INT. J. LANG. COMM. DIS., SEPTEMBER–OCTOBER 2008,
VOL. 43, NO. 5, 570–594

informa
healthcare

A diagnostically promising technique for tallying nominal reference errors in the narratives of school-aged children with Foetal Alcohol Spectrum Disorders (FASD)

John C. Thorne and Truman Coggins

Department of Speech and Hearing Sciences, University of Washington, Seattle, WA, USA

(*Received 22 September 2006; accepted 21 September 2007*)

Abstract

Background: Foetal Alcohol Spectrum Disorders (FASD) include the range of disabilities that occur in children exposed to alcohol during pregnancy, with Foetal Alcohol Syndrome (FAS) on the severe end of the spectrum. Clinical research has documented a range of cognitive, social, and communication deficits in FASD and it indicates the need for diagnostic tools that can identify children with diminished communicative capacities resulting from prenatal alcohol exposure. Previous research indicates that analysis of nominal reference errors within narrative discourse may provide such a tool.

Aims: To demonstrate the potential diagnostic utility of a new tool for tallying nominal reference errors in the oral narratives of school-aged children with FASD by presenting quantitative measurement data that address interrater agreement and predictive accuracy.

Methods & Procedures: Retrospective analysis was conducted on spontaneously produced oral narratives from 32 school-aged children (8;5–11;7) with a range of socio-economic and ethnic profiles. Sixteen of the children had been previously diagnosis with an FASD, including five with full or partial FAS (pFAS). The remaining 16 children were considered typically developing (TD). A range of methods for calculating the rate of nominal reference errors (rNRE) were used to predict which narratives were produced by children from each group. Accuracy (sensitivity and specificity) for two predictions (FASD versus TD, and FAS/pFAS versus all others) was quantified using receiver-operating characteristic curve analyses. Pairwise statistical comparisons were made between methods to determine which had the most diagnostic potential.

Outcomes & Results: The proposed system for calculating the rNRE was highly accurate at predicting which narratives were produced by children with FASD

Address correspondence to: John C. Thorne, Department of Speech and Hearing Sciences, University of Washington, 1417 NE 42nd Street, Box 354875, Seattle, WA 98105-6246, USA; e-mail: jct6@u.washington.edu

International Journal of Language & Communication Disorders
ISSN 1368-2822 print/ISSN 1460-6984 online
http://www.informaworld.com
DOI: 10.1080/13682820701698960

(versus TD, 88% overall accuracy), and which were produced by children with FAS/pFAS (versus all others, 97% overall accuracy), and outperformed all other methods tested. Agreement on coding decisions between independent judges was high (kappa=0.90).
Conclusions: The strong predictive accuracy demonstrated in this study provides empirical evidence that the system proposed in this feasibility study has sufficient sensitivity and diagnostic utility to warrant further development for use with children suspected of prenatal alcohol exposure. It also points to the potential for the tool to be used with other clinical populations that, even in the absence of a confirmed alcohol exposure, share many of the communication challenges of this complex clinical population.

Keywords: Foetal alcohol spectrum disorders, foetal alcohol syndrome, discourse analysis, diagnostic assessment, language disorders.

What this paper adds

What is already known

Previous research of narrative production has indicated that school-aged children with Foetal Alcohol Spectrum Disorders (FASD) may make errors of cohesive reference in their use of nominal phrases at a higher rate than their typically developing peers. For diagnostic purposes, it is unclear whether diagnostic utility is increased by separately measuring and classifying these errors according to the various discourse purposes served by nominal references (i.e. introduction, maintenance/re-introduction). It is also unclear whether the narrative behaviours of children with Foetal Alcohol Syndrome (FAS) differ discriminatively along this dimension from other children with FASD.

What this study adds

The results of this narrative investigation underscore the diagnostic value of considering nominal reference errors of children with FASD as a unified category of behaviour rather than considering errors of introduction separate from errors of maintenance/re-introduction. Results also demonstrate the potential for using higher rates of nominal reference errors as a method for distinguishing children with FAS from those with other FASD.

