Published in final edited form as:

Int J Lang Commun Disord. 2009 ; 44(5): 600–615. doi:10.1080/13682820802259662.

Identifying language impairment in children: combining language test scores with parental report

Dorothy V. M. Bishop and David McDonald

Department of Experimental Psychology, Oxford Study of Children's Communication Impairments, University of Oxford, Oxford, UK

Abstract

Background—Children who meet language test criteria for specific language impairment (SLI) are not necessarily the same as those who are referred to a speech and language therapist.

Aims—To consider how far this discrepancy reflects insensitivity of traditional language tests to clinically important features of language impairment.

Methods & Procedures—A total of 245 twin children, 52 of whom had been referred to a speech and language therapist for assessment or intervention, were studied. They were given a battery of language tests and their parents completed the Children's Communication Checklist — 2 (CCC-2).

Results—Language tests that stressed verbal short-term memory were best at distinguishing clinically referred from other cases; narrative and vocabulary tasks were less effective. A discriminant function analysis identified a combination of language test and parental report measures as giving the best discrimination between referred and non-referred cases. Nevertheless, of 82 children classified as language impaired by the discriminant function, 44 had never been referred to a speech and language therapist. These did not appear to be false-positives; they scored at least as poorly as referred cases on literacy tests. They had significantly lower socio-economic backgrounds than referred cases.

Conclusions & Implications—Language test scores provide important information about which children are at risk of academic failure, though this varies from test to test. Reliance on language tests alone, however, is insufficient; a parental report provides important complementary information in the diagnostic process. Children of low socio-economic status with language problems are particularly likely to have no contact with speech and language therapist services.

Keywords

specific language impairment; diagnosis; language tests; checklists; parent report

© 2009 Royal College of Speech & Language Therapists

Address correspondence to: Dorothy Bishop, Department of Experimental Psychology, Oxford Study of Children's Communication Impairments, University of Oxford, South Parks Road, Oxford OX1 3UD, UK; dorothy.bishop@psy.ox.ac.uk.

Publisher's Disclaimer: This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Declaration of interest: Dorothy Bishop is the author of both ERRNI and CCC-2, which are evaluated in this report.

What this study adds

It is known that there can be a quite marked mismatch between cases of specific language impairment (SLI) identified via epidemiological screening using language tests, and cases who are identified clinically. The study considered reasons for this disagreement, and whether it might be reduced by using different diagnostic instruments.

In a sample of 245 twin children, we looked at extent of agreement between (1) clinical referral for assessment or treatment by a speech and language therapist; and (2) language impairment as identified by low scores on language tests and a parental report measure, the Children's Communication Checklist — 2. The best statistical discrimination between referred and non-referred cases was obtained when language test and parent report data were combined. Nevertheless, many cases whose scores were indicative of language impairment had not been clinically referred. We draw two main conclusions from this study. First, language tests vary in their sensitivity to clinically important language problems, and parental report using a checklist can provide useful complementary information in the diagnostic process. Second, when objective criteria are used to define language impairment, we find many children with significant difficulties have not had contact with speech and language therapy services, and these are particularly likely to be children of low socio-economic status.

Introduction

Identification of children with specific language impairment (SLI)

Psychometric tests play an important role in the identification of children with language impairments. They allow us to observe aspects of language function in a standardized setting, and to relate performance to normative data. The DSM-IV-TR criteria for specific language disorder (American Psychiatric Association 2000) and the equivalent criteria in ICD-10 (World Health Organization 1993) emphasize the need for diagnosis to be based on standardized individually administered measures of both receptive and expressive language. Diagnostic criteria include (1) a mismatch between language and non-verbal abilities, (2) significant interference with academic achievement or social communication, and (3) an absence of pervasive developmental disorder or other exclusionary criteria. To receive a diagnosis in ICD-10, it is specifically required that there is a score on an expressive or receptive language test that is 2 standard deviations (SD) or more below the population mean, i.e. below the third percentile. However, this cut-off is arbitrary and unvalidated, and there is no specification as to the language test that should be used, despite the fact that there can be marked variation in sensitivity of tests to language impairment (Spaulding *et al.* 2006).

In English-speaking countries there is a range of well-standardized language assessments suitable for children. However, the assumption that these provide the most valid and efficient tools for identifying clinically significant language difficulties has seldom been scrutinized. From time to time, doubts have been voiced when researchers have found a mismatch between those children who are identified as language-impaired on standardized tests, and those who are receiving clinical services. As illustrated in figure 1, such mismatch can work in both directions: for instance, in an epidemiological survey, Tomblin *et al.* (1997) found that only 29% of children who met criteria for SLI on a psychometric test battery had previously been identified as language-impaired. In a survey of children referred for psychiatric services, Cohen *et al.* (1989) reported that language testing revealed a high proportion of hitherto unsuspected language impairment. There are at least two explanations for such cases (corresponding to quadrant B in figure 1). The first is that they correspond to

genuine cases of language impairment who remain unidentified because of poor awareness of language impairment, which may in turn be a consequence of lack of provision of services. An alternative possibility is that these cases are 'false-positives' that arise because of imperfect reliability of the language tests that were used as a basis for diagnosis. In addition, even if language tests are adequately reliable, they may lack validity if children who perform poorly do not give any concern to parents or teachers in everyday life.

