). Forty were African American and 55 were White. At the time of the study, all lived in Ascension parish, LA, in areas that

were predominately rural and nonfarm with a variety of chemical/industrial plants located along the banks of the Mississippi River. The 2000 census (U.S. Department of Commerce) reported a total population of 76,627 for the parish, with 20% of the residents listed as African American and 77% listed as White. In some areas along the Mississippi, however, residents listed as African American were reported to be as high as 69%.

All of the 6-year-olds attended public kindergartens where 75% or more of the children qualified for free or reduced lunch. All of the children were also judged to speak a nonstandard dialect of English by at least two examiners. Although other more formal measures of nonmainstream dialect use were not collected, the children were recruited from the same schools as, and 12 months after, those who were studied by Oetting and McDonald (2001, 2002). In these earlier studies, language samples were collected from a different set of 93 children, and within the samples, 35 different nonstandard patterns of English were identified and coded. Some of these patterns included *zero copula and auxiliary be* (e.g., he bad), *zero third person marking* (e.g., she talk to him), *omission of auxiliary does* (e.g., what that say?), and *multiple negation* (e.g., he don't got none). All participants in these earlier studies produced one or more nonstandard English patterns in 3% or more of their utterances (range for African American children = 10%-52%; range for White children = 3%-35%).

Eighteen of the children were classified as 6-year-olds with specific language impairment. The criteria for this classification were (a) currently enrolled in speech-language services in the public schools, (b) designated as exhibiting language skills below his or her peers as determined by the classroom teacher, (c) performed within 1 *SD* of the mean on the Columbia Mental Maturity Scale (CMMS; Burgmeister, Blum, & Lorge, 1972), (d) performed below -1 *SD* of the mean on the Peabody Picture Vocabulary Test—Revised (PPVT-R; Dunn & Dunn, 1981) and on the syntactic quotient of the Test of Language Development—Primary: Second Edition (TOLD-P:2; Newcomer & Hammil, 1988), (e) did not demonstrate frank neurological impairments or social-emotional deficits as documented by school records, and (f) passed a hearing screening within 6 months of the study.

The other children were classified as typically developing, with 40 being 6 years of age and 37 being 4 years of age. Two ages were represented in the typically developing group because some were solicited to serve in the original studies as age controls and others were solicited to serve as younger, language-matched controls. The 6-year-olds were recruited from the kindergarten classrooms of the children with impairment, and the 4-year-olds were recruited from Head Starts, daycares, and preschools that were in close proximity to the kindergartens. To identify these participants, all children who were functioning within the average range per teacher report and without a history of speech or language impairment were given a consent form to take home. Although we do not know the exact number of consent forms that were sent home by the teachers, of the 78 who returned a form, only one 4-year-old was unwilling to participate in the testing battery.

Of the 77 children who completed the battery, 36 presented an age or language profile that matched those of the children who were classified as impaired. These children also earned scores within or above -1 *SD* of the normative average on the PPVT-R, the syntactic quotient of the TOLD-P:2, and the CMMS. Of the others, 25 earned a standard score that was below -1 *SD* on the PPVT-R ($n = 16$), TOLD-P:2 ($n = 16$), and/or CMMS ($n = 6$). For the PPVT-R and TOLD-P:2, these results reflect a 21% fail rate by the typically developing children. This rate is higher than 16%, which is what one would expect based on a normal curve. High fail rates are not inconsistent with previous descriptions of these tools as potentially biased against children from culturally/linguistically diverse backgrounds (for PPVT-R, see Washington & Craig, 1992; for TOLD-P:2, see Hammer, Pennock-Roman,

Rzasa, & Tomblin, 2002; Tomblin et al., 1997, but also Smith, Myers-Jennings, & Coleman, 2000). Participant profiles are provided in Table 2.

Materials

The CSSB is one of five subtests that make up the Verbal Reasoning Composite of the Stanford-Binet Intelligence Scale— Fourth Edition (Thorndike et al., 1986). This subtest begins by showing children a picture of a young boy and asking them to point to six different body parts. Then children are asked questions that require descriptions of objects, actions, and concepts (e.g., “What do people do when they are hungry?”). As testing progresses, the questions become increasingly more abstract and difficult (e.g., “Why are there traffic signs?”). Testing is discontinued after children miss three consecutive questions.