Introduction

For over 30 years ethyl alcohol has been recognized as a teratogen and is the single largest preventable cause of birth defects and mental retardation (Abel and Sokol 1987). The term 'Foetal Alcohol Spectrum Disorders' (FASD) is applied to the full range of disability profiles associated with prenatal ethyl alcohol exposure (Bertrand *et al.* 2005). Sampson *et al.* (1997) estimated that approximately 1% of children are born with an FASD, with Foetal Alcohol Syndrome (FAS) accounting for perhaps 10% of cases. FAS was the first of the FASD to be recognized (Jones and Smith 1973) and is diagnosed based on growth deficiency, central nervous system (CNS) impairment, and a unique cluster of facial anomalies in the context

of prenatal alcohol exposure (Astley 2004, Bertrand *et al.* 2005, Chudley *et al.* 2005).

Diagnostic challenge in FASD

Diagnoses within FASD present a daunting challenge as only a few of the myriad of clinical outcomes are specifically linked to prenatal alcohol exposure (for a discussion, see Astley 2004). All of the major diagnostic guidelines in current use define the subcategories within FASD (e.g. partial-FAS) based on measurements along the same three parameters which identify FAS: growth, facial morphology, and CNS impairment (Astley 2004, Bertrand *et al.* 2005, Chudley *et al.* 2005). Growth and facial morphology are relatively straightforward to measure and their relationship to prenatal alcohol exposure has been extensively studied. While the facial features associated with FAS are highly specific to prenatal alcohol exposure, neither they nor growth deficiency are particularly sensitive markers of that exposure and are not associated with all FASD. It is the brain, rather, that is the organ most sensitive to prenatal alcohol exposure and CNS impairments are the most important sequelae of exposure (Streissguth and O'Malley 2000, Guerri 2002, Cortese *et al.* 2006). These impairments manifest across a wide range of CNS structures and functions (Riley *et al.* 2004, Kodituwakku 2007) making evaluation of CNS status a task best met by an interdisciplinary team. As language is a core domain of CNS functioning, a language specialist is an integral part of this diagnostic team.

CNS damage is readily diagnosed in FASD when significant structural deficits (e.g. microcephaly) or neurological abnormalities (e.g. hydrocephaly) are present. When there is a lack of direct evidence of CNS damage from structural or neurological measures, significant CNS dysfunction — equivalent to two or more standard deviations (SDs) from the mean on valid standardized tests — is used as a sign of CNS damage in FASD diagnoses (Astley 2004, Chudley *et al.* 2005). This strict diagnostic criterion has been adopted to improve specificity for CNS impairment. A lower cut-off on norm-referenced standardized tests would improve sensitivity, but only at the cost of increasing the false-positive rate.

Empirically derived criterions that are 'highly predictive' of CNS impairment can also provide evidence of CNS damage for diagnostic purposes (Astley 2004: 38). Given the wide range of possible functional impairments in FASD, a diagnostic team will increase the chances of successfully identifying meaningful impairments by choosing empirically validated measures of CNS functioning that are sensitive to prenatal alcohol exposure while maintaining acceptable levels of specificity (i.e. are highly predictive). This means there is a need to develop reliable and valid measures of CNS functioning sensitive to the specific kinds of CNS damage caused by prenatal alcohol exposure. These measures will need to be developed across all major domains of functioning, including language functioning, and will be critically important if we are to improve the process of diagnosing FASD.

Language impairments and FASD

Over the last 30 years the cognitive and behavioural functioning of children with FASD has been studied extensively (for a review, see Riley and McGee 2005).

Although no specific profile of cognitive and behavioural functioning has emerged, language and communication impairments are frequently associated with FAS and are among the most commonly reported functional impairments for FASD (e.g. Streissguth *et al.* 1994, Church and Kaltenbach 1997, Sowell *et al.* 2001, Greenbaum *et al.* 2002, Burd *et al.* 2003, Kvigne *et al.* 2004, Cone-Wesson 2005, Kodituwakku 2007).

