JSLHR

Article

Bilingual Language Assessment: A Meta-Analysis of Diagnostic Accuracy

Christine A. Dollaghan[a] and Elizabeth A. Horner[a]

Purpose: To describe quality indicators for appraising studies of diagnostic accuracy and to report a meta-analysis of measures for diagnosing language impairment (LI) in bilingual Spanish–English U.S. children.

Method: The authors searched electronically and by hand to locate peer-reviewed English-language publications meeting inclusion criteria; the authors rated quality features, calculated accuracy metrics and confidence intervals, and generated forest plots.

Results: Of 771 citations (86 unique) located initially, accuracy metrics could be calculated for 17 index measures studied in a total of 100 children with LI and 109 with typical language. Most studies lacked clear descriptions of reference standards, procedures, and controls for subjective bias, making it difficult to rate specific quality features with confidence. Positive likelihood ratios (LR+) for most measures were at least diagnostically suggestive (pooled LR+ = 4.12; 95% CI [2.94, 5.78]). Negative likelihood ratios (LR–) were also generally suggestive, but heterogeneity precluded averaging. For every measure, confidence intervals for LR+ and LR– included diagnostically uninformative values.

Conclusions: The available evidence does not support strong claims concerning the diagnostic accuracy of these measures, but a number appear promising. Several steps are suggested for strengthening future investigations of diagnostic accuracy.

Key Words: bilingual speakers, evidence-based practice, language disorders

Distinguishing between people who have a disorder and people who do not is one of the most fundamental tasks in clinical practice. Accurate diagnosis is a prerequisite to ensuring that costly treatment resources are allocated to all and only those likely to benefit. Despite a growing number of studies of diagnostic accuracy in recent years, to date no meta-analysis of diagnostic accuracy has appeared in the communication disorders literature. Communication disorders is not unusual in this regard; even in the medical literature, meta-analyses of diagnostic accuracy are infrequent. For example, a July 2009 search of The Cochrane Library (www.thecochranelibrary.com) yielded 12 reviews of diagnostic accuracy, as compared with 1,989 reviews concerning clinical trials.

The purpose of the present article was twofold. First, we define studies of diagnostic accuracy and describe three issues that are specific to appraising and quantifying their results. Second, we report a meta-analysis of the diagnostic accuracy of measures for diagnosing language impairment (LI) in bilingual Spanish–English U.S. children. We selected this domain not because we anticipated finding conclusive evidence favoring a single, optimal diagnostic tool for this purpose. Rather, our intent was to illustrate how a meta-analysis can clarify the state of the diagnostic literature, identifying measures that appear promising and pinpointing weaknesses in the evidence to be remedied so that more powerful meta-analyses are possible in the future.

Studies of Diagnostic Accuracy

As described by many (e.g., Battaglia et al., 2002; Bossuyt et al., 2003; Knottnerus & van Weel, 2002; Sackett, Straus, Richardson, Rosenberg, & Haynes, 2000), a study of diagnostic accuracy compares the classification decisions of two indicators, measures, or protocols that are intended to assign individuals to one of two mutually exclusive categories. For clinical purposes, the most common classification questions concern screening, in which individuals are categorized according to whether

[a]University of Texas at Dallas

Correspondence to Christine A. Dollaghan: dollaghan@utdallas.edu

Editor: Janna Oetting
Associate Editor: Ron Gillam

Received April 8, 2010
Revision received September 24, 2010
Accepted November 10, 2010
DOI: 10.1044/1092-4388(2010/10-0093)

they appear to require further evaluation, and diagnosis, in which individuals are categorized as affected (having a disorder) or as unaffected (not having the disorder). Any study of classification accuracy must begin with a reference standard reflecting a well-supported or widely agreed approach to the classification task. A reference standard may consist of an individual indicator, test, or measure or a specified combination of these (e.g., Meehl, 1995); what is important is that the standard be known or widely agreed to classify individuals with high accuracy. In a study of diagnostic accuracy, the reference standard defines each participant's status as affected or unaffected by the disorder of interest.

The other indicator in a study of diagnostic accuracy is known as the index or comparator measure; it is the accuracy of the index measure that is under investigation. Specifically, each person's diagnostic status (affected or unaffected) according to the index measure is compared with his or her diagnostic status based on the reference standard; disagreements are counted as classification errors by the index measure. The sine qua non for a study of diagnostic accuracy, accordingly, is an exact count of the number of individuals whose category status (affected/unaffected) is correctly assigned by the index measure. Investigations that merely contrast mean scores from groups of affected and unaffected individuals on a proposed diagnostic indicator, or show correlations between the proposed indicator and another measure, do not meet this standard and are better thought of as "preaccuracy" studies. Although preaccuracy studies can provide a rationale for future investigations of a new measure's diagnostic accuracy, only a study comparing each person's diagnostic status on the index test with his or her status on the reference standard counts as a study of diagnostic accuracy.

It is worth noting that many assessment tools are not designed to classify or diagnose, but rather to obtain information for planning or monitoring treatment for patients already diagnosed. Studying the effectiveness of an assessment measure for these purposes would require a randomized clinical trial comparing outcomes for patients whose treatment had been based on the measure and patients whose treatment had not (e.g., Knottnerus & van Weel, 2002). Such studies of evaluation tools are important in their own right, but they do not concern diagnostic accuracy and will not be considered further here.

Special Considerations for Appraising Studies of Diagnostic Accuracy

Although some of the criteria used in critical appraisal of treatment studies also apply to studies of diagnostic accuracy, three issues are unique to the latter: research design and sample selection, choice of a reference standard, and effect size metrics. Each is described in turn below.

Research design and sample selection. In studies of treatment, the research design known as a *parallel-groups randomized controlled trial* is valued highly, because randomly assigning participants to groups reduces the potential for preexisting group differences to account for apparent treatment effects. In studies of diagnostic accuracy, random assignment is impossible because participants' status (affected or unaffected by the disorder) generally cannot be manipulated. Because studies of diagnostic accuracy necessarily involve nonrandomized designs in which participants' diagnostic status is fixed before the study begins, the process by which affected and unaffected individuals are selected for inclusion warrants special scrutiny. The two general approaches to participant selection are known as *one-gate* and *two-gate* study designs (Bossuyt & Leeflang, 2008).

In a one-gate design, the study sample is, in a sense, intended to be unselected, comprising a large, broad, and representative swath of individuals that will presumably include some with and some without the disorder. A study in which the reference standard and the index test are administered to every child in a school district, or a study in which both measures are administered to a consecutive series of 100 people presenting at a speech and hearing clinic, are examples of one-gate designs. Diagnostic accuracy studies with one-gate designs are somewhat analogous to cohort studies, although participants in diagnostic studies usually are not followed forward in time.

The other way to select participants for a study of diagnostic accuracy, known as a *two-gate design,* is analogous to what is called the *case-control design* in studies of treatment. In a two-gate study, the investigator actively selects one group of people with the target disorder (cases) and another group free of the disorder (controls). For example, if an investigator recruited cases by clinician referral and controls by advertising for people with no history of the disorder, the study would have a two-gate design.