Dollaghan and Campbell’s (1998) NRT includes 16 nonce items that are equally divided into one, two, three, and four syllables. As reported in Oetting and Cleveland (2006), the NRT was administered by recording the 16 nonce words onto the audio track of a videotape. During the audio presentation, the video track presented a solid blue background. The audiotaped items were presented 5 to 7 s apart. Before each item, a flashing white bar was presented in the corner of the blue background to increase the likelihood that the children were attending when the nonce words were presented. Typically, the NRT is presented via audio only (or in live voice with the mouth of the presenter hidden, as in Conti-Ramsden, 2003, and Conti-Ramsden et al., 2001). We chose the audio track of a video monitor because our word learning studies used video monitors and the children were comfortable with the equipment. No visual information other than the blue background and flashing bar was provided to the children during the NRT.

Procedure

The second author or another graduate student in communication disorders administered the CSSB and NRT to each child in a quiet room at each child’s school as part of the testing battery of the word learning studies. For the CSSB, standard scores were calculated using the test manual and then multiplied by 2. With this method, the normative mean for the CSSB is 100 and the *SD* is 16. For the NRT, children were asked to listen to each word and repeat it. Their responses were audio recorded for later transcription and scoring. Following Dollaghan and Campbell (1998), each child’s percentage of phonemes correct on the task was calculated. Phoneme omissions and substitutions (but not distortions) were marked as errors. Phoneme additions were transcribed but not scored. Interrater reliability for the NRT was 96%.

RESULTS

Preliminary Analyses

Preliminary analyses were completed to determine if the CSSB scores differed as a function of the children’s race (African American vs. White) and clinical status (with vs. without impairment). This was important to do for at least three reasons. First, from a statistical standpoint, a difference between children with and without impairment is a necessary precursor to running the types of discriminant analyses that allow for an examination of a test’s classification accuracy (Dillon & Goldstein, 1984). Second, these analyses were needed to show that our data replicated those of Peña and Quinn (1997). Recall that in their earlier study, the CSSB led to group differences that were related to the children’s clinical status but not their race and/or ethnicity. If our CSSB scores varied as a function of the children’s race, then there would be no reason to examine the tool further. Third, we argued earlier that our analysis of the CSSB can be used as an analog for future considerations of

other tools. In order for this to be true, we needed to demonstrate that our CSSB results were similar to those of other tools that were currently available to clinicians.

CSSB means for the individual groups are reported in Table 3. To examine these data, a 2×2 analysis of variance was completed with race (African American vs. White) and clinical status (with impairment vs. without) as the independent variables. This analysis showed a statistically reliable difference between the children with and without impairment, $F(3, 91) = 23.62$, $p < .001$, $\eta_p^2 = .21$. Neither the main effect for race nor the interaction between race and clinical status was significant, $p > .39$. The difference between the average CSSB scores of the children with and without impairment, while statistically significant, was slightly less than 1 *SD* of the normative sample for this tool. As mentioned earlier, nine of the tests reviewed by Spaulding et al. also reported an average group difference of less than 1 *SD* within their manuals. Together, these findings justify the next set of analyses and show the CSSB to be similar to a number of existing assessment tools.

Group means for the NRT can also be found in Table 3. As reported in Oetting and Cleveland (2006), a similar main effect for clinical status, but not race, was found for this tool. These findings are consistent with the CSSB results, and they also replicate those of Ellis Weismer et al. (2000). In Oetting and Cleveland, we also searched the children's NRT responses for the presence of nonstandard English phonology and examined all production errors as a function of the children's race and clinical status. Although some nonstandard productions (e.g., monophthongal variants of diphthongs, syllable deletion of unstressed initial and medial syllables, and final consonant devoicing of /b/ and /g/) were identified within the data, occurrence of these productions was infrequent (<2%). In addition, the error analysis showed the NRT results to not differ as a function of the children's race or clinical status. Together, these findings justify further consideration of the NRT as a supplement to the CSSB scores.