Although the communication impairments associated with FASD are likely to be the result of an interaction between environment and capacity, the CNS damage caused by prenatal exposure certainly plays an important causal role (Coggins *et al.* 2007). In a recent review, Kodituwakku (2007) notes that:

> there exist only limited data on the effects of prenatal alcohol exposure on specific linguistic processes … [but] studies that obtained positive results employed relatively complex tests of language.
>
> (p. 196)

For example, children with an FASD might have difficulty balancing linguistic and socio-cognitive task demands in conversations (Hamilton 1981) and in narratives (e.g. Coggins *et al.* 1998, 2003, 2007, Thorne *et al.* 2007), and may fail to provide listeners with adequate information. Indeed, caregivers report that children with an FASD often fail to accommodate the perspectives of others during communication and interaction (Timler *et al.* 2005). It should also be noted that studies of FASD with negative findings for language impairment typically involve younger children with lower levels of prenatal alcohol exposure (e.g. Greene *et al.* 1990). On the whole the current research on communication impairments in FASD suggests that later developing, integrative functions involving complex language may be the most vulnerable to the types of CNS damage associated with prenatal alcohol exposure.

If language specialists are to provide useful information in team assessments of individuals suspected of having an FASD, it is essential that appropriate diagnostic tools be developed for measuring the communication abilities of this complex clinical population. Given what is known currently, it is likely that valid measures of complex and integrative language functioning will have the most diagnostic utility for this population (Coggins *et al.* 2003, Timler *et al.* 2005, Thorne *et al.* 2007).

FASD and narrative analysis

Narrative analysis is an ecologically valid method for measuring complex and integrative expressive language abilities in school-aged children and has been shown to discriminate between those with and without language impairment (e.g. Liles 1993, De Villiers 2004). Language produced during narrative tasks is less contextually constrained, more complex, and more cognitively demanding when compared with response requirements for standardized tests, which typically require discrete responses at the sentence level or below (but cf. Gillam and Pearson 2004). Analysis of a complex language task such as narrative — particularly in terms of narrative coherence (i.e. informativeness) and cohesion (i.e. the network of obligatory semantic relationships crossing sentence boundaries that are necessary for interpretation) — seems ideally suited for use with children with prenatal alcohol exposure (for a discussion, see Coggins *et al.* 1998).

In a retrospective survey of clinical records, Coggins *et al.* (2007) found that 149 of 393 school-aged children with an FASD (i.e. 38%) performed $\geqslant$2 SDs below the

mean on standardized language measures, demonstrating a significant impairment according to FASD diagnostic guidelines (Astley 2004). In addition to receiving a standardized language measure, 313 of these children also produced an oral narrative, which was evaluated using a criterion-referenced narrative analysis. The data reveal that 201 of these oral narratives (i.e. 64%) lacked age-appropriate coherence and cohesion. For this clinical population, the criterion-referenced narrative measure was more sensitive to FASD than the various standardized language measures used and identified a language deficit in an additional 52 children. Because it is a criterion-referenced measure, however, the diagnostic utility of this clinical measure is reduced. A quantitative measure of narrative coherence and/or cohesion would allow for greater discrimination of performance, and could provide the basis for a more diagnostically sound metric of functioning.

In a study examining the diagnostic utility of a comprehensive system for quantitative analysis of coherence and cohesion in narratives, Thorne *et al.* (2007) found a narrow measure of cohesive reference in nominal phrases to be the most promising and accurate of 26 measures for identifying school-aged children diagnosed with an FASD. The substantial accuracy of classification supported by the measure was shown to provide diagnostic information above and beyond that available from a standardized language measure. The current study will describe the performance characteristics of a refined system for quantifying these types of errors.

Semantic Elaboration Coding System (SECS)

The Semantic Elaboration Coding System (SECS; Thorne 2004) was designed to quantify ten discourse-level behaviours in the narrative microstructure (cf. Justice *et al.* 2006). The SECS, built on Talmy's (2000) framework of cognitive linguistics, operationalizes measurement of two essential discourse parameters: (1) *elaboration* of concepts, an aspect of coherence; and (2) avoidance of *ambiguity*, an aspect of cohesion. Skilfully manipulating these parameters of discourse requires storytellers to consider the perspective of their listener, an obligation many children with FASD find challenging (Coggins *et al.* 2003).