Two-gate designs are very common in diagnostic studies, but they are generally valued less highly than one-gate studies because of their greater susceptibility to the validity threat known as *spectrum bias* (see Leeflang, Bossuyt, & Irwig, 2009; Reitsma et al., 2009). Spectrum bias is a concern when diagnostic accuracy is calculated from a sample of participants who do not represent the full spectrum of characteristics (e.g., age, severity, clinical history) that would be encountered when the index measure is used in a real-world clinical context. For example, an index test is likely to be more accurate in separating severely affected from clearly unaffected people than in classifying those with borderline performance. If a two-gate study involves participants who represent the extreme ends of the affected-to-unaffected continuum, then diagnostic accuracy of the index test may be inflated relative to its accuracy in a one-gate study with a more representative sample (Battaglia et al., 2002). Spectrum bias must also be considered in appraising one-gate studies because the accuracy of an index measure in one site or sample may not generalize to samples with different mixtures of participant ages, severities, comorbidities, and the like (see Battaglia et al., 2002; Irwig, Bossuyt, Glasziou, Gatsonis, & Lijmer, 2002; Leeflang et al., 2009).

Differential verification or *ascertainment bias* is another special concern with respect to sample selection in diagnostic accuracy studies; this bias can occur if different reference standards are used to establish diagnostic status in the affected and unaffected groups. For example, if clinical cases were identified by means of a test battery but unaffected participants were identified on the basis of an absence of parental concern, then estimates of an index test's diagnostic accuracy could be distorted due to the presence of undiagnosed cases in the unaffected group (Battaglia et al., 2002). To minimize this validity threat, the diagnostic status of affected and unaffected participants should be established by the same reference standard and identical procedures.

A special type of differential verification bias, known as *incorporation bias,* can occur if results from the index test contribute to determining participants' diagnostic status, rather than the reference standard alone. For example, in some of the studies of diagnostic accuracy described below, a discriminant function analysis (specifically, a predictive discriminant analysis; cf. Huberty & Hussein, 2003) of scores on several measures in a sample of affected and unaffected individuals was used to identify the combination of measures that most successfully separated the affected and unaffected groups. The accuracy of the newly derived measure was then tested in a different sample of affected and unaffected individuals. If this second step had not been taken (i.e., if the new measure's accuracy simply had been reported for the same sample from which it had been derived), incorporation bias could have inflated the estimates of its accuracy.

A final methodological issue in studies of diagnostic accuracy concerns the potential for subjective bias to influence results from the reference standard and the index test. To minimize this threat to internal validity, the reference standard and the index test must be administered to each participant by different examiners, and examiners must be unaware of any information, including results from the other test, that could allow them to deduce the diagnostic status of the participant they are testing.

Choice of reference standard. The measure used to determine participants' correct or actual diagnostic status has traditionally been described as the *gold standard,* a phrase with an unfortunate connotation of perfection. Of course perfect diagnostic indicators are rare, particularly in the behavioral sciences in which diagnostic categories and indicators are subject to ongoing debate and refinement (e.g., Rounsaville et al., 2002). The term *reference standard* better reflects the reality of diagnostic classification in such situations, in which the accuracy of proposed new diagnostic tools must be gauged relative to admittedly imperfect measures. Studies of diagnostic accuracy in communication disorders and many other fields often require something like a bootstrapping process, in which promising new diagnostic tools are evaluated by comparison to fallible reference standards that have been defined explicitly and chosen on the basis of a persuasive rationale.

The important point is that a reference standard must be chosen, described, and justified for any study of diagnostic accuracy. If no widely agreed formal reference standard exists for the disorder, then informal standards, such as ratings by expert clinicians (e.g., Shriberg, Aram, & Kwiatkowski, 1997) or prior enrollment in treatment (e.g., Dollaghan & Campbell, 1998), can provide a reasonable starting point for investigation, especially if supported by evidence of their validity and reliability. An absence of information on the reference standard used in a study of diagnostic accuracy, however, is a serious concern because classification errors by the reference standard necessarily threaten the validity of inferences about the accuracy of the index measure.

Effect size. The final special consideration in studies of diagnostic accuracy concerns the effect size metric. In studies of treatment, in which the comparison of interest is between the effects of different treatments or treatment conditions, the effect size metric is most often a standardized group mean difference measure such as Cohen's d or a measure of variance accounted for such as R^2. In studies of diagnostic accuracy, however, the critical measure is the accuracy with which the index measure assigns individuals to the correct diagnostic category (affected or unaffected) as defined by performance on the reference standard. A variety of metrics for quantifying the accuracy of an index measure exist; these include both global measures of test accuracy, such as the diagnostic odds ratio and the area under the receiver operating characteristic curve, and measures of accuracy with respect to category assignments for individuals, such as sensitivity, specificity, likelihood ratios, and positive and negative predictive values. The purposes, relationships, characteristics, advantages, and limitations of these alternatives have been addressed in several sources (e.g., Battaglia et al., 2002; Bossuyt et al., 2003). The most widely used metrics for meta-analyses of diagnostic accuracy are sensitivity, specificity, positive likelihood ratio (LR+) and negative likelihood ratio (LR–).

All diagnostic accuracy metrics have as their starting point a 2 × 2 contingency table in which two cells contain the number of people correctly classified by the index measure as affected (true positives, upper left cell a) or as unaffected (true negatives, lower right cell d). The other two cells contain the number of people misclassified by the index measure, as unaffected (false negatives, lower left cell c) or as affected (false positives, upper right cell b). The frequencies in the contingency table can be used to calculate a variety of accuracy metrics; their 95% confidence intervals (CIs) can be calculated by hand (see Sackett et al., 2000; Straus, Richardson, Glasziou, & Haynes, 2005) or by a reputable online diagnostic accuracy calculator such as can be found at www.cebm.utoronto.ca or http://spph.ubc.ca/sites/healthcare/files/calc/bayes.html. A labeled contingency table and a table of definitions and formulas for calculating six common accuracy metrics are included as supplementary online material.

Many investigators (e.g., Battaglia et al., 2002, Sackett et al., 2000; Straus et al., 2005) have noted the advantages of likelihood ratios over other diagnostic accuracy metrics. As a rule of thumb (Dollaghan, 2007; Sackett, Haynes, Guyatt, & Tugwell, 1991; Sackett et al., 2000), a measure with an LR+ ≥ 10 or, more specifically, an LR+ for which the lower bound of the 95% CI is ≥ 10, is clinically informative for identifying someone with a disorder; a person who scores in the affected range on such a measure is very likely to have the disorder. A measure with an LR– ≤ 0.10 or, more specifically, one for which the upper bound of the 95% CI is ≤ 0.10, is likewise informative for ruling out the presence of the disorder; a person scoring in the unaffected range on such a measure is very likely not to have the disorder. Results from measures with LRs+ around 3.00 and LRs– around 0.30 are viewed only as suggestive, meaning that additional testing would be necessary to diagnose presence or absence of the disorder with confidence. LRs+ or LRs– approximating 1.0 characterize measures that are diagnostically uninformative.