Classification Accuracy: Arbitrary Versus Derived Cutoffs

The classification accuracy of the CSSB was next evaluated using discriminant analysis. Discriminant analysis is a statistical technique for classifying individuals or objects into mutually exclusive and exhaustive groups on the basis of one or more independent variables (Dillon & Goldstein, 1984). The a priori groups included the 6-year-olds with and without impairment. The 4-year-olds were excluded from this analysis because clinical decisions typically are based on how children perform as a function of their age. The first analysis used -1 *SD* (or a standard score of 84) on the CSSB as the arbitrary cut-off score. When this cutoff was used, only one of the typically developing 6-year-olds earned a score that was below -1 *SD* of the normative mean. This <1% fail rate is considerably lower than the 21% fail rate that was observed for the PPVT-R and TOLD: P2. However, 10 of the children (5 African American and 5 White) with impairment were classified as typically developing. This number reflected a 56% false negative rate. Sensitivity was low at .44 but specificity was high at .98. Pushing the CSSB cutoff even lower to -1.5 *SD* and -2 *SD* further reduced the sensitivity values to .11 and .06, although specificity increased to 1.00 (or 100%) in both cases. In other words, the lower cutoffs led to the misclassification of almost all of the children with impairment even though all of the children without impairment scored above these cutoffs.

Next, we ran the discriminant analysis without specifying the clinical cutoff of the CSSB. With this method, an empirically derived cut-off score that maximized the classification accuracy of the CSSB was generated from the analyses. The resulting discriminant function yielded a CSSB standard score of 92, which reflected $-.5$ *SD*, as the optimal clinical cutoff, and with this cutoff, the CSSB yielded an overall classification accuracy rate of 81%; sensitivity and specificity were .72 and .93, respectively. Children who were misclassified

included 5 with impairment and 3 without impairment. When this model is compared to the first, one sees an increase in sensitivity from .44 to .72 and a corresponding decrease in the false negative rate from 56% to 28%. Improvements in these values, without dramatic changes to the tool's specificity value, indicate that the clinical utility of the CSSB could be improved by using the empirically derived cut-off score of $-.5\ SD$.

Classification Accuracy: Single Tools Versus Combined Tools

Discriminant analysis was then used to determine whether the diagnostic accuracy rate of the CSSB could be further improved by combining it with the children's NRT scores. As reported in Oetting and Cleveland (2006), 2 children with impairment and 4 without were unable to complete the NRT due to their unwillingness to attempt the task. Thus, the current analyses were completed with data from 52 participants. Like before, this analysis allowed the discriminant function to determine an empirically derived cutoff that maximized classification of the children with and without impairments. The results are illustrated in Figure 1. As can be seen, the resulting discriminant function (i.e., D score) is a linear equation that requires each child's score on the CSSB and NRT to be weighted and added together. Within this analysis, the empirically derived cutoff was a D score of $-.87$. This D score corresponded to a CSSB standard score of 91 ($-.56\ SD$) and an NRT score of 71% correct, although with D scores, a child's score on one tool can be slightly higher or lower than these values as long as the score on the other tool increases or decreases accordingly. Importantly, when these two tools were considered together and an empirically derived D score of $-.87$ was used as the clinical cutoff, 90% of the children were accurately classified. The false negative rate was 19%, sensitivity was .81, and specificity was .94. From Spaulding et al.'s (2006) perspective, these levels of diagnostic accuracy are high enough to be used for eligibility decisions.

Moreover, when compared to the results of the other discriminant analyses (including one with only the children's NRT scores and another with the CSSB scores from only the 52 children with NRT data), the current discriminant function led to the highest level of diagnostic accuracy, with only 3 children with impairment and 2 without misclassified. The CSSB and NRT were also only moderately correlated to each other ($r = .58$), and within the discriminant function, their pooled within-group correlation was even lower (.28). This low correlation indicates that the children's scores on each tool were providing unique information to the discriminant function.