Other studies have found cases of children receiving clinical services who do not meet psychometric criteria for language impairment, corresponding to quadrant C in figure 1 (e.g., Dunn *et al.* 1996, Conti-Ramsden *et al.* 1997). A particularly striking example comes from Keegstra *et al.* (2007) who found that 35% of preschool children referred to a speech and language clinic in the Netherlands had adequate language development. These children tended to have parents who had a relatively high educational level, and they concluded that more educated parents were prone to be over-concerned about language development. If this is the explanation, then it would be desirable to screen out such cases from those receiving scarce intervention resources. Another possibility is that children may have had language problems that resolved by the time the diagnostic assessment was undertaken, or that they had difficulties, such as voice disorders, that did not involve language impairment. Dunn *et al.* (1996) and Conti-Ramsden and Botting (1999), however, raised a further possibility, which is that standardized tests may miss clinically significant language problems. Dunn *et al.* (1996) studied pre-schoolers who were clinically identified as having SLI yet who had normal scores on language tests and found that they made more morphosyntactic, semantic and pragmatic errors in spontaneous speech than typically developing children, despite producing utterances of normal length for their age. Conti-Ramsden *et al.* (1997) used both language tests and ratings by teachers or speech and language therapists (SALT) in a cluster analysis designed to identify subgroups of children among 7-year-olds given special educational provision for SLI. One cluster contained children who had little evidence of difficulties on the language tests, but who were regarded as having poor comprehension and pragmatic problems by teachers. Presence of speech problems is another salient factor in determining which children receive speech and language therapy services (Zhang and Tomblin 2000, Bishop and Hayiou-Thomas 2008).

The goal for diagnosis should be to establish diagnostic criteria to minimize the numbers of children who fall into quadrants B and C (figure 1), and to maximize those in quadrants A and D. In epidemiological terms, this is equivalent to optimizing the sensitivity, $D/(B+D)$, and specificity, $A/(A+C)$ of the diagnostic process. To achieve this, we need to consider how far agreement between diagnostic criteria and clinical referral can be improved. This will not necessarily give an optimal diagnostic process — to achieve that we would need also to establish whether those who were referred were the children most likely to benefit from such referral. Nevertheless, a first step is to consider carefully the cut-offs used to identify impairment and the instruments that provide the measurements on which diagnoses are based. In the current paper, we use data from a sample of 9–10-year-old children to address three questions. First, can we improve the match with SALT referral by incorporating parental report measures into our diagnostic criteria? Second, will the inclusion of narrative measures enhance our ability to identify children who are likely to have a clinical referral? Third, what are the characteristics of children who appear language-impaired on test and/or parent report, yet are not referred for services?

Parental report

A role for parental report in language assessment has been long-established for toddlers and preschool children. Dale (1997) noted that parent report may be superior to direct assessment of the child in cases where the child is shy or where the presence of an examiner may distort normal communicative patterns. Parent report has been used much less with

school-aged children, where direct assessment is more feasible. Further, where it has been used, agreement between parental rating and language test scores has not been impressive (Massa *et al.* 2008). Nevertheless, Bishop (1998) suggested that parental report may be better than formal testing to detect communication problems that are relatively rare in occurrence, or which are difficult to elicit in a standardized setting. The Children's Communication Checklist (CCC) was developed in an attempt to formalize professional judgements about aspects of communication that may be important clinically, but are not easy to assess using conventional tests. The emphasis of the CCC was on using teacher ratings of pragmatic aspects of language, but a later version, the CCC-2 (Bishop 2003), covered a broader range of communicative skills, and was designed to be completed by parents.

A problem inherent in a checklist is that informants may vary both in their ability to understand the items and in their subjective interpretations and biases. On the other hand, a checklist has the advantage that one can obtain information about day-to-day communication from someone who knows the child well. Bishop *et al.* (2006) found that CCC ratings were as effective as standardized tests at identifying children at risk of language impairment. They suggested that standardized tests may not only miss clinically important features of language impairment but also might identify some children who did poorly because of weak concentration or motivation, but who do not have specifically linguistic difficulties that interfere with everyday life. Thus they raised the possibility that some language tests may lead to 'false-positive' cases of language impairment, even if they are psychometrically acceptable, simply because they are measuring 'the wrong thing' — deficits that do not interfere with communication in everyday life.

Narrative versus other kinds of language measure

Many of the language tests that are widely used in the assessment of children use highly artificial tasks, such as repeating words or sentences, selecting a picture from an array to match a spoken sentence, or completing a sentence with an appropriately inflected word. In recent years, partly stimulated by need to characterize a heritable phenotype, there has been interest in identifying those tests that work best as 'markers' for language impairment, i.e. that show large effect sizes when clinically identified children are compared with a typically developing group. It is noteworthy that, despite the arguments in favour of more naturalistic measures, the best markers for SLI have been tests that involve repeating non-words, repeating sentences, or generating verb inflections (Conti-Ramsden *et al.* 2001, Conti-Ramsden 2003). However, we do not know whether more natural tasks might be as good as, or even better, markers. Narrative tasks have been proposed as particularly promising because they place multiple demands on the child — to generate utterances that are semantically and syntactically appropriate and well-structured, and to blend these into a coherent discourse. In the current study we compared the performance of established markers of SLI, sentence repetition and non-word repetition, with indices from a narrative task, in terms of how well they distinguished children who had been referred for SALT services from the remainder of the sample.

Outline of study

The current study compared CCC-2 ratings with standardized tests, including a recently developed narrative test, to address the question of how different measures compare in their ability to distinguish those who have had contact with SALT services, and to discover the optimal combination of measures for making this distinction. We investigated this question using data from a cohort of 9–10-year-old children from the Twins Early Development Study (TEDS) (Trouton *et al.* 2002), which included many of the twins who were seen at 6 years of age by Bishop *et al.* (2006), plus an additional subset of poor readers from the

TEDS cohort. A detailed account of ascertainment and assessment of this sample is given by Bishop *et al.* (forthcoming).