Is ambiguity the key?

Nominal phrases serve to introduce, maintain, or reintroduce characters and objects into a narrative (for discussions of these discourse functions, see Halliday and Hasan 1976, and Wong and Johnston 2004). When a nominal phrase is ambiguous, failing to meet its discourse function, the cohesion of the narrative is reduced. In the preliminary feasibility study, ambiguous nominal references (ANR) emerged as the SECS's most diagnostically promising measure (Thorne *et al.* 2007). The study showed that when calculated as a function of total words (TW), the rate of ambiguous nominal references (ANR/TW) correctly classified the narratives of 26 of the 32 children in the study (four false-positive, two false-negative, 81% overall accuracy) according to whether or not the storyteller had a diagnosis of an FASD (n=16).

Post-hoc analysis of these results indicated that the majority of incidents of ANR (>92%) involved definite nominal reference forms. No distinction in the SECS allowed a contrast to be made between (1) definite forms used inappropriately to *introduce* concepts into a story versus (2) definite forms used unsuccessfully in an

attempt to make a *referential tie* to maintain/re-introduce a previously introduced concept. Given that referential ties to concepts should significantly outnumber introductions of concepts in most narratives (resulting in a larger number of opportunities for errors to occur) this distinction has the potential to be diagnostically meaningful. To test this potential a new narrative analysis system has been developed that tallies these two aspects of discourse cohesion separately.

Tallying nominal reference errors in narrative

The analysis system described in this paper, Tallying Reference Errors in Narrative (TREIN; Thorne 2006), focuses exclusively on nominal reference strategies within a narrative and classifies strategies as appropriate or inappropriate based on expectations derived from the narrative context. Because the TREIN implements a more rigorous and comprehensive set of operational rules for defining nominal reference errors (NRE) than the SECS, it should make similar, but not identical, judgements regarding strategies for maintaining cohesive nominal reference in a narrative.

In the TREIN, errors of introduction (IE) result from improper marking of the obligatory known/new distinction found in nominal phrases in English. At a discourse level, characters, places, or objects can be introduced into the narrative as 'known' only if they are part of a common ground of shared knowledge. By convention, this common ground includes concepts with situational salience, functional dependency to already introduced concepts, or status as a unique/ubiquitous entity (cf. Heusinger 2003). Known concepts that are part of the common ground can be introduced into the narrative using a definite form (e.g. *the window*, *the boy's puppy*, *the sun*, *the air*). All other concepts need to be introduced into the narrative using an indefinite form (e.g. *a boy*, *some baby frogs*, *a jar*, *trees*). Introduction with a particular form establishes a referential precedent within the narrative (Barr and Keysar 2002). This precedent can then be used (in definite form) later in the narrative to establish cohesive referential ties to the shared concept as the story progresses (Brennan and Clark 1996).

In the TREIN, a referential tie error (TE) is identified when the storyteller breaks a referential precedent, fails to put a precedent in definite form, or fails to elaborate a referential form sufficiently for the listener to distinguish between several concepts already available in the common ground. Additionally, a TE is identified when a storyteller uses a precedent established for one concept to refer to a different concept (e.g. referring to the concept FROG using 'dog' — see the appendix for codes and examples).

Because a relatively limited range of discourse behaviours are tracked by the system, it is easy to learn and use. To get the total NRE for a narrative, the number of IE and TE are simply summed. This total NRE can then be used to calculate a rate of nominal reference errors (rNRE). For an individual trained on the system, obtaining the rNRE for a typical narrative transcript of the length examined here (308 word mean length) takes less than 15 minutes of coding and calculation.

Purpose

The current study examines the reliability and validity of the TREIN by quantifying its diagnostic performance within the same sample of narratives examined in

Thorne *et al.* (2007). All transcripts from that study were recoded using the TREIN, and diagnostic predictions based on the resulting rate of nominal reference errors (rNRE) were tested. The study provides answers to the following research questions:

- Can two independent coders substantially agree on coding decisions needed to complete a TREIN analysis?
- Does the rate of nominal reference errors (rNRE) from a TREIN analysis provide equivalent diagnostic classification to the rate of ambiguous nominal reference errors (ANR/TW) calculated with the SECS system of analysis?
- Do important diagnostic consequences result from different methods for calculating rNRE (i.e. which is the most promising)? This will be examined along two parameters.