Likelihood Ratios

As mentioned earlier, likelihood ratios are indices that can be used by clinicians to apply test classification data to individual children. We now turn to the application of these ratios to the current data set. Table 4 lists the sensitivity, specificity, and likelihood ratios that were calculated from all discriminant analyses. As can be seen, improvements in the CSSB's likelihood ratios improved along with changes that were made to the tool's sensitivity and specificity values. When the cutoff of $-1\ SD$ was employed, the positive likelihood ratio was 22 and the negative ratio was .57. These likelihood ratios changed to 10 and .30 when an empirically derived cutoff of $-.5\ SD$ for the CSSB was used, and they changed again to 14 and .20 when the children's CSSB scores were combined with their NRT scores. As can also be seen, other sets of likelihood ratios that were calculated when the CSSB and NRT were used in isolation were lower than these.

Using Mant's (1999) recommendations of a positive likelihood ratio of 10 to "rule in" a diagnosis and a .10 to "rule out" a diagnosis, clinicians can interpret these likelihood ratios in the following way. If a child scored below $-1\ SD$ on the CSSB, clinicians would have sufficient evidence to rule in a diagnosis of impairment because the positive likelihood ratio

at this cutoff is above 10. On the other hand, if this same child scored above -1 *SD* on the CSSB, clinicians would be unable to rule out a diagnosis because the negative likelihood ratio of this tool at this cutoff is not as low as it should be. In fact, a .57 negative likelihood ratio means that children without impairment are only about twice as likely as children with impairment to earn a score above this cutoff. Adjusting the cutoff to $-.5$ *SD* on the CSSB improves clinicians' ability to rule out a diagnosis without significantly reducing their ability to rule in a diagnosis; however, if clinicians consider a child's CSSB and NRT scores together and use a *D* score of $-.87$ as the cutoff, they can be even more confident in their eligibility decisions. Recall that this combination of tools with this cutoff generated positive and negative likelihood ratios of 14 and .20. These likelihood ratios mean that *D* scores below $-.87$ are 14 times more likely to come from children with impairment, and those above this cutoff are five times more likely to come from children without impairment.

DISCUSSION

We began this article by reviewing Spaulding et al.'s (2006) concerns with current eligibility guidelines and highlighting their use of classification accuracy rates to examine tools for clinical practice. As part of their review, they also identified a small number of tools with adequate classification rates and many others with unavailable or unacceptable rates. For those with adequate rates, Spaulding et al. recommended the use of empirically derived cutoffs that are test specific rather than a single low cut-off score. For those without adequate rates, they recommended that these tools not be used to make eligibility decisions. We then presented data from the CSSB (and NRT) to both support and expand on Spaulding et al.'s recommendations. Below, we summarize the four main findings of the study.

The first finding was that the CSSB, when used with a -1 *SD* cutoff, presented limited classification accuracy because 56% of the children with impairment were incorrectly classified as typically developing. This was unfortunate because preliminary analysis of the CSSB showed statistically reliable group differences as a function of the children's clinical status without score differences being tied to the children's race. Had we not considered the classification accuracy of the CSSB, we may have incorrectly viewed this tool (and others like it) as being clinically useful with a -1 *SD* cutoff for children from culturally/linguistically diverse backgrounds.

The second finding was that the classification accuracy of the CSSB could be greatly improved by using an empirically derived cutoff score. In fact, the classification accuracy of the tool went from 69% to 81% just by moving the cutoff to $-.5$ *SD*. Both of these findings support Spaulding et al.'s test review paper. The first shows how important it is for clinicians working with culturally/linguistically diverse children to consider classification accuracy rates when evaluating tests, and the second illustrates the value of using empirically derived cutoffs to improve assessment practices for these children.

The third finding of the study was that the classification accuracy of the CSSB could be improved further by combining it with the children's NRT scores. When we did this, classification accuracy went from 81% to 90%. This classification rate is good, and it is also better than what was found for most of the tools that were included within Spaulding et al.'s review. Recall that before these statistical manipulations, the classification accuracy of the CSSB was extremely low and the average group score difference between the children with and without impairment was slightly less than 1 *SD* of the normative mean. These findings aligned the CSSB with nine other tools that Spaulding et al. reviewed, including the Diagnostic Evaluation of Language Variation (Seymour et al., 2003), a tool that has been developed for children who speak a range of nonstandard English dialects. Whereas Spaulding et al. recommended that these nine tools and others like them not be used to

determine eligibility, our approach offers an alternative that allows for additional consideration of these tools. This alternative involves the development and validation of empirically derived combinations of tools and cutoffs. For clinicians who work with culturally/linguistically diverse children, this is an important option because the number of available tests for children from diverse backgrounds is limited.

