Considerations for Appraising Diagnostic Studies of Communication Disorders

Thomas Klee

Newcastle University, UK

Evidence-Based Communication Assessment and Intervention, *2*, 34-45, 2008

Current address:

Thomas Klee

Department of Communication Disorders

University of Canterbury

Private Bag 4800

Christchurch 8140

New Zealand

E-mail: thomas.klee@canterbury.ac.nz

Abstract

Clinicians have always been aware of the importance of using clinical tests and measures that are valid and reliable – and avoiding those that are not. However, the choice of which tests and measures to use is often more a matter of personal preference, arising out of knowledge of the test's psychometric properties and one's experience with the test, rather than on a systematic critical appraisal of assessment tools. This paper outlines a proposal for how clinical assessments in the speech and language sciences can be critically appraised for the purpose of deciding whether they are likely to be informative in diagnosing individuals with communication disorders. QUADAS, a 14-item evidence-based critical appraisal tool (Whiting et al., 2003), originally designed to assess the quality of diagnostic accuracy studies used in systematic reviews in medicine, is presented with an example of how it can be applied in the field of communication disorders.

Keywords: *evidence-based practice, diagnostic accuracy, critical appraisal, assessment*

Considerations for Appraising Diagnostic Studies of Communication Disorders

Accurate assessment of individuals with communication disorders provides the foundation for clinicians' decisions about diagnosis and intervention. Without accurate assessment, decisions about whether to offer intervention, what type to offer, and whether it was effective, are not possible. This paper is concerned with the question of how clinicians[1] can go about the challenging task of evaluating the usefulness of specific clinical assessment tools (e.g., standardized tests, language sample measures, instrumental measurements) designed to diagnose individuals with communication disorders, using a framework grounded in evidence-based practice (EBP). The focus here is on assessments intended for diagnosis (or classification) by "detecting or excluding disorders, by increasing diagnostic certainty as to their presence or absence" (Knottnerus & van Weel, 2002, p. 3) rather than on assessments designed for other purposes, such as measuring treatment progress, assessing prognosis or tracking the clinical course of a disorder for clients already diagnosed. The paper is presented from the perspective of the clinician faced with the task of assessing individuals suspected of having communication disorders, rather than from the perspective of those developing tests and measures, and draws on ideas presented previously by Klee, Wong, Stokes, Fletcher and Leonard (in press).

In that paper, we argued that evidence-based assessment is the joint product of those who develop clinical assessment measures and those who use them. Although the current paper focuses on the latter, and in particular on how clinicians can decide whether a particular test or measure is likely to be useful to them in their daily practice, many of the concepts are equally relevant to test developers. The focus is on evaluating current tests and measures rather than on a hypothetical set of ideal

measures that may be developed at some point in the future. This is not to suggest that developing new methods and models of clinical assessment is unnecessary; it plainly is. Rather, this paper outlines a proposal for how to critically appraise new clinical tests and measures appearing in the literature or ones that are already in use. Like all proposals, it is one that should be debated. The proposal is based on a model originating in evidence-based medicine and one that is specifically concerned with how to evaluate the diagnostic accuracy of tests and measures. This will be introduced later in the paper. Following that, we present an evidence-based tool that can be used by clinicians to assess the quality of diagnostic accuracy studies. First, we briefly review the standard psychometric assessment model that has been used to develop new assessment measures in the fields of psychology, education and speech and language sciences.

Psychometric Assessment of Tests and Measures

Traditionally, when a new test or measure is developed, data are collected from a research sample or from a representative sample of individuals selected from the population; the latter is usually referred to as a *normative group*. In the clinical setting, a client's performance on the test or measure is then compared to the range of scores produced by the normative group, usually by converting the raw score (e.g., a test score or a language sample measure such as mean length of utterance or percent consonants correct) to one of several standardized scores (e.g., percentile, z-score, standard score, or in the case of a child, an age-equivalent score). Those whose standardized score falls below a certain criterion (e.g., 1.5 standard deviations [SD] below the group mean or z = -1.50) on a test that has been shown to be psychometrically sound may then be considered for clinical services. This paradigm is

also typically used in research studies to operationally define individuals from a particular clinical population. For instance, children with specific language impairment (SLI) [2] may be selected for an experimental group on the basis of one or more standardized language scores falling below a pre-defined level, such as 1.5 SDs below the mean (after ruling out certain other conditions such as hearing loss, cognitive impairment, and social and emotional problems), whereas those in the control group may be selected on the basis of their scores not falling below that level.

The framework for assessing the psychometric adequacy of many of the clinical tests and measures used in the US and elsewhere in the speech and language sciences, psychology and education is presented in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999). It describes in detail technical standards for test construction and evaluation, including various aspects of test validity and reliability, the contents of the technical and user's manuals, and guidelines relating to professional standards of test use. Using these, McCauley and Swisher (1984) formulated a review of 30 developmental speech and language tests in use at the time, based on a set of 10 key psychometric criteria relating to the adequacy of the normative sample, the presentation of test procedures and norms, test validity and test reliability. They concluded from their review that half of the tests met no more than two criteria while only three tests met more than four criteria.