Methods

Participants

Same-sex twin pairs from TEDS were recruited for this study when the twins were aged from 9 to 10 years of age. TEDS is a community sample of twins born in England and Wales between 1994 and 1996 (Trouton *et al.* 2002). Bishop *et al.* (2006) assessed 196 twin pairs from the TEDS sample at 6 years of age; this sample included a large number of children who had been at risk of language impairment (LI) at 4 years of age, according to parental report. Of the original 196 twin pairs seen at 6 years, 151 (77%) were seen for the current study at 9–10 years. These included 101 pairs from the LI risk group, and 50 pairs from a low risk comparison group, where neither twin had shown evidence of language difficulties at 4 years of age. An additional 78 twin pairs were selected from the main TEDS cohort on the basis that one or both twins had evidence of reading difficulties (at least 1.33 SD below average) on the Test of Word Reading Efficiency (Torgesen *et al.* 1999) when assessed at 7 years of age (Harlaar *et al.* 2005). Note that for the current analyses, we did not subdivide children according to these selection criteria, but rather pooled the whole sample and considered how different diagnostic criteria related to clinical referral status.

We excluded cases where non-verbal ability was more than 1 SD below the mean (for details of the assessment, see below), or where the language impairment was associated with sensorineural hearing loss, physical handicap, autism, or another syndrome affecting cognitive development. Children who failed a hearing screen when assessed (average hearing threshold for frequencies 500–4000 Hz more than 26 dB in the better ear) were also excluded, as well as families where English was not the only language spoken in the home. Because molecular genetic studies are planned with the TEDS cohort, participants were selected to be white in order to reduce heterogeneity associated with ethnicity. Here we report data only from children whose parents completed the CCC-2. This information was obtained for 134 girls and 131 boys. Ethics approval for this study was obtained from Oxford University's Experimental Psychology Research Ethics Committee.

Although this is a twin sample, we do not report any genetic analyses in this paper. The goal here is to compare different sources of diagnostic information, rather than to consider aetiology of language impairment.

Psychometric assessment

The psychometric tests administered at 9–10 years are shown in table 1. When selecting children falling below a cut-off for language impairment, it is important to have a common standard, not influenced by differences in the normative samples of the different tests. To ensure that standardized scores gave an accurate reflection of statistical abnormality of scores in the base population from which the sample was taken all tests were therefore restandardized with mean=100 and SD=15, using a weighted subgroup of 98 children from the main sample consisting of 88% from the low risk group and 12% from the LI risk group (i.e., reflecting the proportions of these two subgroups in the original population). As discussed by Bishop *et al.* (forthcoming), this normative sample had mean scores on IQ tests very close to the average in the original standardization sample, offering reassurance that they were an average ability sample. Note that the battery includes some reading measures as well as language and non-verbal tests. These were not used to identify language impairment, but were used to determine whether language impairment was associated with poor literacy.

The measure of parental report was the CCC-2 (Bishop 2003). This checklist was based on the original CCC, but with modifications to make it more suitable for parents (for sample items, see table 2). CCC-2 has been standardized on 542 children and young people aged from 4 to 16 years (Norbury *et al.* 2004). On most of the scales, such as speech and syntax, scores approach ceiling by school age, and so the distribution of scores is skewed, though a composite measure, the general communication composite (GCC), gives a distribution that approximates normality. Another composite index, the social interaction deviance composite (SIDC) identifies children whose pragmatic difficulties are disproportionate in relation to their structural language skills (Norbury *et al.* 2004).

Parents were asked to complete the CCC-2 for both twins. Published norms were used to convert CCC-2 scores to standard scores with mean=10 and SD=3. The GCC is computed as the sum of scales A–H. The SIDC is the sum of scales A–D subtracted from the sum of E–J, so that a negative score indicates disproportionate problems with pragmatic and social aspects of communication, and a positive score indicates disproportionate difficulties with non-pragmatic aspects of language (speech, syntax, and semantics).

Categorization of speech and language status

Close to the time of the 9-year-old assessment, parents were asked if their child had been referred to a SALT, and/or had a statement of special educational needs. Comparable information was available from an earlier wave of data collection at 7 years. These sources were combined, with children categorized into three groups: the not referred (NR) group, consisting of those who had no contact with SALT services ($n=193$) the clinically referred (CR) group, i.e., those who had been assessed or treated by a SALT ($n=52$), and the other group, those who had a statement of special educational needs (SEN), but with no indication of a speech and language problem ($n=20$). The latter group contained children with a variety of conditions such as attention deficit hyperactivity disorder, and they were excluded from further consideration.

Results

Language tests and CCC-2 subtests in relation to speech and language status

Table 3 shows mean scores on language tests and CCC-2 scales for the NR and CR groups. The interest here is in the size of difference of mean scores of the two groups, which is expressed as Cohen's d , a standardized measure of effect size. Because data from two members of a twin pair are not independent, multilevel modelling was used to compute adjusted F ratios when comparing means. This allows the degrees of freedom to be adjusted to take into account the extent of correlation between twins (Kenny *et al.* 2006).

For the language battery and for the CCC-2 subscales, there is wide variation from one measure to another in terms of how well they discriminate between the two groups of children. Cohen's d , the mean standardized difference between the two groups, directly reflects the overlap between groups. An effect size of 0.5 is regarded as medium in size, and corresponds to a situation where there is 33% non-overlap between groups on the measure in question. For the language battery, effect sizes greater than 0.5 were seen for understanding directions, memory for names, non-word repetition, and oromotor skills. The measures from the narrative task, ERRNI, and the Vocabulary measure gave weaker effects.