A Meta-Analysis of Diagnostic Accuracy in Bilingual Language Assessment

We next present a meta-analysis of the diagnostic accuracy of measures intended to identify specific or primary LIs (language disorders unaccompanied by other developmental deficits; Tomblin, Zhang, Buckwalter, & O'Brien, 2003) in U.S. children who are bilingual Spanish–English speakers. The diagnosis of LI in bilingual Spanish–English children was chosen for two reasons. First, the number of school-age children designated as English-language learners, the majority of whom are Latinos (Linan-Thompson & Ortiz, 2009), is increasing; data suggest that nearly 70% of Hispanic or Latino children between the ages of 5 and 17 speak a language other than English (U.S. Census Bureau, 2004). Second, a variety of methods for diagnosing language disorders in bilingual Spanish–English children have been reported, and a synthesis of the evidence concerning their diagnostic accuracy values could be helpful to practitioners making decisions about measures for clinical use as well as to researchers planning and conducting studies of diagnostic accuracy for these or other measures.

In mature areas of investigation, meta-analyses of diagnostic accuracy may synthesize evidence across multiple studies of one or a small number of measures such that an optimal diagnostic tool can be identified and/or the variables influencing diagnostic results can be identified (e.g., Chan, Ang, Bryant, Zamora, & Naik, 2009). However, when diagnostic tools are essentially being bootstrapped in the context of discovery, as is the case for diagnosing LI in bilingual children and many other diagnostic tasks (e.g., Alldred, Alfirevic, Deeks, & Neilson, 2008), a meta-analysis across the full range of measures can suggest the measures that appear most promising and identify gaps and flaws in the evidence that should be addressed in future diagnostic studies. Because the literature on diagnosing LI in bilingual children is relatively new, we did not anticipate that the number and quality of studies would be sufficient to support strong conclusions about an optimal diagnostic approach for this purpose. A meta-analysis does, however, offer a systematic way to take stock of the current state of the evidence and to pinpoint the steps that can be taken by investigators to increase the quality of the evidence in future research. The meta-analysis reported below illustrates many of the points described above and draws heavily on the framework proposed in the in-progress *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* (Bossuyt & Leeflang, 2008; de Vet et al., 2008; Reitsma et al., 2009; Smidt, Deeks, & Moore, 2008).

Method

Search Methods

Two electronic databases were searched in July 2009 using the following search terms: *bilingual language disorders, bilingual language impairment, bilingual child language disorders, bilingual developmental language disorders, bilingual language assessment, bilingual language testing,* and *bilingual language tests*. In PubMed, a database incorporating the approximately 5,200 biomedical journals indexed for MEDLINE from approximately 1949 to the present, a broad, sensitive search was conducted for each term using the Clinical Queries function; according to PubMed, the full filter for such searches is "(sensitiv*[Title/Abstract] OR sensitivity and specificity[MeSH Terms] OR diagnos*[Title/Abstract] OR diagnosis[MeSH:noexp] OR diagnostic * [MeSH:noexp] OR diagnosis,differential[MeSH:noexp] OR diagnosis[Subheading:noexp])". The same seven terms were searched in EBSCO Academic Search Complete, which includes citations from PsycINFO, the Psychology and Behavioral Sciences Collection, Education Research Complete, PsycARTICLES, and Communication and Mass Media Complete dating back to at least 1950. The reference lists of relevant publications, including books and book chapters, were searched by hand. Conference presentations and sources published in languages other than English were excluded.

The titles and abstracts of studies were examined to exclude those that clearly did not concern the identification of language disorders in bilingual Spanish–English U.S. children. Full texts of the remaining potentially relevant studies were read to determine whether they (a) concerned methods for identifying language disorders in bilingual Spanish–English children 3–15 years of age acquiring both languages simultaneously or acquiring English as a second language; (b) included at least five children with primary LI and five children with typical language (TL) skills; and (c) reported cell frequencies for a 2 × 2 contingency table directly or reported sensitivity, specificity, and the numbers of children with LI and TL so that cell frequencies could be calculated. Studies meeting these criteria were included irrespective of their designs, index tests, and reference standards, which were extracted and reported separately.

Data Extraction and Analysis

The authors of the present study independently extracted information from each study concerning participant recruitment, ages, description of bilingual and language status, and reference standard. The authors also independently rated each study with respect to (a) research design (one-gate or two-gate), (b) whether all participants underwent the same reference standard and index test, (c) whether different examiners administered the reference standard and index test to each participant, and (d) whether examiners were blinded to information concerning the language status (LI or TL) of the participants they tested. Finally, the authors independently determined contingency table cell frequencies for each index measure.

Contingency table values were entered into Meta-DiSc (Zamora, Abraira, Muriel, Khan, & Coomarasamy, 2006), which yielded point estimates; associated 95% CIs; and forest plots (graphic displays) of sensitivity, specificity, and LR+ and LR– values across all index measures.

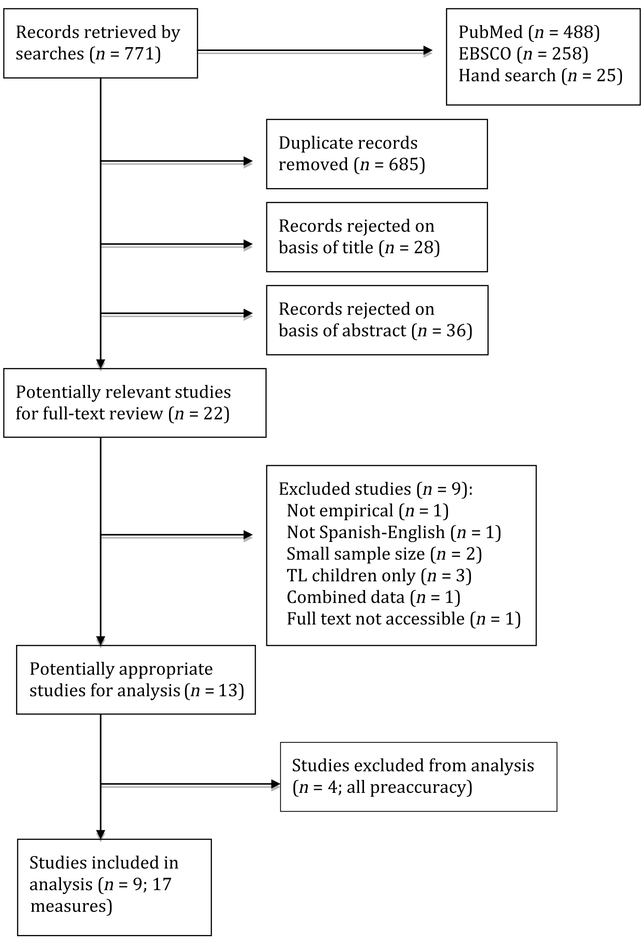
Results

Search Results

Figure 1 shows the search results in the graphic format recommended by de Vet et al. (2008). The initial search yielded 771 citations, 86 of which were unique. Of these, 28 were excluded on the basis of their titles alone. The authors of the present study independently reviewed the abstracts of the remaining records for potential relevance, agreeing on 50/58 (86%) of these decisions. The 36 records judged unanimously as not potentially relevant were excluded, as was one record for which the full text was not yet accessible electronically. Full texts of remaining records were reviewed, and eight additional studies were judged unanimously as not meeting the inclusion criteria. Of the remaining 13 studies, four were excluded as preaccuracy studies lacking the information necessary to calculate accuracy metrics.