A decade later, Plante and Vance (1994) updated McCauley and Swisher's review in their examination of 21 preschool language tests, employing the same set of psychometric criteria as were used previously. Then, in what was a ground-breaking departure from the traditional psychometric assessment paradigm in use at the time,

they examined how accurate four of the tests were in differentiating children with and without SLI. They reported the sensitivity and specificity of each test based on cut-off scores that had been empirically determined, rather than relying on an arbitrary cut-off score (e.g., z = -1.50). Children in the SLI group were recruited from clinics and schools serving those children while children without SLI came from a local preschool. The cut-off *z*-score that most accurately discriminated the groups varied between +0.06 to -3.35 across the tests. They concluded that only one of the tests demonstrated an acceptable level of classification accuracy, in that it correctly identified 90% of the children in each group relative to a clinical diagnosis based on standardized test performance and clinical judgment. Subsequently, further studies examining the diagnostic accuracy of other norm-referenced language tests have come to be published (Gray, Plant, Vance, & Henrichsen, 1999; Merrell & Plante, 1997; Oetting, Cleveland & Cope, 2008; Perona, Plante, & Vance, 2005; Plante & Vance, 1995). For a recent review of 43 commercially available tests of child language, see Spaulding, Plante and Farinella (2006). In reporting test sensitivity and specificity, these were some of the first assessment studies in the field of child language disorders that moved beyond the standard psychometric assessment paradigm. The concern for diagnostic accuracy of assessment measures was also recognized by Dollaghan and Campbell (1998), who reported likelihood ratios for a new nonword repetition task designed for use with children. Measures of diagnostic accuracy (e.g., sensitivity, specificity, likelihood ratios) have been in use since the 1980s in studies of screening instruments designed to identify children with speech and language delay and two systematic reviews of these instruments have been published (Law, Boyle, Harris, Harkness, & Nye, 2000; Nelson, Nygren, Walker, & Panoscha, 2006). The conclusions of both of these systematic reviews were drawn from a sub-set of

screening studies that had been selected on the basis of methodological quality. Schlosser, Wendt and Sigafoos (2007) discuss ways in which studies can be filtered for inclusion in systematic reviews, as well as how various sources of bias can be reduced in constructing systematic reviews.

An Evidence-Based Framework for Assessing Diagnostic Accuracy

Haynes, Sackett, Guyatt and Tugwell (2006) identified what they called a *diagnostic quartet* of design features that should be incorporated into studies evaluating diagnostic tests and measures. For them, "A valid diagnostic study (1) assembles an appropriate spectrum of patients; (2) applies both the diagnostic test and reference standard to all of them; (3) interprets each blind to the other; and (4) repeats itself in a second, independent ('test') set of patients" (p. 275). They go on to present the methods and outcome measures used in such studies and discuss design flaws that conspire to make the results of diagnostic accuracy studies less than convincing to readers. See Chapter 8 of the third edition of their book (Haynes et al., 2006) and Chapter 4 of the second edition (Sackett, Haynes, Guyatt, & Tugwell, 1991) for a full discussion of the issues and the quantitative methods involved. The quartet of study design features, although it was developed with medical tests in mind, can be applied to diagnostic studies in speech and language sciences. These provide the basis for the approach taken in this paper. We now briefly describe how each of these features is relevant to diagnostic research involving communication disorders, with examples from assessment of child language disorders.

The first feature requires that for a test or measure to be useful clinically, it should be examined with an appropriate spectrum of clients. Most investigations designed to evaluate the potential of either clinical measures (e.g., standardized tests

or language sample measures) or experimental procedures (e.g., language processing tasks) are conducted by recruiting a group of children who are known to have a particular disorder (e.g., SLI) and then comparing them to a group of children known to be free of the disorder. For example, children selected as possible candidates for an SLI group are almost always recruited from a clinical population by asking clinicians to refer children from their caseloads who meet certain criteria. Typically, the children in these studies have already been assessed, diagnosed, and sometimes even treated. Similarly, children are selected for the control group by recruiting from the general, non-clinical, population of typically-developing children. After this initial screening process, potential candidates for each group are then usually, but not always, tested and selected by virtue of their standardized test scores falling above (for the control group) or below (for the SLI group) a pre-determined cut-off level. While Haynes et al. (2006) suggest that this "may be appropriate while initially exploring the potential of a diagnostic test, [it is] almost certain to be misleading if you are trying to establish test properties for use in clinical practice" (p. 290). So *ultimately*, if a measure is to have real-world clinical value in diagnosing SLI, for example, it should be evaluated using an appropriate spectrum of children who are representative of what might be encountered in a clinical setting, such as individuals who are referred for speech and language evaluation by parents, teachers and doctors. *Ultimately* is used here in its literal sense, to mean that a series of studies can be designed in such a way that the first principle of the diagnostic quartet occurs during the final phase of evaluating a measure's diagnostic accuracy. A hierarchy of research investigations for how this can be achieved is outlined in the next section of the paper (see *Four Phases of Diagnostic Research*).

The second feature of the diagnostic quartet requires that the test or measure (referred to as the *index test*) and the reference standard are applied to all clients recruited to the study. In reviewing the literature on child language disorders, it appears that this condition is usually met. Occasionally however, studies of diagnostic accuracy have been reported in which only children in the affected group (e.g., SLI) received the reference standard while those in the control group (e.g., typically-developing children) did not. Presumably, those in the control group are assumed to be developing normally on the basis of parent or teacher report. The problem here is that unless the reference standard is applied to every individual in the control group as well, one cannot be certain that they are free of the condition.