Using similar criteria for the CCC-2, there were seven scales that gave effect sizes greater than 0.5: Speech, Syntax, Semantics, Coherence, Stereotyped Language, Use of Context and Social Relations. The composite measure, GCC, gave the largest effect size, 0.804. The SIDC is an index of mismatch between pragmatic impairment and structural language skills and tends to be negative in children with features of autistic spectrum disorder. Children in

the CR group had significantly higher scores than those in the NR group on the SIDC, indicating that they tended to have more difficulty with structural aspects of language than with pragmatic and social communication. Note, however, that children with autistic spectrum disorders had been excluded from the sample, and relatively few children had negative SIDC scores suggestive of pragmatic language impairment (4% scored -15 or less).

Identifying optimal criteria for distinguishing clinically referred versus not referred cases

We next used a stepwise discriminant function analysis (DFA) to identify the best combination of measures for allocating children to CR and NR groups. This is a procedure that considers how accurately children can be allocated to CR and NR groups on the basis of their scores; the measure that gives the best discrimination is entered first, and then the next best is added, and so on, until there is no improvement in the discrimination. The procedure will not necessarily use the measures with the largest effect sizes, because it also takes into account the extent to which measures overlap; thus two measures which are weakly correlated but each with moderate effect size might give better discrimination than two other measures that have large effect sizes but are strongly intercorrelated. All the psychometric test scores from table 3 were entered in the analysis, plus the GCC and SIDC. The individual CCC-2 scales were not included because they are less reliable than the composites and not independent of them. When describing results from the DFA, we shall refer to the child's *predicted* status from the discriminant function as LI or unimpaired.

The optimal discriminant function included two language test measures: NEPSY Oromotor and ERRNI story recall, and the two CCC-2 indices: GCC and SIDC. For the function with all four variables Wilks's lambda (a measure of the proportion of variance in the four variables that is unaccounted for by group) was 0.821, d.f.=4, $p<0.001$. The standardized discriminant function was:

$$0.469 \times \text{oromotor } z + 0.351 \times \text{ERRNI story recall } z + 0.475 \times \text{GCC } z - 0.383 \times \text{SIDC } z$$

Those with a discriminant score below -0.34 were categorized as predicted LI.

Although the classification was significantly better than chance, overall the amount of agreement between predicted status and obtained referral category was not impressive, with only 72.2% agreement. Correctly classified cases included 149 children in the NR group who were predicted as unimpaired on the DFA (corresponding to quadrant A in figure 1) and 38 in the CR group who were predicted as LI (quadrant D). Misclassified cases included 44 cases who were in group NR but predicted as having LI (quadrant B), and 14 cases predicted as unimpaired who were in the CR group (quadrant C). As shown in table 4, this translates to a specificity of 0.91 and sensitivity of 0.46.

In addition to the discriminant analysis, specificity and sensitivity were calculated using simpler diagnostic criteria based on ICD-10, using a 2 SD cut-off on the language measures and/or CCC-2 subtests (table 4). This shows the numbers of children categorized as LI on the basis of having (1) at least one language test score less than -2 SD below the mean; (2) at least one CCC-2 scale less than -2 SD below the mean; (3) having both a language test and a CCC-2 scale -2 SD below the mean; and (4) having either a language test or a CCC-2 scale -2 SD below the mean. As with the discriminant analysis, specificity was highest when a combination of the two types of measures was used, but the overall agreement between test classification and clinical referral status was not impressive and sensitivity was poor, i.e., many children identified as cases of LI were not in the CR group. Other less extreme cut-offs were explored, but did not improve the agreement between referral and diagnostic categorization, because they led to reduced sensitivity (i.e. more NR cases

identified as LI). The best combination of sensitivity and specificity was obtained when LI was defined as having at least one measure from both the language battery and the CCC-2 subtests lower than -2 SD below the mean.

Children who appear language-impaired but are not clinically referred: true cases of impairment or false-positives?

We found significant numbers of children who appeared to have LI on test criteria (either from the discriminant function or from using ICD-10 criteria) but who have not been referred, i.e. those in quadrant B in figure 1. A key question is whether they should be regarded as 'false-positives', i.e. cases who would not merit intervention, or whether they are children who have difficulties comparable to LI children in the CR group (quadrant D) but who have for some reason failed to access clinical services. A possibility is that these children simply have less severe problems. One way to address this is to consider whether such children meet the additional criterion of having poor academic achievement. Table 5 shows scores on the four literacy measures for children who fell in the four quadrants, using the DFA classification for predicted LI status. One-way multilevel analyses of variance (ANOVAs) were used to compare children falling in the four quadrants on the four literacy measures, with Sidak tests for post-hoc comparisons. Highly significant group differences ($p < 0.001$) were seen on all reading tests. On the first three reading measures, the same pattern was observed: the two groups of children identified as LI on the DFA did worse than the non-referred, unimpaired children but did not differ from one another. On reading comprehension, the non-referred LI children scored significantly lower than all other groups, and the referred LI children scored lower than the non-referred, unimpaired children.

These data support the conclusions that (1) the discriminant function used here identifies children whose language problems affect academic achievement and (2) children who were identified as LI on the discriminant function but who had not been referred for SALT services were not those with relatively mild problems — on the contrary, on a test of reading comprehension, they scored lower than LI children who had been referred.