    1 What diagnostic consequences result from tallying inappropriate strategies for introduction (IE) separately from inappropriate strategies for creating referential ties (TE)?
    2 What diagnostic consequences result from calculating the rate of each measure as a function of opportunities rather than as a function of story length (i.e. total words)?

Methods

Participants

This study retrospectively examined the same set of 32 transcripts of oral narratives utilized in Thorne *et al.* (2007). Storytellers ranged in age from 8;5 to 11;7 years (mean=9;11 years) and presented a range of socio-economic and ethnic profiles. Sixteen of the children (nine females, seven males) had previously been diagnosed with an FASD while the remaining 16 children (six females, ten males) were considered typically developing (TD). For a more complete description of participants, see Thorne *et al.* (2007).

Children with an FASD

The 16 participants in the FASD group had a diagnosis of either (1) full or partial-FAS (FAS=3, pFAS=2), or (2) a confirmed alcohol exposure accompanying static encephalopathy or neurobehavioral disorder. The children were diagnosed by an interdisciplinary team at the University of Washington Foetal Alcohol Syndrome Diagnostic and Prevention Network using The Four-Digit Diagnostic Code (Astley 2004). These 16 participants were selected from a larger FASD research database containing 42 children with FASD who had undergone a standard battery of cognitive and language testing as part of an intervention trial for children with FASD and behaviour/social problems (Carmichael-Olson and Astley 2005). Baseline testing provided the narratives and other measures used in selection for the current study.

Participants were selected for the current research based on their performance on the only standardized language measure available for all 42 children, the

Recreating Sentences subtest of the Test of Language Competence (RS-TLC; Wiig & Secord 1989). Of the initial 42 children, nine demonstrated average performance on the RS-TLC with standard scores within 1 SD of the mean (range=7–10), while seven had standard scores ⩾2 SDs below the mean (range=3–4). Those 26 children falling between 1 and 2 SDs below the mean were excluded from the study dichotomizing the sample into those with clearly impaired performance (i.e. with a 'significant' performance deficit; Astley 2004: 38) and those clearly without. The resulting 16 participants had scores ranging from 79 to 130 (mean=101) on the Matrices subtest from the Kaufman Brief Intelligence Test (Kaufman and Kaufman 1990), the available measure of non-verbal problem-solving ability.

Typically developing peers (TD)

Each participant with FASD was paired with a TD peer matched on chronological age (±12 months, mean difference=3.5 months). Thirteen TD age-matched peers also matched the gender of their FASD counterpart. The TD storytellers came from an existing database collected as part of a normative study of narrative production (Coggins 1995). No intelligence or standardized language measures were available for TD participants. However, a school psychologist familiar with the 16 children and with the profile of FASD screened school records for each child with respect to school performance, social ability, and general behaviour. Based on this review of available records, each was judged to be following a typical developmental course due to their unremarkable behaviour and adequate yet unexceptional school achievement. The TD participants did not undergo the same interdisciplinary assessment as the children with FASD.

Materials

Self-generated narratives were selected from two independent databases generated by two previous research studies (Coggins 1995, Carmichael-Olson and Astley 2005). All narratives were elicited using *Frog, Where Are You?* (Mayer 1969).

Procedures

Narrative collection

In both source studies participants were tested individually and received the same instructions. Each child was instructed to look through the wordless picture book to become familiar with the story line. When the child completed previewing the storybook, the examiner exhorted the participant to tell the best story possible while using the picture book as a visual prompt. In each case, examiners were seated across the room from the child with the storybook out of their line of sight.

Question 1: Intercoder agreement

All primary coding in this study was completed by the first author utilizing the TREIN protocol (Thorne 2006). To determine if an independent coder could substantially agree on the coding decisions involved in completing a TREIN analysis, a graduate student in speech and hearing sciences served as a secondary

coder. Coder competence was established when intercoder agreement between the primary and secondary coder reached a kappa (κ) of 0.7 or better for each nominal category code in the system on a set of five training narratives taken from the CHILDES databank (MacWhinney 2000). This training phase included approximately 10 hours of face-to-face training and 40 hours of coding practise. The secondary coder achieved criterion for coding accuracy after coding 18 narrative transcripts.