A further aspect of the second feature concerns the nature of the reference standard used to define the condition (e.g., SLI). As has been argued elsewhere, there is at present no universally-agreed gold standard for defining SLI (Leonard, 1987; Mervis & Robinson, 2005; Tager-Flusberg, 2005). However, Tomblin, Records and which Zhang (1996) have proposed an empirically-determined measure, the epiSLI standard, could be considered a candidate. This is not unlike the situation with respect to a number of medical and psychiatric conditions. In the absence of a gold standard, a *reference standard* can be defined. One approach to this has been to define a cut-off score on a test which is used in conjunction with a set of relevant exclusionary criteria. Although this has the advantage of being replicable, the diagnostic accuracy of the reference test itself may not be known.

The third feature requires that the index test and the reference standard are each interpreted blind to the other. If the diagnostic status of the child is known to the examiner at the time the index test is administered or scored, then the third feature of the diagnostic quartet is compromised. To avoid this, steps must be taken to blind the

examiner as to the diagnostic status of the child at the time the testing or language sample takes place, since *a priori* knowledge of the child's diagnostic status may inadvertently bias the way in which the test is administered or the way the language sample is taken, particularly if the interlocutor during the language sample is the examiner. Steps must also be taken to blind any data coders down the line with respect to the diagnostic status of the child. For example, if a transcriber is overtly aware of the diagnostic status of the child (or the age or sex of the child, if the measure varies as a function of these) at the time the language sample is transcribed and analyzed, this could bias the results in various ways. If the data coder is aware that the child they are listening to, transcribing, or coding is language disordered – and also aware, for example, that such children are likely to omit certain morphological markers in their speech with greater frequency than typically-developing children – that has the potential for biasing the way they 'hear', transcribe, or code the presence or absence of such morphemes. Although having a second transcriber independently re-transcribe some of the data for reliability purposes offers a degree of control, the potential for bias is still there. The only way to minimize bias in this situation is to completely blind the examiner and any data coders with respect to the diagnostic status, age and sex of the child. Authors of diagnostic accuracy studies need to exert the same level of control for bias that authors of randomized controlled trials (RCT) strive for.

The fourth feature requires that the investigation is repeated on a second, independent set of clients. This is discussed in the next section, which presents a hierarchy for doing systematic diagnostic research.

Four Phases of Diagnostic Research

Sackett and Haynes (2002a, b) proposed that diagnostic research might take place in a series of four phases, with each phase designed to answer a specific question. These are outlined below, along with suggestions for how they might be applied to diagnostic studies of communication disorders. As before, while the examples are drawn from child language disorders, the questions can be applied to any clinical population.

Phase I question: Do clients with the target disorder have different test results from normal individuals? This question is asked during the initial stage of the research, where the investigator is trying to find out whether a new measure or procedure has the potential for distinguishing affected from unaffected individuals. In this phase, it is reasonable to recruit children from a known clinical population (e.g., children with SLI who are on a waiting list to receive intervention) and compare them to children from a non-clinical population (e.g., children in a preschool or nursery who are developing language normally). The developmental status of each child would then be confirmed by administering the components making up the reference standard (e.g., standardized language test, audiological screening, and any other measure thought to be relevant to the disorder of interest). The standardized test or experimental measure being investigated (referred to as the *index test*) would then be administered and an appropriate statistical procedure (e.g., a test of mean differences) would be used to determine whether the two groups differed in their performance on the index test. Many studies of the diagnostic accuracy of tests and measures in child language have only ever been evaluated as a phase I question (e.g., Klee, 1992), with the research never progressing to the next phase. An advantage of performing a phase I study is that it is "quick and relatively cheap to carry out, and a negative result saves

having to proceed to the tougher, more time consuming, and costlier questions of phases II-IV" (Sackett and Haynes, 2002b, p. 539).

Phase II question: Are clients with certain test results more likely to have the target disorder than clients with other test results? If the answer to the phase I question was *yes*, the research can proceed to phase II. In this phase, the goal is to establish whether clients with certain results on the index test are more likely to have the disorder than clients below, or above, that level. Sackett and Haynes (2002a, b) indicate that at times the dataset that was used in phase I can also be used in phase II, but that the goal in the second phase is to find the cut-off score that differentiates affected from unaffected individuals with the highest level of accuracy. It is beyond the scope of this paper to review the methods and measures by which this is done, but interested readers are directed to Chapter 4 of Sackett et al. (1991) and Chapter 8 of Haynes et al. (2006). For convenience however, a list of the main outcome measures, along with definitions, is provided in Table 1.

-- Insert Table 1 about here. --

Phase III question: Does the test distinguish clients with and without the target disorder among clients in whom it is clinically reasonable to suspect that the disorder is present? The next phase of the research evaluates whether the test or measure is capable of differentiating those with and without the target disorder in a clinical sample. Previously, phase I will have established that the index test produced different results, on average, in groups with and without the disorder, while phase II will have established the optimal cut-off level for the test. At phase III, the index test would be trialled on a sample of clients who have been referred because of concerns raised by parents, teachers, doctors or others. In the case of an index test for SLI, for example, the referred sample might contain children with a wide range of conditions

associated with speech and language difficulties, including conditions such as cognitive impairment, hearing impairment, and autism spectrum disorders. In the case of an instrument designed to screen children for language delay, the target sample might be either a particular group of individuals known to be at high-risk of having the condition or a group of unselected individuals in the general population who are pre-symptomatic at the time of screening (e.g., Klee, Carson, Gavin, Hall, Kent & Reece, 1998; Laing, Law, Levin & Logan, 2002; Stokes, 1997). Thus, the "correct" target group may differ depending on the intended purpose of the index test.