The fourth major finding of the study related to the likelihood ratios. Spaulding et al. (2006) did not include likelihood ratios within their test review, but these indices are generated from the same data that are used to calculate sensitivity and specificity values. As was shown with the CSSB (and NRT) data, likelihood ratios improve as the classification accuracy of a tool or combination of tools improves. Given this, clinicians can have greater confidence in their diagnoses if the tools they use present with high levels of classification accuracy. However, when an optimal testing battery is not available (as in the case of assessments with culturally/linguistically diverse children), likelihood ratios provide an evidence-based means by which clinicians can augment a test score(s) with information about the odds at which that score(s) supports a diagnosis of impairment. Obviously, the long-term goal of the field is to have optimal assessment batteries for all children, but until this becomes a reality, likelihood ratios help clinicians talk about the uncertainty that exists within testing.

Although our data support the use of classification accuracy rates and show the value of developing empirically derived combinations of tools and cutoffs within the field, it is important to point out that it is not always the case that more tools lead to more accurate diagnoses. Gray, Plante, Vance, and Henrichsen (1999) demonstrated this in their evaluation of four vocabulary tests for standard English-speaking preschoolers. In their study, all four vocabulary tests showed higher sensitivity and specificity values when used in isolation than when combined. Interestingly, though, like the current illustrative case, empirically derived cutoffs led to higher sensitivity and specificity values than traditional $-1\ SD$, $-1.5\ SD$, and $-2\ SD$ cutoffs. In addition, all of their empirically derived cutoffs were also above the $-1\ SD$ level. In fact, the optimal cutoff for the Peabody Picture Vocabulary Test—Third Edition (Dunn & Dunn, 1997) was $+.26\ SD$, and for the Expressive Vocabulary Test (Williams, 1997), it was $-.20\ SD$. These findings and ours underscore the importance of rigorously testing the usefulness of different combinations of tools and cutoffs instead of blindly believing that diagnostic accuracy will always improve with the administration of more tools.

With this caveat noted, we close by citing a few other studies within the field that have explored empirically derived testing batteries. These include works by Catts, Fey, Tomblin, and Zhang (2002); Craig and Washington (2000); Klee, Pearce, and Carson (2000); Peña et al. (2006); Peña, Iglesias, and Lidz (2001); and Perona, Plante, and Vance (2005). In most, if not all, of these studies, classification accuracy rates have been highest when tools are combined and cutoffs are derived. Together, the current illustrative case and these studies call for the field to begin using empirically derived cutoffs when they are available and to pursue empirically derived combinations of tools and cutoffs as one option within an evidence-based framework.

Acknowledgments

The CSSB was initially examined as part of the second author's master's thesis. The thesis and analyses of diagnostic accuracy were made possible by an LEQSF grant from the Louisiana Board of Regents and an RO3 grant from the National Institute on Deafness and Other Communication Disorders (DC03609) awarded to the first author. Appreciation is extended to the teachers, parents, and children who participated in the research, and to Lesley Eyles, Anita Hall, and Karen Lynch for help with data collection and coding.