Given that the severity of the child's problems does not appear to be a factor determining SALT referral, we considered three other variables: non-verbal IQ, socio-economic background, and gender. For each variable, we selected all children who were identified as LI on the discriminant function, and compared the CR and NR cases. These two groups did not differ on non-verbal IQ: mean for NR=99.8, SD=10.29, mean for CR=104.2, SD=12.00, $F(1, 80)=3.12$, $p=0.08$. A measure of socio-economic background was available for 72 children; this was a standard score based on occupational status and educational qualifications of parents and age of the mother at birth of the eldest children (Petrill *et al.* 2004). This measure differed significantly for CR and NR cases: mean for CR=-0.02, SD=0.78, mean for NR=-0.40, SD=0.63, $F(1, 70)=5.52$, $p=0.022$. Thus for children whose language scores and CCC-2 ratings were indicative of LI, those of lower socio-economic status were less likely to be clinically referred. There was a trend for a higher proportion of girls to be in the NR group (21 of 44=47%) than in the CR group (11 of 38=29%), but this did not reach statistical significance ($\chi^2=3.02$, $p=0.08$).

Discussion

The preliminary analysis, reported in table 3, compared effect sizes of different measures when comparing mean scores for referred versus non-referred children. Overall, these effect sizes were small in comparison to those reported by Spaulding *et al.* (2006) in their survey of published data on validity of standardized language tests. However, this could reflect the fact that our validation sample was identified in terms of referral for assessment or treatment by a speech and language therapist (SALT), whereas validation samples in test manuals

typically focus on children for whom a diagnosis of specific language impairment (SLI) has been established using standardized tests.

Among the language tests, the largest effect sizes were found for repeating non-words, understanding directions, oromotor skills and memory for names. The finding on the repeating non-words test is consistent with previous studies showing this to be as a particularly good marker of language impairment (Conti-Ramsden *et al.* 2001). It is noteworthy that the best markers in this study all placed demands on verbal short-term memory, although non-word repetition and oromotor skills also make demands on speech production. This is consistent with findings of Zhang and Tomblin (2000) and Bishop and Hayiou-Thomas (2008) that children with disorders affecting speech production are especially likely to be identified clinically, perhaps because their difficulties are particularly obvious in everyday situations. Another factor that may be important is the extent to which a test requires the child to process novel sound sequences, as this will minimize the role of prior learning on performance: this applies to both non-word repetition and oromotor sequences. This study, then confirms the effectiveness of some of the markers for SLI that have previously been identified, although it should be noted that we did not include a measure of verb morphology, which has also been shown to be a sensitive marker (Conti-Ramsden *et al.* 2001) and seems largely independent of phonological short-term memory (Bishop *et al.* 2006).

The narrative measures derived from ERRNI were among the weakest predictors of referral to a SALT. There were differences between the clinically referred and other children on these measures, but the effect size was relatively small. This was a surprising finding, as one might have expected that a narrative task would provide a particularly good index of language ability, because it requires the child to integrate semantic, syntactic and discourse skills, and previous research has found narrative generation to be more sensitive than standardized language tests in distinguishing language impaired from typically developing children (Fey *et al.* 2004). Furthermore, a narrative task, while not totally naturalistic, is closer to 'real-world' use of language than tests that involve repeating sentences, adding morphological endings to words, or pointing to pictures. Nevertheless, for children of this age, given the particular framework provided by ERRNI (that of telling a story from a series of pictures), narrative indices were less good than other measures at identifying children's referral status. This may be because this narrative task does not 'stretch' children sufficiently, in that they can choose how to tell the story and there are no right or wrong answers. The lack of sensitivity may arise if children *without* language impairments tend to use simpler language and include fewer story details than they are capable of. Some narrative tests require the child to retell a story that has been told by an adult — such tasks have been shown to be good predictors of outcome in SLI (e.g., Botting *et al.* 2001) and it is possible that this would be a more sensitive task because it sets a standard that the child's retold narrative has to match. Narrative tasks may also be more discriminating in younger children, whose narrative skills are still developing (cf. Newman and McGregor 2006). Furthermore, measures of grammatical errors (which were not included in our narrative task) may be more useful than measures of content or complexity (cf. Liles *et al.* 1995). Nevertheless, although table 3 shows only modest effect sizes for ERRNI indices, it is noteworthy that the story recall index was one of only four measures to enter the discriminant function. This reflects the fact that this test provides distinctive information about group status that was not captured by the other measures.

Parent report indices from the CCC-2 did as well or better than psychometric tests in distinguishing groups. Furthermore, the discriminant function analysis indicated that information from parental report complemented test results, so that a function that combined both sources of information was most effective in predicting referral status.

A study like this raises thorny issues about diagnostic criteria. In particular, it is clear that we lack a 'gold standard' for the diagnosis of SLI. We found that over half the children who met a criterion for language impairment — either based on ICD-10 criteria, or using the DFA function — had not been referred for assessment or treatment by a SALT. In this regard our results are comparable to those of Tomblin *et al.* (1997). These children (meeting psychometric criteria but not clinically referred) correspond to quadrant B in figure 1. The question arises as to whether they *should* be a focus of clinical attention. Psychometric tests are not perfectly reliable, and some children will score below the cut-off for impairment on one occasion but not on another, and may be regarded as 'false-positives' rather than genuine cases for concern. Another possibility is that children who are not referred as in need of services could be those who do not meet the DSM-IV-TR criterion of there being 'significant interference with academic achievement or social communication'. This explanation is suggested by a study of children referred for psychiatric services by Cohen *et al.* (1998); they found that children whose language impairments had gone undetected tended to have better academic achievement than those with identified language impairments. However, in our study, non-referred children with LI had literacy problems that were at least as bad as those of children in quadrant D, whose problems had been identified as meriting referral. Further analyses suggested that children who had language problems that were not identified were more likely than identified cases to be of lower socio-economic status. It is unclear whether this reflects a lack of provision of SALT services in less affluent regions, a reluctance of some parents to seek help for their child's communication problems, or a lack of concern about such difficulties.