Phase IV question: Do clients who undergo the diagnostic test fare better (in their ultimate health outcomes) than similar clients who are not tested? The last phase of evaluating an index test poses the difficult question of whether individuals who receive the index test have better (speech, language, or functional) outcomes than those who were not given the test. This question has long been recognized as being part of the process by which screening measures are evaluated (e.g., Cochrane & Holland, 1971; Wilson & Jungner, 1968), but has rarely been asked of assessments in speech and language sciences. This phase of the research will usually require an RCT. Law et al. (2000) presented a model for how this might be done in screening studies and de Koning et al. (2004) conducted a study attempting to do just that.

Critical Appraisal of Studies of Diagnostic Accuracy

In the opening pages of her very readable introduction to evidence-based medicine, Trisha Greenhalgh summarizes five key steps originally proposed by David Sackett and colleagues (e.g., Sackett & Rosenberg, 1995) with regard to developing an evidence-based practice. These are (1) "to convert our information needs into answerable questions (i.e., to formulate the problem); (2) to track down, with

maximum efficiency, the best evidence with which to answer those questions…; (3) to appraise the evidence critically (i.e., weigh it up) to assess its validity (closeness to the truth) and usefulness (clinical applicability); (4) to implement the results of this appraisal in our clinical practice; and (5) to evaluate our performance." (Greenhalgh, 2006, p. 2; see also Straus, 2007). From here, the current paper focuses on the third of these – critically appraising the evidence regarding tests and measures used to assess and diagnose individuals with communication disorders. Critical appraisal involves an assessment of methodological quality (Greenhalgh, 2006); it provides a means for judging the quality of the evidence that will be brought to bear on deciding which clinical assessment tools are worthwhile and which are not.

Several critical appraisal checklists have been developed for evaluating the methodological quality of studies of diagnostic tests and measures. A selection of these follows:

- The website of the Scottish Intercollegiate Guidelines Network (http://www.sign.ac.uk) contains six critical appraisal checklists, with notes on how to use them, for evaluating various types of studies, including systematic reviews and meta-analyses, RCTs, cohort studies, case-control studies, diagnostic studies, and economic evaluations. The checklist for diagnostic studies (Scottish Intercollegiate Guidelines Network, 2004) contains 22 items, including questions relating to internal validity (i.e., methodological quality) and study details, such as sample size and measures of diagnostic accuracy (e.g., sensitivity, specificity, positive predictive value [PPV], negative predictive value [NPV], and likelihood ratios [LR]).
- The website of the Centre for Evidence-Based Medicine at Oxford University (http://www.cebm.net) contains critical appraisal worksheets for systematic

reviews, RCTs, and diagnostic studies. Items on the worksheets are accompanied by notes and calculations to help the user answer the question.

- Greenhalgh (2006) provides an appendix containing checklists for finding, appraising and incorporating evidence into clinical practice. Included in this is a 10-item checklist with questions specifically related to papers about diagnostic or screening tests.
- CADE (*Critical Appraisal of Diagnostic Evidence;* Dollaghan, 2007) is a 13-point checklist for evaluating the methodological quality of diagnostic accuracy studies, including an evaluation of the strength and precision of two study outcome measures (see also Dollaghan, 2004). One particularly useful feature of this checklist is that reviewers are asked to formulate the explicit diagnostic question, called a *Foreground Question,* which was addressed by the study under review. An example of such a question is, *"For identifying preschoolers in need of further evaluation, what is the accuracy of the new screening test as compared with the results of a diagnostic evaluation?"* (Dollaghan, 2007, p. 103).
- QUADAS (*Quality Assessment of Diagnostic Accuracy Studies;* Whiting et al., 2003) is a 14-point checklist that will be discussed further below.

While these checklists have a lot in common with one another, they differ with respect to their breadth of coverage and length. Consequently, clinicians and reviewers may come to different conclusions about the quality of a diagnostic study partly as a consequence of using different checklists. This brings us back to a point made earlier, that evidence-based assessment is the joint product of those who develop assessment measures and those who use them. Despite the availability of a number of different critical appraisal checklists, at times it is difficult to judge the

quality of evidence in published studies because of the way in which they have been reported. Evidence-based assessment can be facilitated if authors aim for complete and accurate reporting of key features in their papers and if clinicians have a means of critically appraising those studies in a consistent and comprehensive way.

Two evidence-based checklists have recently been developed for reporting and critically appraising diagnostic tests and measures. One is aimed at test developers and the other at test users. The STARD checklist (*Standards for Reporting of Diagnostic Accuracy*) and accompanying flow chart were designed "to improve the accuracy and completeness of reporting of studies of diagnostic accuracy" (Bossuyt et al., 2003a, p. 41). To date over 50 journals, including the journals of the American Speech-Language-Hearing Association, require authors of diagnostic studies to use STARD when preparing manuscripts for submission to those journals. The STARD checklist contains 25 items covering key points within the title, abstract, and keywords of the article and the introduction, methods, results, and discussion sections. Authors are also encouraged to provide a flow chart that tracks the sequence and progression of participants through the study. Further information regarding each item on the checklist may be found in a companion paper (Bossuyt et al., 2003b) and on the STARD website (http://www.stard-statement.org). While STARD was developed for authors of diagnostic accuracy studies, it was not intended to be used by readers as a critical appraisal checklist (http://www.stard-statement.org/website%20stard/, accessed December 3, 2007).