Two of the nine remaining studies (Restrepo, 1998; Simon-Cereijido, & Gutiérrez-Clellen, 2007) included a

Figure 1. Search results.


Records retrieved by searches (n = 771)
PubMed (n = 488)
EBSCO (n = 258)
Hand search (n = 25)
Duplicate records removed (n = 685)
Records rejected on basis of title (n = 28)
Records rejected on basis of abstract (n = 36)
Potentially relevant studies for full-text review (n = 22)
Excluded studies (n = 9):
Not empirical (n = 1)
Not Spanish-English (n = 1)
Small sample size (n = 2)
TL children only (n = 3)
Combined data (n = 1)
Full text not accessible (n = 1)
Potentially appropriate studies for analysis (n = 13)
Studies excluded from analysis (n = 4; all preaccuracy)
Studies included in analysis (n = 9; 17 measures)

phase in which exploratory discriminant function analyses were used to identify maximally informative candidate diagnostic indicators from relatively larger samples of bilingual children with LI and TL, and a phase in which the accuracy of these derived measures was tested in independent and smaller samples of participants with LI and TL. Because of the threat of incorporation bias, we considered only the accuracy results from the second phase of these studies.

Accordingly, across these nine studies accuracy metrics and CIs could be calculated for a total of 100 children with LI and 109 children with TL. The mean age across studies was 6;8 (years;months), with a range from 4;0 to 9;9.

Data Extraction and Analysis Results

Definitions of bilingualism and LI. Table 1 shows how bilingual status and LI were determined in each study. Inconsistencies in reporting prevented the planned independent extraction of this information, so Table 1 reflects a consensus reached by the investigators after close reading and discussion of each study. In most studies, participants were defined as Spanish–English bilinguals on the basis of some combination of parent reports, teacher reports, placement in Spanish-language educational settings, and performance on language measures. In most studies, Spanish was described as the child's first, primary, native, or dominant language and English as the child's second language; in one study, participants were described as Spanish–English bilingual, and in another, participants were characterized as English-dominant bilingual. Some studies included subsets of monolingual Spanish and/or English children, but these data were excluded from the diagnostic accuracy analyses reported below.

The process by which children were identified as LI and TL also appeared to vary widely across studies, and here, too, descriptions of the process were rarely well specified. This is understandable given the lack of diagnostic tools for which English and Spanish versions exist, but it complicates efforts to determine whether children with LI and TL underwent the same evaluation process and reference standard. The reference standard for LI generally appeared to involve some combination of clinical judgment of experienced bilingual professionals (speech-language pathologists or educators) and parent and/or teacher concern; in some studies, professional judgments were verified by subsequent testing. Children with TL were identified on the basis of a lack of concern by professionals and parents, in some studies again verified by testing. In one study, the term *language delay* was used to refer to the target condition; the others used either *language impairment* or *specific language impairment* for this purpose. Irrespective of label, all studies provided direct or indirect evidence that both TL and LI

Table 1. Description and determination of bilingual status, and determination of language impairment (LI), for each study.

Study	Description and determination of bilingual status	Determination of LI
Girbau & Schwartz (2008)	L1 = Spanish; L2 = English; by teacher report, testing, and parent questionnaire	SLP report confirmed by testing, parent, teacher report; being treated for LI or scheduled for evaluation of suspected LI
Gutiérrez-Clellen & DeCurtis (1999)	L1 = Spanish; L2 = English; tested as limited English proficient; receiving bilingual instruction	Clinical judgments of qualified professionals; on caseloads of bilingual SLPs
Gutiérrez-Clellen et al. (2006)	L1 = Spanish; L2 = English by parent, teacher questionnaires on language exposure and use	Parent, teacher concern and SLP judgments
Gutiérrez-Clellen & Simon-Cereijido (2007)	English-dominant bilingual by parent, teacher questionnaires on language exposure and use	Parent, teacher concern and SLP judgments
Gutiérrez-Clellen et al. (2008)	Spanish–English bilingual by parent, teacher questionnaires on language exposure and use	Parent, teacher concern and SLP judgments
Jacobson & Schwartz (2005)	L1=Spanish; L2=English; bilingual classroom placement re: state criteria	Bilingual SLP evaluation, parent agreement with diagnosis, intervention for ≥ 2 yr
Restrepo (1998)	L1=Spanish; L2=English by parent, teacher questionnaires, in classroom with Spanish for content instruction	Bilingual SLP diagnosis, teacher judgment, in intervention
Roseberry & Connell (1991)	L1=Spanish; L2=English, designated as limited English proficient by school personnel	Diagnosis by bilingual SLP, test results and teacher judgments
Simon-Cereijido & Gutiérrez-Clellen (2007)	L1=Spanish; L2=English by parent, teacher questionnaires on language exposure and use, limited English proficiency	Parent, teacher concern, bilingual SLP judgment

Note. L1 = first language; L2 = second language; SLP = speech-language pathologist.

Table 2. Quality feature ratings for each study.

Study	Design	Same tests to all	Independent testing	Blinded testing
Girbau & Schwartz (2008)	2-gate	Yes	Yes	No
Gutiérrez-Clellen & DeCurtis (1999)	2-gate	Yes	Unclear	No
Gutiérrez-Clellen et al. (2006)	2-gate	Unclear	Unclear/Yes	No
Gutiérrez-Clellen & Simon-Cereijido (2007)	2-gate	Yes	Unclear	No
Gutiérrez-Clellen et al. (2008)	2-gate	Yes	No/Unclear	No
Jacobson & Schwartz (2005)	2-gate	Yes/No	No	No
Restrepo (1998)	2-gate	Yes/No	Yes	Yes
Roseberry & Connell (1991)	2-gate	Yes	Unclear/No	No
Simon-Cereijido & Gutiérrez-Clellen (2007)	2-gate	Yes	No	No

Note. In instances of both raters agreeing, a single rating is listed. In instances of raters disagreeing, the individual ratings are listed. If a feature could not be judged with confidence, it was rated as unclear.

participants had hearing, cognition, and/or academic skills consistent with their ages.

Ratings of quality features. Table 2 shows results of independent ratings of the nine studies with respect to four quality features; raters agreed on 31 of these 36 decisions (86%). Two-gate designs were used in all studies, in which participants in the TL and LI groups were selected to meet specified criteria. Most studies stated or implied that all participants were administered both the index measure and the reference standard, but it was difficult in most cases to determine whether the measures were administered by independent examiners. Only one study (Restrepo, 1998) clearly indicated that examiners were blinded to relevant information about the children they tested.

Index measures and accuracy metrics. Table 3 shows for each study and index measure the number of bilingual children with LI and TL, their mean ages or age ranges, and the contingency table cell frequencies. The nine studies provided diagnostic accuracy information for 15 different index measures, one of which was examined in three different age groups (Gutiérrez-Clellen,

Table 3. Number and mean ages of participants with primary LI and typical language (TL), index tests, and contingency table values from each study.