Quadrant C contains children who show the opposite pattern of mismatch between LI status and referral, i.e. they were referred for SALT yet did not appear impaired on test criteria. As noted in figure 1, there are several possible reasons for such mismatch. A study by Broomfield and Dodd (2004) found that in a group of mostly preschool children referred for speech and language therapy assessment, just under 9% had voice or fluency disorders (which would not be identified by our assessment methods), and just under 12% were regarded as having normal speech and language (presumably reflecting either a resolved problem, or parental over-concern). To the extent that such cases are included in our sample, they will reduce the specificity of the diagnostic criteria.

In sum, this study has shown that there are two major problems with the conventional approach to diagnosis enshrined in DSM-IV and ICD-10, with its emphasis on use of individually administered language tests. First, it is evident that tests differ widely in their ability to discriminate referred from non-referred cases, and second, it is clear that language tests may miss features of language impairment that are of clinical importance. We suggest that integration of information from parental report with data from language tests that act as good markers could provide a basis for defining objective diagnostic criteria that would agree better with clinical judgement. Nevertheless, even when parental report is incorporated in diagnostic criteria, it is evident that in current practice, many children with clinically significant language impairments are not identified as in need of SALT services.

Acknowledgments

All royalties from these assessments are donated to charity. The authors alone are responsible for the content and writing of the paper.

References

- American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 4th edn. American Psychiatric Association; Washington, DC: 2000. text revn

- Bishop, DVM. The Children's Communication Checklist, Version 2 (CCC-2). Psychological Corporation; London: 2003.
- Bishop, DVM. Expression, Reception and Recall of Narrative Instrument ERRNI. Psychological Corporation; London: 2004.
- Bishop DVM, Hayiou-Thomas ME. Heritability of specific language impairment depends on diagnostic criteria. *Genes, Brain and Behavior*. 2008; 7:365–372.
- Bishop DVM, Laws G, Adams C, Norbury CF. High heritability of speech and language impairments in 6-year-old twins demonstrated using parent and teacher report. *Behavior Genetics*. 2006; 36:173–184. [PubMed: 16485179]
- Bishop DVM, McDonald D, Bird S, Hayiou-Thomas ME. Children who read accurately despite language impairment: who are they and how do they do it? *Child Development*. forthcoming.
- Botting N, Faragher B, Simkin Z, Knox E, Conti-Ramsden G. Predicting pathways of specific language impairment: what differentiates good and poor outcome? *Journal of Child Psychology and Psychiatry*. 2001; 42:1013–1020. [PubMed: 11806682]
- Broomfield J, Dodd B. Children with speech and language disability: caseload characteristics. *International Journal of Language and Communication Disorders*. 2004; 39:303–324. [PubMed: 15204443]
- Cohen NJ, Barwick MA, Horodezky NB, Vallance DD, Im N. Language, achievement, and cognitive processing in psychiatrically disturbed children with previously identified and unsuspected language impairments. *Journal of Child Psychology and Psychiatry*. 1998; 39:865–877. [PubMed: 9758195]
- Cohen NJ, Davine M, Kelly-Meloche M. Prevalence of unsuspected language disorders in a child psychiatric population. *Journal of the American Academy of Child and Adolescent Psychiatry*. 1989; 28:107–111. [PubMed: 2914822]
- Conti-Ramsden G. Processing and linguistic markers in young children with specific language impairment. *Journal of Speech, Language and Hearing Research*. 2003; 46:1029–1037.
- Conti-Ramsden G, Botting N. Characteristics of children attending language units in England: a national study of 7-year-olds. *International Journal of Language and Communication Disorders*. 1999; 34:359–366. [PubMed: 10884906]
- Conti-Ramsden G, Botting N, Faragher B. Psycholinguistic markers for specific language impairment SLI. *Journal of Child Psychology and Psychiatry*. 2001; 42:741–748. [PubMed: 11583246]
- Conti-Ramsden G, Crutchley A, Botting N. The extent to which psychometric tests differentiate subgroups of children with SLI. *Journal of Speech Language and Hearing Research*. 1997; 40:765–777.
- Dale, PS. Parent report assessment of language and communication. In: Cole, KN.; Dale, PS.; Thal, DJ., editors. *Assessment of Communication and Language*. Paul H. Brookes; Baltimore, MD: 1997. p. 161–182.
- Dunn M, Flax J, Sliwinski M, Aram D. The use of spontaneous language measures as criteria for identifying children with specific language impairment: an attempt to reconcile clinical and research findings. *Journal of Speech and Hearing Research*. 1996; 39:643–654. [PubMed: 8783141]
- Fey ME, Catts HW, Proctor-Williams K, Tomblin JB, Zhang X. Oral and written story composition skills of children with language impairment. *Journal of Speech, Language, and Hearing Research*. 2004; 47:1301–1317.
- Harlaar N, Spinath FM, Dale PS, Plomin R. Genetic influences on early word recognition abilities and disabilities: a study of 7-year-old twins. *Journal of Child Psychology and Psychiatry*. 2005; 46:373–384. [PubMed: 15819646]
- Keegstra AL, Knijff WA, Post WJ, Goorhuis-Brouwer SM. Children with language problems in a speech and hearing clinic: background variables and extent of language problems. *International Journal of Pediatric Otorhinolaryngology*. 2007; 71:815–821. [PubMed: 17353056]
- Kenny, DA.; Kashy, DA.; Cook, WL. *Dyadic Data Analysis*. Guilford; New York, NY: 2006.
- Korkman, M.; Kirk, U.; Kemp, SI. *NEPSY: A Developmental Neuropsychological Assessment*. Psychological Corporation; San Antonio, TX: 1998.