REFERENCES

- Bishop DVM, North T, Donlan C. Nonword repetition as a behavioural marker for inherited language impairment: Evidence from a twin study. *Journal of Child Psychology and Psychiatry*. 1996; 37:391–403. [PubMed: 8735439]
- Burgmeister, B.; Blum, H.; Lorge, I. *Columbia Mental Maturity Scale*. The Psychological Corporation; San Antonio, TX: 1972.
- Campbell T, Dollaghan C, Needleman H, Janosky J. Reducing bias in language assessment: Processing-dependent measures. *Journal of Speech, Language, and Hearing Research*. 1997; 40:519–525.
- Catts HW, Fey ME, Tomblin JB, Zhang X. A longitudinal investigation of reading outcomes in children with language impairments. *Journal of Speech, Language, and Hearing Research*. 2002; 45:1142–1157.
- Conti-Ramsden G. Processing and linguistic markers in young children with specific language impairment. *Journal of Speech, Language, and Hearing Research*. 2003; 46:1029–1037.
- Conti-Ramsden G, Botting N, Faragher B. Psycholinguistic markers for specific language impairment. *Journal of Child Psychology and Psychiatry*. 2001; 42:741–748. [PubMed: 11583246]
- Crago M, Goswami U. Psychological and linguistic studies across languages and learners. *Applied Psycholinguistics*. 2006; 27(4):513–613.
- Craig H, Washington J. An assessment battery for identifying language impairments in African American children. *Journal of Speech, Language, and Hearing Research*. 2000; 43:366–379.
- Dillon, WR.; Goldstein, M. *Multivariate analyses: Methods and applications*. John Wiley & Sons; New York: 1984.
- Dollaghan C, Campbell T. Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*. 1998; 41:1136–1146.
- Dunn, L.; Dunn, L. *Peabody Picture Vocabulary Test—Revised*. AGS; Circle Pines, MN: 1981.
- Dunn, L.; Dunn, L. *Peabody Picture Vocabulary Test—Third Edition*. AGS; Circle Pines, MN: 1997.
- Ellis Weismer S, Tomblin J, Zhang X, Buckwalter P, Chynoweth J, Jones M. Nonword repetition performance in school-age children with and without language impairment. *Journal of Speech, Language, and Hearing Research*. 2000; 43:865–878.
- Gardner, M. *Expressive One-Word Picture Vocabulary Test*. Academic Therapy; Novato, CA: 1979.
- Gardner, M. *Expressive One-Word Picture Vocabulary Test—Revised*. Academic Therapy; Novato, CA: 1990.
- Gathercole S, Baddeley AD. Phonological memory deficits in language disordered children: Is there a causal connection? *Journal of Memory and Language*. 1990; 29:336–360.
- Gathercole, S.; Baddeley, AD. *The Children's Test of Non-word Repetition*. The Psychological Corporation; London: 1996.
- Gillam, R.; Pearson, N. *Test of Narrative Language*. Pro-Ed; Austin, TX: 2005.
- Gray S, Plante E, Vance R, Henrichsen M. The diagnostic accuracy of four vocabulary tests administered to preschool-age children. *Language, Speech, and Hearing Services in Schools*. 1999; 30:196–206.
- Hammer CS, Pennock-Roman M, Rzasa S, Tomblin JB. An analysis of the Test of Language Development—Primary for item bias. *American Journal of Speech-Language Pathology*. 2002; 11:274–284.
- Horohov J, Oetting J. Effects of input manipulations on the word learning abilities of children with and without specific language impairment. *Applied Psycholinguistics*. 2004; 25:43–67.
- Klee T, Pearce K, Carson D. Improving the positive predictive value of screening for developmental language. *Journal of Speech, Language, and Hearing Research*. 2000; 43:821–833.
- Laing S, Kamhi A. Alternative assessment of language and literacy in culturally and linguistically diverse populations. *Language, Speech, and Hearing Services in Schools*. 2003; 34:44–55.
- Mant, J. Studies assessing diagnostic tests. In: Dawes, M.; Davies, P.; Gray, A.; Mant, J.; Seers, K.; Snowball, R., editors. *Evidence-based practice: A primer for health care professionals*. Churchill Livingstone; New York: 1999. p. 67-78.