Study	LI *n* (age)	TL *n* (age)	Index test(s)	TP	FP	FN	TN
Girbau & Schwartz (2008)	11 (8;10)	11 (9;1)	Nonword repetition	9	1	2	10
Gutiérrez-Clellen & DeCurtis (1999)	8 (9;9)	9 (9;8)	Formal word definition	4	0	4	9
Gutiérrez-Clellen et al. (2006)	9 (4;0–5;1)	9 (4;0–5;1)	Spanish Morphosyntax Test	7	0	1	8
	7 (5;2–5;11)	7 (5;2–5;11)	Spanish Morphosyntax Test	7	0	0	7
	7 (6;1–7;0)	7 (6;1–7;0)	Spanish Morphosyntax Test	3	0	4	7
Gutiérrez-Clellen & Simon-Cereijido (2007)	10 (5;1)	10 (5;1)	English Morphosyntax Test	8	1	2	9
Gutiérrez-Clellen et al. (2008)	11 (5;7)[a]	16 (5;7)[a]	Finite English verb marking	10	6	1	10
			Obligatory English subject	1	0	10	16
Jacobson & Schwartz (2005)	12 (8;1)	15 (8;0)	Past-tense, regular English verb	11	2	1	13
			Past-tense, irregular English verb	11	4	1	11
			Past-tense, nonsense verb	12	4	0	11
Restrepo (1998)	8 (6;1)	8 (6;2)	PRSLP + NETU[b]	7	0	1	8
			PRSLP + NETU + MLTU + FHSLP[b]	7	0	1	8
Roseberry & Connell (1991)	13 (5;6)	13 (5;8)	Invented English morpheme	10	1	3	12
Simon-Cereijido & Gutiérrez-Clellen (2007)	5 (4;5)	5 (4;5)	MLU + UNGRAMM[b]	4	1	1	4
			CLITIC + VERB + ART[b]	4	1	1	4
			MLU + THEME + DITRAN[b]	5	1	0	4

Note. The language of index tests was Spanish except where English is noted. Age appears in years;months. TP = true positive; FP = false positive; FN = false negative; TN = true negative; PRSLP = parent report of speech and language problems; NETU = number of errors per T-unit; MLTU = mean length of T-unit; FHSLP = family history of speech and language problems; MLU = mean length of utterance in words; UNGRAMM = percentage of utterances with grammatical errors; CLITIC = clitic errors; VERB = verb errors; ART = article errors; THEME = percentage of correct use of theme arguments; DITRAN = percentage of ditransitive verbs.

[a] Age not reported separately for LI and TL. [b] Measure derived from discriminant function analyses.

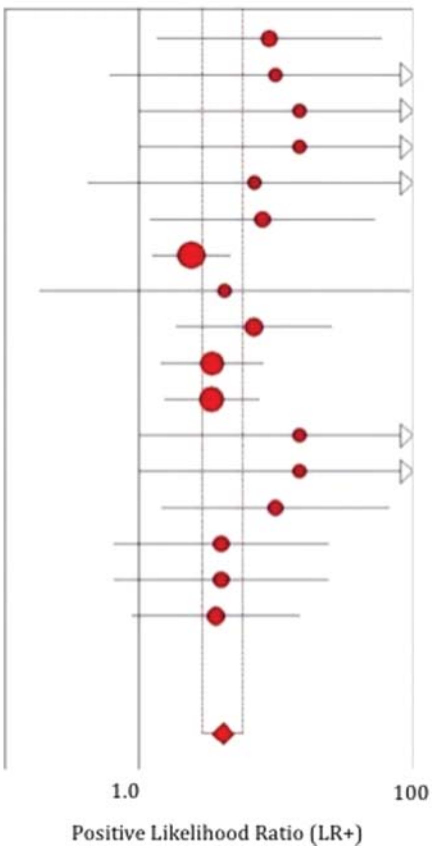
Restrepo, & Simon-Cereijido, 2006). Accordingly, complete contingency tables could be constructed for 17 measures, five of which had been derived from exploratory discriminant function analyses. The measures spanned a number of aspects of language. Morphosyntactic skills were represented singly or in combination with other skills in 13 of the 15 different index measures. Nonword repetition, invented morpheme learning, family history of speech-language problems, and parental concerns about speech and language were also represented singly or in combination.

Likelihood ratios and CIs. The contingency table values shown in Table 3 were entered into Meta-DiSc to generate sensitivity, specificity, likelihood ratios, and associated 95% CIs for these 17 index measures. A value of ½ was added to cells with a frequency of zero in order to enable CIs to be calculated (Zamora et al., 2006). A table of point estimates and 95% CIs for sensitivity and specificity values for each index measure is available online as supplementary materials.

Figure 2 and Figure 3 are forest plots showing LR+ and LR–, respectively, for each index measure. Each filled circle represents the LR for the measure listed to its right; the horizontal line through the circle represents its 95% CI, and arrows indicate CIs exceeding values on the graph. The numeric values for each measure also appear to its right. The pooled (average) LR across measures, calculated using a random effects model, is shown by a filled diamond near the bottom of the graph; vertical dashed lines extending throughout the plot show the 95% CI for the pooled value and facilitate comparisons of individual measures to the pooled value.

On both graphs, the *x*-axis labels the LR value for diagnostically uninformative measures (1.0), shown via

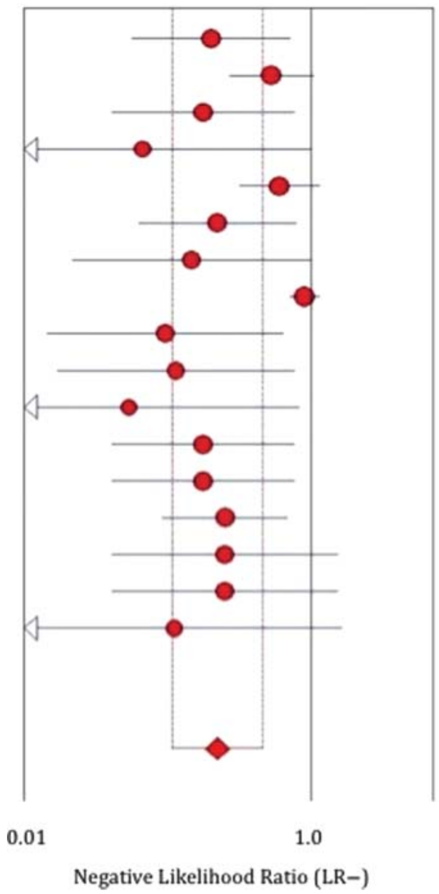
Figure 2. Forest plot of positive likelihood ratio (LR+) values and confidence intervals for each index measure, alphabetized by author(s) of study. [a]Girbau & Schwartz (2008). [b]Gutiérrez-Clellen & DeCurtis (1999). [c]Gutiérrez-Clellen et al. (2006). [d]Gutiérrez-Clellen & Simon-Cereijido (2007). [e]Gutiérrez-Clellen et al. (2008). [f]Jacobson & Schwartz (2005). [g]Restrepo (1998) measures derived from parent report of speech-language concerns (PR) and grammatical errors per T-unit (ETU); and from PR, ETU, mean length of T-unit (LTU), and positive family history (FH). [h]Roseberry & Connell (1991). [i]Simon-Cereijido & Gutiérrez-Clellen (2007), measures derived from mean length of utterance (MLU) and ungrammatical utterances (UNGRAMMs); from clitics (C), verbs (V), and articles (ART); and from MLU, themes, and ditransitives.