- Liles BZ, Duffy RJ, Merritt DD, Purcell SL. Measurement of narrative discourse ability in children with language disorders. *Journal of Speech and Hearing Research*. 1995; 38:415–425. [PubMed: 7596107]
- Massa J, Gomes H, Tartter V, Wolfson V, Halperin JM. Concordance rates between parent and teacher Clinical Evaluation of Language Fundamentals Observational Rating Scale. *International Journal of Language and Communication Disorders*. 2008; 43:99–110. [PubMed: 17852530]
- Neale, MD. *Neale Analysis of Reading Ability*. 2nd edn. NFER-Nelson; Windsor: 1997.
- Newman RM, McGregor KK. Teachers and laypersons discern quality differences between narratives produced by children with or without SLI. *Journal of Speech, Language and Hearing Research*. 2006; 49:1022–1036.
- Norbury CF, Nash M, Bishop DVM, Baird G. Using parental checklists to identify diagnostic groups in children with communication impairment: a validation of the Children's Communication Checklist — 2. *International Journal of Language and Communication Disorders*. 2004; 39:345–364. [PubMed: 15204445]
- Petrill SA, Pike A, Price T, Plomin R. Chaos in the home and socio-economic status are associated with cognitive development in early childhood: Environmental mediators identified in a genetic design. *Intelligence*. 2004; 32:445–460.
- Spaulding TJ, Plante E, Farinella KA. Eligibility criteria for language impairment: is the low end of normal always appropriate? *Language, Speech and Hearing Services in Schools*. 2006; 37:61–72.
- Tomblin JB, Records NL, Buckwalter P, Zhang X, Smith E, O'Brien M. Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research*. 1997; 40:1245–1260.
- Torgesen, JK.; Wagner, R.; Rashotte, C. *Test of Word Reading Efficiency TOWRE*. Psychological Corporation; New York, NY: 1999.
- Trouton A, Spinath FM, Plomin R. Twins Early Development Study TEDS: a multivariate, longitudinal genetic investigation of language, cognition and behaviour problems in childhood. *Twin Research*. 2002; 5:444–448. [PubMed: 12537874]
- Wechsler, D. *Wechsler Abbreviated Scale of Intelligence*. Psychological Corporation; San Antonio, TX: 1999.
- Woodcock, RW.; McGrew, KS.; Mather, N. *Woodcock Johnson III*. Riverside; Itasca, IL: 2001.
- World Health Organization. *The ICD-10 Classification for Mental and Behavioural Disorders: Diagnostic Criteria for Research*. World Health Organization; Geneva: 1993.
- Zhang X, Tomblin JB. The association of intervention receipt with speech and language profiles and social-demographic variables. *American Journal of Speech–Language Pathology*. 2000; 9:345–357.

		Meets diagnostic criteria	
		NO	YES
Referred to speech- language therapy services	NO	A truly unaffected	B false positive OR true case unidentified by services
	YES	C over-concern OR resolved problem OR problem missed by diagnostic method	D truly affected

Figure 1.
Relationship between diagnostic identification and speech and language therapy referral.

Table 1

Language, literacy and ability tests administered at 9–10 years

Instrument	Content
Wechsler Abbreviated Scale of Intelligence (WASI) Block Design (non-verbal) (Wechsler 1999)	Match pictured patterns using blocks
WASI Vocabulary (Wechsler 1999)	Provide definitions for spoken words
Woodcock–Johnson III: Understanding Directions subtest (Woodcock <i>et al.</i> 2001)	Carry out verbal instructions of increasing complexity
Expression, Reception and Recall of Narrative Instrument (ERRNI) (Bishop 2004)	Tell a narrative from pictures, answer questions about it, and recall it from memory. Provides indices of semantic content from story telling and story recall, a measure of mean length of utterance in words (MLU), and a comprehension score
<i>NEPSY</i> (Korkman <i>et al.</i> 1998)	
Sentence repetition	Repeat sentences of increasing length and complexity
Non-word repetition	Repeat meaningless sequences of two to five syllables
Oromotor sequences	Accurately repeat tongue-twisters
Memory for names	Recall name–face associations, immediately and after a delay
Test of Word Reading Efficiency (TOWRE) (Torgesen <i>et al.</i> 1999)	Rapidly read real words and non-words
Neale Analysis of Reading Ability (NARA-II): stories 1–4 (Neale 1997)	Accuracy and comprehension for passage reading

Table 2

CCC-2 scales with sample items

Scale	Sample item ^a
A. Speech	Pronounces words in a babyish way, such as 'chimbley' for 'chimney' or 'bokkle' for 'bottle'
B. Syntax	(+) Produces sentences containing 'because' such as 'John had a cake because it was his birthday'
C. Semantics	Mixes up words that sound similar, e.g. might say 'telephone' for 'television' or 'magician' for 'musician'
D. Coherence	It is hard to make sense of what s/he is saying (even though the words are clearly spoken)
E. Inappropriate initiation	Talks repetitively about things that no one is interested in
F. Stereotyped language	Repeats back what others have just said. For instance, if you ask, 'What did you eat?', s/he might say, 'What did I eat?'
G. Use of context	(+) Appreciates the humour expressed by irony. Would be amused rather than confused if someone said 'Isn't it a lovely day!' when it is pouring with rain
H. Non-verbal communication	Does not look at the person s/he is talking to
I. Social relations	With familiar adults s/he seems inattentive, distant or preoccupied
J. Interests	Moves the conversation to a favourite topic, even if others do not seem interested in it

^a Each item is rated as being observed: 0, less than once a week (or never); 1, at least once a week, but not every day; 2, once or twice a day; and 3, several times (more than twice) a day (or always). Items marked (+) are reverse scored.