- Newcomer, P.; Hammil, D. Test of Language Development— Primary: Second Edition. Pro-Ed; Austin, TX: 1988.
- Oetting, JB. Children's use of prepositions to learn verbs. Louisiana State University; Baton Rouge: 2003. Unpublished manuscript
- Oetting JB, Cleveland LH. The clinical utility of nonword repetition for children living in the rural south of the US. *Clinical Linguistics and Phonetics*. 2006; 20:553–561. [PubMed: 17056486]
- Oetting JB, McDonald J. Nonmainstream dialect use and specific language impairment. *Journal of Speech, Language, and Hearing Research*. 2001; 44:207–223.
- Oetting JB, McDonald J. Methods for characterizing participants' nonmainstream dialect use within studies of child language. *Journal of Speech, Language, and Hearing Research*. 2002; 45:505–518.
- Peña E, Gillam RB, Malek M, Ruiz-Felter R, Resediz M, Fiestas C, Sabel T. Dynamic assessment of school-age children's narrative ability: An investigation of reliability and validity. *Journal of Speech, Language, and Hearing Research*. 2006; 49:1037–1057.
- Peña E, Iglesias A, Lidz C. Reducing test bias through dynamic assessment of children's word learning ability. *American Journal of Speech-Language Pathology*. 2001; 10:138–154.
- Peña E, Quinn R. Task familiarity: Effects on the test performance of Puerto Rican and African American children. *Language, Speech, and Hearing Services in Schools*. 1997; 28:323–332.
- Perona K, Plante E, Vance R. Diagnostic accuracy of the Structured Photographic Expressive Language Test: Third Edition. *Language, Speech, and Hearing Services in Schools*. 2005; 36:103–115.
- Plante E, Vance R. Diagnostic accuracy of two tests of preschool language. *American Journal of Speech-Language Pathology*. 1994; 4:70–76.
- Rice, M.; Wexler, K. Rice/Wexler Test of Early Grammatical Impairment. The Psychological Corporation; San Antonio, TX: 2001.
- Rodekohr R, Haynes W. Differentiating dialect from disorder: A comparison of two processing tasks and a standardized language test. *Journal of Communication Disorders*. 2001; 34:1–18.
- Roid, G. Stanford-Binet Intelligence Scale—Fifth Edition. Riverside; Chicago: 2003.
- Semel, E.; Wiig, EH.; Secord, WA. Clinical Evaluation of Language Fundamentals—Fourth Edition. The Psychological Corporation; San Antonio, TX: 2003.
- Seymour, HN.; Roeper, T.; de Villiers, JG. Diagnostic evaluation of language variation: Criterion-referenced. The Psychological Corporation; San Antonio, TX: 2003.
- Smith T, Myers-Jennings C, Coleman T. Assessment of language skills in rural preschool children. *Communication Disorders Quarterly*. 2000; 21:114–124.
- Spaulding TJ, Plante E, Farinella KA. Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools*. 2006; 37:61–72.
- Stokes SF, Wong A, Fletcher P, Leonard LB. Nonword repetition and sentence repetition as clinical markers of specific language impairment: The case of Cantonese. *Journal of Speech, Language, and Hearing Research*. 2006; 49:219–236.
- Thal DJ, Miller S, Carlson J, Vega MM. Nonword repetition and language development in 4-year-old children with and without a history of early language delay. *Journal of Speech, Language, and Hearing Research*. 2005; 48:1481–1495.
- Thorndike, R.; Hagen, E.; Sattler, J. Stanford-Binet Intelligence Scale—Fourth Edition. Riverside; Chicago: 1986.
- Tomblin JB, Records N, Buckwalter P, Zhang X, Smith E, O'Brien M. Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research*. 1997; 40:1245–1260.
- U.S. Department of Commerce Bureau of the Census. Statistical Abstract of the United States. 110th ed.. Author; Washington, DC: 2000.
- Washington J, Craig H. Performances on low-income African American preschool and kindergarten children on the Peabody Picture Vocabulary Test—Revised. *Language, Speech, and Hearing Services in Schools*. 1992; 23:329–333.
- Wiig, EH.; Secord, WA. Test of Language Competence— Expanded Edition. The Psychological Corporation; San Antonio, TX: 1989.

- Williams, KT. Expressive Vocabulary Test. AGS; Circle Pines, MN: 1997.
- Zimmerman, IL.; Steiner, VG.; Pond, RE. Preschool Language Scales—Fourth Edition. The Psychological Corporation; San Antonio, TX: 2002.