Measure	LR+	95% CI
Nonword repetition[a]	9.00	1.36, 59.54
Formal word definition[b]	10.00	0.62, 161.11
Spanish Morphosyntax Test – age 4[c]	15.00	1.00, 225.33
Spanish Morphosyntax Test – age 5[c]	15.00	1.02, 220.92
Spanish Morphosyntax Test – age 6[c]	7.00	0.43, 114.71
English Morphosyntax Test[d]	8.00	1.21, 52.69
Finite verb marking[e]	2.42	1.25, 4.69
Obligatory Subject[e]	4.25	0.19, 95.68
Past tense – regular verb[f]	6.88	1.87, 25.26
Past tense – irregular verb[f]	3.44	1.46, 8.09
Past tense – nonce verb[f]	3.42	1.55, 7.54
PR+ETU[g]	15.00	1.00, 225.33
PR+ETU+LTU+FH[g]	15.00	1.00, 225.33
Invented morpheme[h]	10.00	1.49, 67.29
MLU+UNGRAM[i]	4.00	0.66, 24.37
CLITIC+V+ART[i]	4.00	0.66, 24.37
MLU+THEME+DITRAN[i]	3.67	0.90, 14.97
Pooled LR+ (Random effects model)	4.12	2.94, 5.78

Inconsistency (I-square) = 0.0%; Cochrane-Q = 11.97 (df = 16, *p* = .75)

Figure 3. Forest plot of negative likelihood ratio (LR–) values and confidence intervals for each index measure, alphabetized by author(s) of study. [a]Girbau & Schwartz (2008). [b]Gutiérrez-Clellen & DeCurtis (1999). [c]Gutiérrez-Clellen et al. (2006). [d]Gutiérrez-Clellen & Simon-Cereijido (2007). [e]Gutiérrez-Clellen et al. (2008). [f]Jacobson & Schwartz (2005). [g]Restrepo (1998) measures derived from PR and ETU; and from PR, ETU, mean LTU, and positive FH. [h]Roseberry & Connell (1991). [i]Simon-Cereijido & Gutiérrez-Clellen (2007) measures derived from MLU and UNGRAMMs; from C, V, and ART; and from MLU, themes, and ditransitives.



Measure	LR−	95% CI
Nonword repetition[a]	0.20	0.06, 0.71
Formal definition[b]	0.53	0.27, 1.03
Spanish Morphosyntax Test – age 4[c]	0.18	0.04, 0.77
Spanish Morphosyntax Test – age 5[c]	0.07	0.00, 0.98
Spanish Morphosyntax Test – age 6[c]	0.60	0.32, 1.13
English Morphosyntax Test[d]	0.22	0.06, 0.78
Finite verb marking[e]	0.15	0.02, 0.98
Obligatory Subject[e]	0.90	0.72, 1.13
Past tense – regular verb[f]	0.10	0.01, 0.63
Past tense – irregular verb[f]	0.11	0.02, 0.76
Past tense – nonce verb[f]	0.05	0.00, 0.82
PR+ETU[g]	0.18	0.04, 0.77
PR+ETU+LTU+FH[g]	0.18	0.04, 0.77
Invented morpheme[h]	0.25	0.09, 0.68
MLU+UNGRAM[i]	0.25	0.04, 1.52
CLITIC+V+ART[i]	0.25	0.04, 1.52
MLU+THEME+DITRAN[i]	0.11	0.01, 1.64
Pooled LR− (Random effects model)	0.22	0.11, 0.46

Inconsistency (I-square) = 86.7%; Cochrane-Q = 119.93 (df = 16, *p* = .00)

a solid vertical line. On Figure 2, the rightmost *x*-axis label shows a positive LR value of 100, well above the value (10) that would enable high confidence in using scores in the LI range to diagnose LI. In Figure 3, the leftmost *x*-axis label corresponds to a negative LR of 0.01, a value well below that (0.10) which would enable high confidence in using normal-range scores to rule out the presence of LI. In evaluating any point estimate, such as an LR, it is crucial to consider its 95% CI. Ideally, the lower bound of the 95% CI around an LR+ would be greater than 10, indicating a very low (5%) chance that the true LR+ value falls below this level. For the same reason, the upper bound of the 95% CI around an LR– would ideally be lower than 0.10.

Figure 2 shows that all but one of the index measures had LR+ values exceeding the value (3) deemed suggestive; six measures met or exceeded the level (10) deemed clinically informative for diagnosing LI. In addition, LR+ values for the individual measures aligned reasonably well with the pooled LR+ value of 4.12, suggesting a lack of significant heterogeneity among this set of measures. This visual impression is buttressed by the low Inconsistency (I^2) value, which indicates that none (0.0%) of the variation in LR+ values was due to between-measure heterogeneity (Higgins & Thompson, 2002). Similarly, a Cochrane-Q test of the null hypothesis that variations between the results for individual measures are due to chance was not significant.

The fact that most of these measures had LR+ values in the suggestive range appears somewhat encouraging, but it is important to note that CIs for all measures were wide, due at least in part to the small sample sizes in these studies. Importantly, the lower bounds of the 95% CIs for all measures fell near, and in some cases below, the value (1.0) at which scores are diagnostically uninformative for identifying LI. These broad CIs vitiate the clinical utility of these measures for diagnosing LI in bilingual Spanish–English children at present, but taken as a whole, the forest plot suggests that most of these measures are worthy of further investigation.

The LR– values shown in Figure 3 are somewhat more variable and less encouraging than the LR+ results. LR– values for most of these measures fell at the level viewed as suggestive for identifying the absence of LI (0.30), but the upper bound of the 95% CI for every measure was considerably higher, and several measures had upper bounds near or above 1.0, indicating that a score in the unaffected range on the measure has no value for identifying typical language. The validity of the pooled LR– value is doubtful because of the excessive heterogeneity between measures that is reflected both in the high I^2 value and in the significant Cochrane-Q (Zamora et al., 2006, p. 3). The number of studies was too small, however, to support a heterogeneity analysis aimed at identifying the specific factor(s) contributing to the variations between the LR– results for these measures.

Discussion

We presented an overview of studies of diagnostic accuracy and reported results of a meta-analysis of diagnostic accuracy in measures intended to identify language disorders in bilingual Spanish–English U.S. children. Although searches yielded a large number of publications addressing the general topic of diagnosing LI in bilingual Spanish–English children, accuracy metrics could be calculated for only nine studies reporting evidence on 15 different index measures, one of which was studied at three different ages. Index measures varied widely in format, emphasis, and origin; they included parent reports, spontaneous and elicited expressive language production, novel word and morpheme learning, and combinations of measures derived from discriminant function analyses. The reference standards against which accuracy of the index tests were judged appeared to be similarly diverse.