Table 3

Mean (standard deviation) scores for clinically referred (CR) and not referred (NR) subgroups, with effect size (d) and analysis of variance (ANOVA) statistics

	NR ($n=193$)	CR ($n=52$)	Cohen's d	F ratio	$d.f.^2$	p
Block design	104.7 (12.41)	102.6 (11.76)	0.172	1.17	219.6	0.281
Vocabulary	96.7 (16.12)	89.9 (17.37)	0.410	4.35	232.4	0.038
Understanding directions	99.6 (14.74)	90.7 (17.57)	0.562	15.27	220.9	<0.001
ERRNI tell	100.3 (14.83)	95.6 (12.80)	0.325	3.55	202.5	0.061
ERRNI recall	99.6 (15.00)	93.2 (14.22)	0.426	6.33	186.2	0.013
ERRNI comp.	98.9 (15.41)	92.0 (15.36)	0.437	7.10	203.4	0.008
ERRNI MLU	100.1 (17.08)	93.4 (17.30)	0.382	4.94	189.4	0.027
Sentence repetition	96.7 (14.84)	88.6 (18.52)	0.508	8.64	237.7	0.004
Non-word repetition	97.4 (14.56)	87.2 (13.69)	0.684	15.20	212.9	<0.001
Oromotor	96.7 (15.54)	84.6 (16.42)	0.736	14.82	222.6	<0.001
Memory for names	100.4 (14.52)	92.0 (16.20)	0.551	10.49	212.9	0.001
<i>CCC-2</i>						
A. Speech	10.1 (2.93)	8.1 (3.26)	0.642	19.08	199.2	<0.001
B. Syntax	10.3 (2.79)	7.8 (3.64)	0.781	26.19	222.7	<0.001
C. Semantics	9.5 (3.59)	6.7 (3.28)	0.744	16.91	241.9	<0.001
D. Coherence	10.0 (2.96)	7.6 (3.14)	0.748	24.17	228.7	<0.001
E. Inappropriate initiation	9.5 (3.39)	8.4 (3.13)	0.344	4.52	240.9	0.035
F. Stereotyped	10.4 (2.89)	8.8 (2.78)	0.543	16.89	229.0	<0.001
G. Use of context	9.2 (3.55)	7.0 (2.9)	0.617	20.9	241.8	<0.001
H. Non-verbal	10.2 (3.02)	8.5 (3.01)	0.566	19.52	241.7	<0.001
I. Social relations	9.7 (2.78)	8.2 (2.76)	0.512	6.87	214.7	0.009
J. Interests	9.2 (3.06)	8.4 (2.75)	0.284	0.57	240.7	0.452
GCC	79.2 (19.51)	63.0 (17.26)	0.805	35.95	243.0	<0.001
SIDC	-1.2 (7.15)	3.1 (9.5)	-0.544	12.82	206.3	<0.001

GCC, general communication composite; SIDC, social interaction deviance composite.

Table 4

Sensitivity and specificity of diagnostic criteria in relation to actual referral status

Referral status: Meeting LI criterion (quadrant)	Measures	NR,		CR,		Specificity	Sensitivity	Optimality index*
		no (A)	yes (B)	no (C)	yes (D)			
Discriminant function	See the text	149	44	14	38	0.91	0.46	0.46
One measure lower than -2 SD	Language tests	159	34	25	27	0.86	0.44	0.43
	CCC-2 scales	155	38	23	29	0.87	0.43	0.42
	Both	181	12	35	17	0.84	0.59	0.56
	Either	133	60	13	39	0.91	0.39	0.39

* With sensitivity plotted versus specificity, it is computed as 1 minus the length of the vector from perfect sensitivity and specificity, i.e. the higher the optimality index, the closer the agreement with perfect sensitivity and specificity.

CR, clinically referred; NR, not referred; SD, standard deviation.

Table 5

Mean (standard deviation) scores on literacy tests in relation to which criteria met for language impairment, for clinically referred (CR) and not referred (NR) groups

Referral × DFA LI categorization	TOWRE words	TOWRE non-words	NARA accuracy	NARA comprehension
<i>Not referred group (NR)</i>				
A. Unimpaired on DFA (<i>n</i> =149)	98.3 (10.25)	99.1 (16.43)	98.8 (15.92)	100.0 (13.38)
B. LI on DFA (<i>n</i> =44)	81.0 (12.74)	80.5 (13.92)	78.8 (11.55)	81.1 (7.77)
<i>Clinically referred group (CR)</i>				
C. Unimpaired on DFA (<i>n</i> =14)	97.5 (19.15)	98.8 (17.15)	96.7 (19.84)	97.0 (16.57)
D. LI on DFA (<i>n</i> =38)	86.9 (15.57)	86.7 (17.12)	86.4 (15.64)	90.2 (13.73)

DFA, discriminant function analysis; LI, language impairment.