The QUADAS tool (Whiting et al., 2003) was designed for appraising the methodological quality of diagnostic tests and measures. This checklist was developed for assessing studies of diagnostic accuracy included in systematic reviews, but it can also be used by clinicians for assessing the methodological quality of individual

studies. For example, QUADAS was used recently to critically appraise a screening study in this journal (Klee, 2007). The QUADAS tool is itself evidence-based, in that its face validity was assessed using a formal consensus method (a Delphi procedure) by a panel of experts after conducting reviews of the diagnostic literature to locate potential items. It was then put through field trials to assess its consistency, construct validity, and usability. A recent empirical evaluation of QUADAS found that the interrater agreement across checklist items was very good (Whiting, Weswood, Rutjes, Reitsma, Bossuyt, & Kleijnen, 2006). The tool contains 14 items and is reproduced in Table 2.

-- Insert Table 2 about here. --

Critical Appraisal using QUADAS: an example

The remainder of this paper will illustrate the use of QUADAS in critically appraising a screening study. The decision to use QUADAS was taken on the basis of the extensive field work that has been undertaken in developing this tool (Whiting, 2006) as well as on the decision to separate methodological quality from study outcomes. Although QUADAS focuses on the former, users may decide to then examine the study's outcomes provided the methodological quality of the study is sufficiently high. As in any research study, a study's results are undermined if methodological quality is not first demonstrated.

To illustrate its use, a screening study is critically appraised (Klee, Pearce, & Carson, 2000). This study involved a further analysis of data originally presented in Klee et al. (1998). In the earlier study, the *Language Development Survey* (LDS; Rescorla, 1998) was mailed to parents at the time of their child's second birthday. The LDS was designed as a screening questionnaire to detect the presence of early

language delay in children. It contains a vocabulary checklist and a question about whether the child has begun to combine words. It also contains questions about the child's sex, birth order, prematurity, number of ear infections, and family history of late talking, among other things. Parents were also asked whether they had any concerns about their child's language, communication or hearing abilities. A total of 306 parents completed the questionnaire. Follow-up clinical evaluations were then offered to the parents of all children who screened positive on the LDS (n = 45), as defined by Rescorla's Delay 3 criterion (<50 words or no word combinations), and a random sample of those who screened negative. The parents of 64 children (17 screen positives; 47 screen negatives) agreed to participate in the follow-up evaluation.

The 2000 study had two goals, but for the purpose of this review, only the first is critically appraised. The first goal of that study was "to explore whether the positive predictive value of the screening program we reported earlier (Klee et al., 1998) could be increased while maintaining the high sensitivity, specificity, and NPV of the original screening criterion. We attempted to refine the pass-fail criterion we originally used (viz., Rescorla's [1989] Delay 3 criterion) by taking into account other information provided by parents on the screening questionnaire" (p. 824).

It is worth mentioning that neither STARD nor QUADAS were available to the authors at the time the study was conducted and written up. The study is evaluated below by answering *Yes*, *No*, or *Unclear* to each of the 14 QUADAS items, following the definitions of each given in Whiting et al. (2003). Each decision is accompanied by supporting evidence which takes the form of a passage and page number from either the 1998 or 2000 paper. For some items, responses are supplemented by a reviewer comment, although this is not strictly part of the tool's protocol.

1. *Was the spectrum of patients representative of the patients who will receive the test in practice?* **No:** "Screening questionnaires...were mailed to 650 families over a period of 25 consecutive months; 582 were successfully delivered, and 306 (53%) were completed and returned by parents" (2000, p. 824). **Reviewer comment:** The spectrum in a screening study is considered to be the population. The spectrum in this study is represented by a biased sample, in that nearly half of the available population did not participate in the screen (screening yield = 53%). While "no information was available regarding those parents who did not return the [screening] survey", several explanations were provided (1998, p. 630).
2. *Were selection criteria clearly described?* **Yes.** "Most children were initially identified from birth announcements in local newspapers. These were then cross-referenced with the local telephone directory to provide current addresses. In addition, postcards were distributed to parents of 18-to 24-month-old children through local physicians' offices, County Public Health offices, the Women Infant and Children (WIC) Program office, and 28 home- and center-based day care facilities. The postcards notified parents of a free speech and language screening and indicated that they could return the postcard if they were interested in having their child screened when the child turned 2. Names, addresses, and birthdates from both these sources were entered into a computer database that was used to generate a list of children to be screened each month" (1998, p. 629).
3. *Is the reference standard likely to correctly classify the target condition?* **Yes.** "Following the clinical evaluation, each child's overall language status was categorized as either normal (LN) or delayed (LD). Categorization was based on the clinical judgement of two independent examiners after reviewing all the data collected during the clinical evaluation. For children categorized as LD, clinical

concern had to be registered by each of the examiners. In addition, the child's performance on at least one of three standardized language measures had to fall one standard deviation (SD) or more below the population mean. These included the LRO *(Language Receptive Organization; italics added by this author)* and LEO *(Language Expressive Organization)* scales of the MSEL *(Mullen Scales of Early Learning)* and mean length of utterance in morphemes.... The LD group consisted of children for whom intervention was recommended...as well as those for whom re-evaluation was recommended in order to monitor development.... For children categorized as LN, no clinical concern was expressed by either examiner, and all three standardized language measures were within normal range (LRO, LEO, and MLU all above -1 SD from the mean" (2000, p. 824). **Reviewer comment:** The target condition was in this case defined as *language delay* rather than language disorder. The reference standard used in this study is in line with other studies of late talkers and is appropriate given that there is no commonly agreed upon gold standard for what constitutes a *developmental language disorder* in toddlers who are otherwise developing normally.