Studies were rated with respect to several features associated with high-quality evidence on diagnostic accuracy. All studies were susceptible to spectrum bias by virtue of their two-gate research designs, but such designs are common during early-stage research aimed at identifying candidate diagnostic measures worthy of future investigation. Of greater concern was a lack of information needed to appraise certain key quality features, such as whether participants with LI and TL underwent identical reference standards and testing procedures and whether the reference standard and index test were administered to individual children by independent examiners. Subjective bias appears to have been a serious threat to the validity of most studies; only one report clearly indicated that examiners were blinded to information on the diagnostic status of the children they tested. With respect to accuracy metrics, LR+ and LR– values suggested that many of these diverse diagnostic measures might hold promise for identifying bilingual Spanish–English children with LI and TL, but 95% CIs for all measures included values considered diagnostically uninformative.

Our primary purpose in undertaking this meta-analysis was to synthesize and appraise the existing evidence, acknowledging strengths as well as identifying areas of weakness that can guide future investigations. Before addressing our findings in this vein, however, it is a fair question to ask whether this meta-analysis has any implications for clinical practice. Because we recognized at the outset that the literature on bilingual language assessment is relatively recent, we did not anticipate that our results would be sufficient to support strong recommendations about “best practice” in diagnosing LI and TL in bilingual Spanish–English U.S. children. Indeed, the paucity of meta-analyses on diagnosis suggests that there are few disorders, whether medical or behavioral, for which the requisite quality and quantity of diagnostic evidence is available to support such a recommendation. Although strong conclusions about clinical practice are premature, however, we suggest that our results might support clinical reasoning in the following ways.

First and foremost, our findings show that no measure stands out as the optimal method for identifying LI or TL in bilingual Spanish–English children. At best the accuracy metrics for these measures fall in the suggestive range, meaning that every measure would need to be supplemented by additional, and unspecified, information in order to identify children as LI or TL. Figure 2 does show that 10 measures have LR+ values that exceed the upper bound of the CI of the average LR+, so a clinician might reasonably focus on these 10 measures as somewhat better candidates for identifying LI. However, the fact that the CIs for LR+ values overlap for all 17 measures and the fact that all measures but one fall within the CI for the averaged value mean that there is little reason to strongly prefer one to the others at present. Grounds for any preferences among measures for identifying children with TL are even weaker, based not only on the overlapping CIs of the LR– values shown in Figure 3 but also on excessive between-measure heterogeneity. In short, until more, stronger, and more precise evidence is available, clinicians can justify using most of these measures in efforts to identify LI or TL in bilingual Spanish–English U.S. children, but the results of any single measure must be viewed as no more than somewhat suggestive of diagnostic status.

The present meta-analysis suggests another possible implication for clinicians. Specifically, although we found no single, widely agreed reference standard for diagnosing presence or absence of LI in bilingual Spanish–English children, the impressions of parents, teachers, and speech-language pathologists about whether a child's language skills warrant concern seemed to be part of the reference standard in many of the studies we analyzed. Clinical practitioners could have an important role to play in specifying the characteristics that cause such concerns, and in compiling data on the accuracy of such impressions using the evidence-based practice perspective. Such "practice-based evidence" (e.g., Nutting, Beasley, & Werner, 1999) is increasingly viewed as a critical source of clinical insights to be tested in future research.

Apart from these necessarily tentative clinical implications, the present meta-analysis has several implications for efforts to improve the quality of evidence concerning the diagnosis of LI in bilingual children. All of the studies we examined provided direct evidence concerning diagnostic accuracy, an advance over preaccuracy studies in which only group mean comparisons or correlational findings are reported. Examining these important, albeit early-phase, accuracy studies suggests two areas in which relatively simple steps could substantially increase the quality of the evidence.

The first area concerns research design and participant selection issues. Similar to early-phase studies of treatment, in which case-control designs are acceptable despite their limitations, early-phase studies of diagnostic accuracy such as those we examined may depend on the use of two-gate designs. Describing participant recruitment and selection procedures is important in any study, and it is crucial for studies with two-gate designs. Our analysis revealed room for considerable improvement in this regard; substantially more information about recruitment and selection of both affected and unaffected participants is needed in order to evaluate the extent to which accuracy metrics could have been distorted by spectrum, differential verification, and/or incorporation biases. The reference standards used in the studies we analyzed seemed to have reasonable face validity, but whether and how they were used for selecting unaffected participants often was not clearly described. In addition, it appeared that some participants may have been included in more than one of the studies we examined; this information should be clearly stated in future investigations.

The second general area for improvement concerns controlling the potential for subjective bias to distort findings. We found no study in which it was clearly stated that the reference standard and the index test had been administered by independent examiners, and only one in which examiners were reported to have been blinded to information on diagnostic status. Future investigations should use, and specify, steps to ensure that these criteria are met to minimize the potential impact of subjective bias; evidence of adequate interexaminer reliability for each measure would also strengthen future studies.

This meta-analysis itself had several limitations, including the small number of eligible studies and the small total number of affected and unaffected participants in whom diagnostic accuracy was examined. It was possible to conduct independent ratings of only a subset of quality features due to inconsistency within and between studies in reporting the requisite information. Finally, the small number of studies made it impossible to blind raters to the identity of individual studies.

These limitations notwithstanding, the present meta-analysis illustrates some of the issues to be considered in planning, conducting, and reporting future studies and meta-analyses of diagnostic accuracy. Stronger evidence on diagnostic accuracy is needed for many if not all behavioral disorders (e.g., Kupfer, First, & Regier, 2002). The information presented here is intended to facilitate more, and more rigorous, efforts to provide such evidence concerning the diagnosis of communication disorders.

Acknowledgment

The impetus for this study was an undergraduate honors thesis completed by the second author under the direction of the first in April of 2009. We thank Tom Campbell for his helpful comments.

References

Alldred, S. K., Alfirevic, Z., Deeks, J. J., & Neilson, J. P. (2008). Antenatal screening for Down's syndrome [diagnostic test accuracy protocol]. *Cochrane Database of Systematic Reviews, 4,* (Art. No.: CD007384). doi: 10.1002/14651858.CD007384.

Battaglia, M., Bucher, H., Egger, M., Grossenbacher, F., Minder, C., & Pewsner, P. (2002). *The Bayes library of diagnostic studies and reviews* (2nd ed.). Retrieved from http://www.medepi.net/meta/guidelines/BAYES_Library.

Bossuyt, P. M., & Leeflang, M. M. (2008). Chapter 6: Developing criteria for including studies. In J. J. Deeks, P. M. Bossuyt, & C. Gatsonis (Eds.), *Cochrane handbook for systematic reviews of diagnostic test accuracy* (Version 0.4; updated September 2008). The Cochrane Collaboration, 2008. Available from www.srdta.cochrane.org.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., … de Vet, H. C., for the STARD group. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *Radiology, 226,* 24–28.