Figure 1.

Results from discriminant analysis with the CSSB and NRT.^a

^aUnstandardized model: $D_{\text{score}} = -10.273 + .055 (\text{CSSB score}) + 6.156 (\text{NRT score})$.

^bReflects average CSSB standard score at the group centroids.

^cReflects average percentage of phonemes correct on the NRT at the group centroids.

Table 1

Contingency table and formulas for calculating indices of diagnostic accuracy.

	<i>True cases of children with impairment</i>	<i>True cases of children who are typically developing</i>	<i>Total</i>
Positive test result	True positive a	False positive b	a+b
Negative test result	False negative c	True negative d	c+d
Total	a+c	b+d	a+b+c+d

Note. Sensitivity (Se) = $a/(a + c)$; specificity (Sp) = $d/(d + b)$; likelihood ratio for a positive test result = $Se/(1 - Sp)$; likelihood ratio for a negative test result = $(1 - Se)/Sp$.

Table 2

Description of study participants

	<i>Children with LI</i>		<i>Typically developing children</i>			
	<i>6-year-olds</i>		<i>6-year-olds</i>		<i>4-year-olds</i>	
	AA	W	AA	W	AA	W
<i>N</i>	8	10	14	26	18	19
Age in months	76.37 (5.42)	73.30 (6.34)	73.50 (6.07)	71.26 (4.80)	56.11 (6.45)	55.00 (7.33)
CMMS	92.50 (5.74)	92.60 (3.95)	101.43 (11.52)	103.00 (8.96)	92.80 (8.93)	100.95 (14.73)
PPVT-R	70.50 (4.9)	77.10 (5.92)	88.57 (9.68)	101.77 (13.71)	84.27 (12.43)	95.21 (10.96)
TOLD-P:2	74.38 (6.52)	71.80 (10.79)	92.35 (5.78)	99.73 (12.12)	90.11 (9.65)	94.68 (9.83)

Note. AA = African American; W = White; CMMS = Columbia Mental Maturity Scale (Burgmeister, Blum, & Lorge, 1972) standard score, $M = 100$, $SD = 15$; PPVT-R = Peabody Picture Vocabulary Test—Revised (Dunn & Dunn, 1981) standard score, $M = 100$, $SD = 15$; TOLD-P:2 = Test of Language Development—Primary: Second Edition (Newcomer & Hammill, 1988) syntactic quotient, $M = 100$, $SD = 15$.

Comprehension subtest of the Stanford-Binet Intelligence Scale—Fourth Edition (CSSB; Thorndike, Hagen, & Sattler, 1986) and nonword repetition task (NRT; Dollaghan & Campbell, 1998) scores by group and race.

Table 3

	Children with SLI		Typically developing children			
	6-year-olds		6-year-olds		4-year-olds	
	AA	W	AA	W	AA	W
CSSB ^a	87.25 (16.21)	84.80 (4.92)	102.43 (11.45)	108.00 (13.97)	101.37 (9.45)	93.33 (9.26)
NRT ^b	65.83 (8.84)	65.85 (12.82)	77.45 (9.90)	84.07 (7.14)	73.02 (8.75)	78.99 (10.11)

^aReported as standard scores that take into consideration a child's age. Normative mean of CSSB is 100 and SD is 16.

^bReported as percentage of phonemes correct. Scores are not standardized and do not take into consideration a child's age. Maximum score is 100%.

Table 4

Classification accuracy indices using different cutoffs and combinations of tools.

	Accuracy	Se	Sp	Positive Likelihood ratio	Negative Likelihood ratio
CSSB with -1 SD cutoff	69%	.44	.98	22	.57
CSSB with -1.5 SD cutoff	83%	.11	1.00	-	.89
CSSB with -2 SD cutoff	81%	.06	1.00	-	.95
CSSB with empirically derived -1.5 SD cutoff	81%	.72	.93	10	.30
NRT with empirically derived 70% cutoff	81%	.56	.92	7	.48
CSSB and NRT with empirically derived cutoff of $D = -.87$ (CSSB = 91; NRT = 71%)	90%	.81	.94	14	.20