4. *Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?* **Yes.** "A double-blind clinical evaluation was conducted on a sample of the children 1 month after screening..." (2000, p. 824). **Reviewer Comment:** Since the natural history of the target condition is not known at this time, it is possible that some of the children changed in the interval between screening and follow-up (e.g., vocabulary spurt).
5. *Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?* **No.** "We attempted to evaluate all the

children who screened positive (17 of 45) and a random sample of those who screened negative (47 of 261)" (2000, p. 824). **Reviewer comment:** Of those who screened positive, only 38% were given follow-up evaluations: "...some could not be assessed further because of scheduling conflicts and illnesses or because parents either could not be contacted or were unwilling to participate" (1998, pp. 630-1).

6. *Did patients receive the same reference standard regardless of the index test result?* **Yes.** "The clinical evaluation included a parent interview, standardized audiological and development testing (Mullen Scales of Early Learning, MSEL; Mullen, 1993; originally published in Infant and Preschool versions), direct observation of the child with both familiar (parent) and unfamiliar individuals (examiner), and a quantitative analysis of a sample of conversational language during play" (2000, p. 824).
7. *Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?* **Yes.** The index test (LDS) did not form part of the reference standard; see the evidence presented for item 6.
8. *Was the execution of the index test described in sufficient detail to permit replication of the test?* **Yes.** "To screen positive, the child had to meet the Delay 3 criterion *and* the parent either had to indicate concern for the child's language development *or* report that the child had experienced six or more ear infections during the first 2 years of life. This combination of factors [was] referred to as the Delay 3+ criterion..." (2000, p. 826). **Reviewer comment:** The original Delay 3 criterion was defined as less than 50 words of expressive vocabulary or no evidence of word combinations on the LDS (Rescorla, 1989).

9. *Was the execution of the reference standard described in sufficient detail to permit its replication?* **Yes.** Refer to the evidence provided for items 3 and 6.
10. *Were the index test results interpreted without knowledge of the results of the reference standard?* **Yes.** "A double-blind clinical evaluation was conducted...1 month after screening..." (2000, p. 824).
11. *Were the reference standard results interpreted without knowledge of the results of the index test?* **Yes.** "At the time of testing, both the examiners and the child's parent(s) were blind with respect to the screening outcome" (1998, p. 631).
12. *Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?* **Not relevant. Reviewer comment:** The instructions for QUADAS allow for this item to be omitted in cases where "the interpretation of the index test is fully automated and involves no interpretation" (Whiting et al., 2003). Deciding whether the screening outcome was positive or negative is straightforward in the case of the LDS.
13. *Were uninterpretable/intermediate test results reported?* **Yes. Reviewer comment:** No uninterpretable or intermediate test results were reported. A proposed modification to the QUADAS background document indicates that "If it is clear that all test results...are reported, then this item would be scored as 'yes' " (Whiting et al., 2006, Table 4).
14. *Were withdrawals from the study explained?* **Yes. Reviewer comment:** Table 4 of Whiting et al. (2006) also indicates that "If is clear what happened to all patients who entered the study... then this item would be scored as 'yes' ".

Having completed critically appraising this paper using QUADAS and deciding that the methodological quality was sufficiently high to warrant examining the study outcomes, a comment will be made about these. While this study

demonstrated that the revised Delay 3+ criterion resulted in improved point estimates of diagnostic accuracy (specificity, PPV, LRs) relative to the original Delay 3 criterion, these were accompanied by large confidence intervals, probably due to the small size of the clinical follow-up sample (see Table 1). Although the screening results look promising on the basis of the evidence presented, it is far from clear whether they would hold up in a population screening program. This led the authors to warn that "Further investigation involving a larger sample…would allow us to place more confidence in the screening model proposed here" (Klee et al., 2000, p. 831).

Notice that the assessment of methodological quality was done without any reference to the results of the study (which are presented in Table 1). Herein lies another benefit of using QUADAS over some of the other critical appraisal checklists. Unless the methods used in the research are sound to begin with, the results cannot be trusted. Whether the results can be trusted in the case of the study reviewed here will be left up to the reader to decide. A caveat is worth mentioning here. It is possible that since the reviewer also happened to be one of the authors of the paper being reviewed, the review itself is biased. Interested readers are therefore encouraged to do their own review of this study, using QUADAS and the item definitions in Whiting et al. (2006), and compare their evaluation to the one above. Of course, investigators are not likely to critically appraise their own studies in the post-hoc fashion done here, although this has been a worthwhile exercise in self-reflection for this investigator! Use of the STARD checklist by investigators of diagnostic accuracy studies, coupled with rigorous research design, will help ensure that readers can properly evaluate their work when it comes time to using QUADAS for critical appraisal.

Summary and Conclusions

As we wrote in an earlier paper (Klee et al., in press), the deputy editor of the BMJ suggested that "while evidence based treatment is well on the way to being sorted out, evidence based diagnosis is still in the dark ages" (Delamothe, 2006). Although he was referring to diagnosis in medicine, the same could be said of diagnosis in speech-language pathology. Much stands to be gained if authors of tests and measures used to diagnose communication disorders design their investigations so that their diagnostic accuracy can be determined. Further, authors can improve the way in which they communicate the details of their investigations to readers if they follow the STARD guidelines. Finally, clinicians and reviewers can critically appraise the methodological quality of these studies by using the QUADAS tool. Taken together, these three things have the potential for moving the state of the art in speech and language assessment closer to the ideal of an evidence-based practice.