Chan, K. K. L., Ang, C., Bryant, A., Zamora, J., & Naik, R. (2009). Sentinel node biopsy for diagnosis of pelvic lymph node involvement in early stage cervical cancer. *Cochrane Database of Systematic Reviews, 3,* (Art. No.: CD007925). doi: 10.1002/14651858.CD007925.

de Vet, H. C. W., Eisinga, A., Riphagen, I. I., Aertgeerts, B., & Pewsner, D. (2008). Chapter 7: Searching for studies. In J. J. Deeks, P. M. Bossuyt, & C. Gatsonis (Eds.), *Cochrane handbook for systematic reviews of diagnostic test accuracy* (Version 0.4; updated September 2008). The Cochrane Collaboration, 2008. Available from www.srdta.cochrane.org.

Dollaghan, C. A. (2007). *The handbook for evidence-based practice in communication disorders*. Baltimore, MD: Brookes.

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research, 41,* 1136–1146.

Girbau, D., & Schwartz, R. G. (2008). Phonological working memory in Spanish-English bilingual children with and without specific language impairment. *Journal of Communication Disorders, 41,* 124–145.

Gutiérrez-Clellen, V. F., & DeCurtis, L. (1999). Word definition skills in Spanish-speaking children with language impairment. *Communication Disorders Quarterly, 21,* 23–31.

Gutiérrez-Clellen, V. F., Restrepo, M. A., & Simon-Cereijido, G. (2006). Evaluating the discriminant accuracy of a grammatical measure with Spanish-speaking children. *Journal of Speech, Language, and Hearing Research, 49,* 1209–1223.

Gutiérrez-Clellen, V. F., & Simon-Cereijido, G. (2007). The discriminant accuracy of a grammatical measure with Latino English-speaking children. *Journal of Speech, Language, and Hearing Research, 50,* 968–981.

Gutiérrez-Clellen, V. F., Simon-Cereijido, G., & Wagner, C. (2008). Bilingual children with language impairment: A comparison with monolinguals and second language learners. *Applied Psycholinguistics, 29,* 3–19.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21,* 1539–1558.

Huberty, C. J., & Hussein, M. H. (2003). Some problems in reporting use of discriminant analyses. *Journal of Experimental Education, 71,* 177–191.

Irwig, L., Bossuyt, P., Glasziou, P., Gatsonis, C., & Lijmer, J. (2002). Designing studies to ensure that estimates of test accuracy are transferable. *BMJ, 324,* 669–671.

Jacobson, P. F., & Schwartz, R. G. (2005). English past tense use in bilingual children with language impairment. *American Journal of Speech-Language Pathology, 14,* 313–323.

Knottnerus, J. A., & van Weel, C. (2002). General introduction: Evaluation of diagnostic procedures. In J. A. Knottnerus (Ed.), *Evidence base of clinical diagnosis* (pp. 1–18). London, England: BMJ Publishing Group.

Kupfer, D. J., First, M. B., & Regier, D. A. (Eds.). (2002). *A research agenda for DSM-V.* Washington, DC: American Psychiatric Association.

Leeflang, M. M. G., Bossuyt, P. M. M., & Irwig, L. (2009). Diagnostic test accuracy may vary with prevalence: Implications for evidence-based diagnosis. *Journal of Clinical Epidemiology, 62,* 5–12.

Linan-Thompson, S., & Ortiz, A. A. (2009). Response to intervention and English-language learners: Instructional and assessment considerations. *Seminars in Speech and Language, 30,* 105–120.

Meehl, P. E. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist, 50,* 266–275.

Nutting, P. A., Beasley, J. W., & Werner, J. J. (1999). Practice-based research networks to answer primary care questions. *Journal of the American Medical Association, 281,* 686–688.

Reitsma, J. B., Rutjes, A. W. S., Whiting, P., Vlassov, V. V., Leeflang, M. M. G., & Deeks, J. J. (2009). Chapter 9: Assessing methodological quality. In J. J. Deeks, P. M. Bossuyt, & C. Gatsonis (Eds.), *Cochrane handbook for systematic reviews of diagnostic test accuracy* (Version 1.0.0). The Cochrane Collaboration, 2009. Available from http://srdta.cochrane.org/.

Restrepo, M. A. (1998). Identifiers of predominantly Spanish-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research, 41,* 1398–1411.

Roseberry, C. A., & Connell, P. J. (1991). The use of an invented language rule in the differentiation of normal and language-impaired Spanish-speaking children. *Journal of Speech and Hearing Research, 34,* 596–603.

Rounsaville, B. J., Alarcon, R. D., Andrews, G., Jackson, J. S., Kendell, R. E., & Kendler, K. (2002). Basic nomenclature issues for DSM-V. In D. J. Kupfer, M. B. First, & D. A. Regier (Eds.), *A research agenda for DSM-V* (pp. 1–29). Washington, DC: American Psychiatric Association.

Sackett, D. L., Haynes, R. B., Guyatt, G. H., & Tugwell, P. (1991). *Clinical epidemiology: A basic science for clinical medicine*. Boston, MA: Little, Brown.

Sackett, D. L., Straus, S. E., Richardson, W. S., Rosenberg, W., & Haynes, R. B. (2000). *Evidence-based medicine: How to practice and teach EBM*. Edinburgh, Scotland: Churchill Livingstone.

Shriberg, L. D., Aram, D. M., & Kwiatkowski, J. (1997). Developmental apraxia of speech: III. A subtype marked by inappropriate stress. *Journal of Speech, Language, and Hearing Research, 40,* 313–337.

Simon-Cereijido, G., & Gutiérrez-Clellen, V. F. (2007). Spontaneous language markers of Spanish language impairment. *Applied Psycholinguistics, 28,* 317–339.

Smidt, N., Deeks, J., & Moore, T. (Eds.). (2008). Chapter 4: Guide to the contents of a Cochrane review and protocol. *Cochrane handbook for systematic reviews of diagnostic test accuracy* (Version 0.4). The Cochrane Collaboration, 2008.

Straus, S. E., Richardson, W. S., Glasziou, P., & Haynes, R. B. (2005). *Evidence-based medicine* (3rd ed.). Edinburgh, Scotland: Elsevier.

Tomblin, J. B., Zhang, X., Buckwalter, P., & O'Brien, M. (2003). The stability of primary language disorder: Four years after kindergarten diagnosis. *Journal of Speech, Language, and Hearing Research, 46,* 1283–1296.

U.S. Census Bureau. (2004, March 16). *Characteristics of children under 18 years by age for the United States, regions, states, and Puerto Rico (Census 2000, special tabulation).* Retrieved from http://www.census.gov/population/www/cen2000/briefs/phc-t30/tables/tab01.pdf.

Zamora, J., Abraira, V., Muriel, A., Khan, K., & Coomarasamy, A. (2006). Meta-DiSc: A software for meta-analysis of test accuracy data. *BMC Medical Research Methodology, 6,* 31. Retrieved from http://www.biomedcentral.com/1471-2288/6/31.

Bilingual Language Assessment: A Meta-Analysis of Diagnostic Accuracy

Christine A. Dollaghan, and Elizabeth A. Horner

J Speech Lang Hear Res 2011;54;1077-1088; originally published online Nov 24, 2010;
DOI: 10.1044/1092-4388(2010/10-0093)