Endnotes

[1] Throughout the paper, the term *clinician* will be used since the professionals who typically work with individuals having communication difficulties are referred differently throughout the world (e.g., *Speech-Language Pathologist* in the USA, *Speech and Language Therapist* in the UK, *Speech Pathologist* in Australia, *Speech Therapist* in Hong Kong). Similarly, the terms *client* and *patient* will be used interchangeably when referring to individuals with speech and language difficulties.

[2] Although the examples throughout the paper will be from child language disorders, the same principles apply to any area of speech-language pathology.

[3] Calculations were carried out to four decimal places and rounded to two at the last stage for the figures reported in Table 1.

References

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

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., et al. (2003a). Towards a complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ (British Medical Journal), 326*, 41-44.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., et al. (2003b). The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clinical Chemistry, 49*, 7-18.

Cochrane, A., & Holland, W. (1971). Validation of screening measures. *British Medical Bulletin, 27*, 3-8.

de Koning, H. J., de Ridder-Sluiter, J. G., van Agt, H. M. E., Reep-van den Bergh, C. M. M., van der Stege, H. A., Korfage, I. J., et al. (2004). A cluster-randomised trial of screening for language disorders in toddlers. *Journal of Medical Screening, 11*, 109-116.

Delamothe, T. (2006). Diagnosis--the next frontier. *BMJ (British Medical Journal), 333*, 0-f.

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

Dollaghan, C. A. (2004). Evidence-based practice in communication disorders: what do we know, and when do we know it? *Journal of Communication Disorders, 37*, 391-400.

Dollaghan, C. A. (2007). Appraising diagnostic evidence. In C. A. Dollaghan (Ed.), *The handbook for evidence-based practice in communication disorders* (pp. 81-104). Baltimore, MD: Paul H. Brookes.

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

Greenhalgh, T. (2006). *How to read a paper: the basics of evidence-based medicine* (3rd ed.). Oxford: Blackwell.

Haynes, R. B., Sackett, D. L., Guyatt, G. H., & Tugwell, P. (2006). *Clinical epidemiology: how to do clinical practice research* (3rd ed.). Philadelphia: Lippincott Williams & Wilkins.

Klee, T. (1992). Developmental and diagnostic characteristics of quantitative measures of children's language production. *Topics in Language Disorders, 12*, 28-41.

Klee, T. (2007). Screening 3-year-olds for language delay using selected parent-report measures: the jury is still out. *Evidence-Based Communication Assessment and Intervention, 1*, 58-59.

Klee, T., Carson, D. K., Gavin, W. J., Hall, L., Kent, A., & Reece, S. (1998). Concurrent and predictive validity of an early language screening program. *Journal of Speech, Language, and Hearing Research, 41*, 627-641.

Klee, T., Pearce, K., & Carson, D. K. (2000). Improving the positive predictive value of screening for developmental language disorder. *Journal of Speech, Language, and Hearing Research, 43*, 821-833.

Klee, T., Wong, A. M.-Y., Stokes, S. F., Fletcher, P., & Leonard, L. B. (in press). Assessing Cantonese-speaking children with language difficulties from the perspective of evidence-based practice: current practice and future directions. In S.-P. Law, B. Weekes & A. M.-Y. Wong (Eds.), *Language disorders in Chinese*.

Knottnerus, J. A., & van Weel, C. (2002). General introduction: evaluation of diagnostic procedures. In J. A. Knottnerus (Ed.), *The evidence base of clinical diagnosis* (pp. 1-17). London: BMJ Books.

Laing, G. J., Law, J., Levin, A., & Logan, S. (2002). Evaluation of a structured test and a parent led method for screening for speech and language problems: prospective population based study. *BMJ (British Medical Journal), 325*, 1152-1156.

Law, J., Boyle, J., Harris, F., Harkness, A., & Nye, C. (2000). The feasibility of universal screening for primary speech and language delay: findings from a systematic review of the literature. *Developmental Medicine & Child Neurology, 42*, 190-200.

Leonard, L. B. (1987). Is specific language impairment a useful construct? In S. Rosenberg (Ed.), *Advances in applied psycholinguistics: disorders of first-language development* (pp. 1-39). Cambridge: Cambridge University Press.

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

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

Mervis, C. B., & Robinson, B. F. (2005). Designing measures for profiling and genotype/phenotype studies of individuals with genetic syndromes or developmental language disorders. *Applied Psycholinguistics, 26*, 41-64.

Nelson, H. D., Nygren, P., Walker, M., & Panoscha, R. (2006). Screening for speech and language delay in preschool children: systematic evidence review for the US Preventive Services Task Force. *Pediatrics, 117*, 298-319.

Oetting, J. B., Cleveland, L. H., & Cope, R. F., III. (2008). Empirically derived combinations of tools and clinical cutoffs: an illustrative case with a sample of culturally/linguistically diverse children. *Language, Speech, and Hearing Services in Schools, 39*, 44-53.

Perona, K., Plante, E., & Vance, R. (2005). Diagnostic accuracy of the Structured Photographic Expressive Language Test: Third Edition (SPELT-3). *American Journal of Speech-Language Pathology, 36*, 103-115.

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

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

Rescorla, L. (1989). The Language Development Survey: A screening tool for delayed language in toddlers. *Journal of Speech and Hearing Disorders, 54*, 587-599.

Sackett, D. L., & Haynes, R. B. (2002a). The architecture of diagnostic research. *BMJ (British Medical Journal), 324*, 539-541.

Sackett, D. L., & Haynes, R. B. (2002b). The architecture of diagnostic research. In J. A. Knottnerus (Ed.), *The evidence base of clinical diagnosis* (pp. 19-38). London: BMJ Books.

Sackett, D. L., Haynes, R. B., Guyatt, G. H., & Tugwell, P. (1991). *Clinical epidemiology: a basic science for clinical medicine* (2nd ed.). Boston: Little, Brown.

Sackett, D. L., & Rosenberg, W. M. C. (1995). The need for evidence-based medicine. *Journal of the Royal Society of Medicine, 88*, 620-624.

Schlosser, R. W., Wendt, O., & Sigafoos, J. (2007). Not all systematic reviews are created equal: considerations for appraisal. *Evidence-Based Communication Assessment and Intervention, 1*, 138-150.

Scottish Intercollegiate Guidelines Network. (2004). Methodology checklist 5: studies of diagnostic accuracy [Electronic Version]. Retrieved November 30, 2007 from http://www.sign.ac.uk/methodology/checklists.html.

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

Stokes, S. F. (1997). Secondary prevention of paediatric language disability: a comparison of parents and nurses as screening agents. *European Journal of Disorders of Communication, 32*, 139-158.

Straus, S. E. (2007). Evidence-based health care: challenges and limitations. *Evidence-Based Communication Assessment and Intervention, 1*, 48 - 51.

Tager-Flusberg, H. (2005). Designing studies to investigate the relationships between genes, environments, and developmental language disorders. *Applied Psycholinguistics, 26*, 29-39.

Tomblin, J. B., Records, N. L., & Zhang, X. (1996). A system for the diagnosis of specific language impairment in kindergarten children. *Journal of Speech and Hearing Research, 39*, 1284-1294.

Whiting, P. (2006). *Quality of diagnostic accuracy studies: the development, use and evaluation of QUADAS.* Unpublished PhD thesis, University of Amsterdam.

Whiting, P., Rutjes, A. W. S., Reitsma, J. B., Bossuyt, P. M. M., & Kleijnen, J. (2003). The development of QUADAS: a tool for the quality of assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology, 3*, 25.

Whiting, P. F., Weswood, M. E., Rutjes, A. W. S., Reitsma, J. B., Bossuyt, P. N. M., & Kleijnen, J. (2006). Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Medical Research Methodology, 6*, 9.

Wilson, J. M. G., & Jungner, G. (1968). *Principles and practice of screening for disease*. Geneva: World Health Organisation.

Table 1

Performance of the LDS as a screening measure for language delay in 24-month-old children, using the revised screening criterion (Klee, Pearce & Carson, 2000).

		Clinical outcome		
		Language delayed	Language normal	Totals
Screening outcome (LDS Delay 3+)	Positive	10	2	12
	Negative	1	51	52
	Totals	11	53	64

Outcome measures (with 95% confidence intervals)[3] and definitions:

Sensitivity = 10/11 = **.91** (.62 - 1.00); proportion of those with the target condition who have a positive test result.

Specificity = 51/53 = **.96** (.87 - . 99); proportion of those without the target condition who have a negative test result.

Positive predictive value (PPV) = 10/12 = **.83** (.55 - . 95); proportion of those with a positive test result who have the target condition.

Negative predictive value (NPV) = 51/52 = **.98** (.90 - 1.00); proportion of those with a negative test result who are free of the target condition.

Likelihood ratio for a positive test (LR+) = .91/(1.00 - .96) = **24.1** (6.1 - 95.0); the number of times a positive screening outcome is likely to occur in those with, as opposed to those without, the target condition.

Likelihood ratio for a negative test (LR-) = (1-.91)/.96 = **0.09** (0.02 - 0.61); the number of times a negative screening outcome is likely to occur in those with, as opposed to those without, the target condition.

Table 2. QUADAS checklist for assessing the quality of studies of diagnostic accuracy.

Item		Yes	No	Unclear
1.	Was the spectrum of patients representative of the patients who will receive the test in practice?	( )	( )	( )
2.	Were selection criteria clearly described?	( )	( )	( )
3.	Is the reference standard likely to correctly classify the target condition?	( )	( )	( )
4.	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	( )	( )	( )
5.	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?	( )	( )	( )
6.	Did patients receive the same reference standard regardless of the index test result?	( )	( )	( )
7.	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	( )	( )	( )
8.	Was the execution of the index test described in sufficient detail to permit replication of the test?	( )	( )	( )
9.	Was the execution of the reference standard described in sufficient detail to permit its replication?	( )	( )	( )
10.	Were the index test results interpreted without knowledge of the results of the reference standard?	( )	( )	( )
11.	Were the reference standard results interpreted without knowledge of the results of the index test?	( )	( )	( )
12.	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	( )	( )	( )
13.	Were uninterpretable/intermediate test results reported?	( )	( )	( )
14.	Were withdrawals from the study explained?	( )	( )	( )

Used with permission. From Whiting et al. (2003). The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology, 3,* 25.