After the training phase was completed, the primary coder scored all 32 of the study narratives while the secondary coder independently scored 25% of the narratives ($n=8$) randomly selected using SPSS for Windows 9.0 (SPSS, Inc., Chicago, IL, USA). Transcripts included in this subsample included those from children identified as typically developing (four of 16 selected), children with a diagnosed FASD who performed within the average range on the RS-TLC (two of nine selected), and those that performed ⩾2 SDs below the mean on the RS-TLC (two of seven selected). To avoid coder bias the second author supervised the collection and transcription of all transcripts and assured that all were stripped of any and all identifying information regarding age, gender, and diagnostic status before coding. The second author maintained the link between the transcripts and participants and kept the primary author blind to these links throughout both the previous study and the initial phases of the current study (analysis conducted by the first author for Thorne *et al.* 2007, used only summary and corpus data without linking specific transcripts to participant identifiers).

Question 2: Comparison of the rNRE with ANR/TW

To establish the relationship between the rate of nominal reference errors (rNRE) provided by the TREIN analysis and the rate of Ambiguous Nominal Reference (ANR/TW) from a SECS analysis several comparisons were made. First, the concordance correlation coefficient (ρc) between the two measures was calculated using MedCalc (Schoonjans 2006). Based on a scatter plot correlating two continuous measures, ρc evaluates the degree to which the relationship between observations is explained by a 45° line through the origin (i.e. a line with slope 1.0 through the origin). The measure combines precision ρ and accuracy *Cb*:

$$\rho c = \rho Cb,$$

where ρ is the Pearson correlation coefficient quantifying deviation from the best-fit line, a measure of precision; and *Cb* is a bias correction factor quantifying how far this best-fit line deviates from the 45° line through the origin, and is a measure of accuracy.

Next, to assess differences between identification of nominal reference errors by the two systems for individual words in the narratives, kappa (κ) statistics between equivalent codes in the two systems were calculated for all 32 transcripts. Agreement was obtained when both systems placed a word in an equivalent category bin. Three equivalent-code category bins were defined as follows:

- ANR=NRE.
- NREF: successful nominal ties plus successful introductions in the SECS (general nominals + specific nominals + nominal reference)=successful

nominal ties plus successful introductions in the TREIN (nominal ties + indefinite introductions + definite introductions + possessive introductions).

- OTHER: all other coding categories (including null).

Finally, to establish how well rNRE identified the diagnostic status of participants, two tests were made. The first test used rNRE to predict which narratives were produced by children with a previously diagnosed FASD (FASD versus TD). This analysis parallels that used for ANR/TW in Thorne *et al.* (2007). The second test used rNRE to predict which narratives were produced by children previously diagnosed with either FAS or pFAS (for diagnostic category definitions, see Astley 2004).

For each prediction, sensitivity (true positive rate) and specificity (true negative rate) were calculated and plotted against each other for obtained values of rNRE creating empirical receiver-operating characteristic (ROC) curves. The area under the ROC curve (AUC), a widely accepted effect-size metric for accuracy, was then used to quantify the diagnostic utility of the measure for each prediction. This analysis was repeated using ANR/TW for each prediction, and pairwise statistical comparisons were made between the AUC for each measure (Hanley and McNeil 1983). All ROC and AUC analyses were conducted using MedCalc.

Question 3: Diagnostic consequences of using various measures

A TREIN analysis provides for the separate tallying of nominal reference tie errors (TE) and nominal introduction errors (IE) that are summed to get the total nominal reference errors (NRE). NRE is divided by the total words (TW) in the narrative when calculating rNRE. If the independent components of NRE (TE and IE) are not strongly correlated with each other and with diagnostic status, their combination may reduce diagnostic accuracy when compared with one or the other operating independently. In addition, given that the number of characters, places, and objects that are likely to be introduced into the story is relatively less sensitive to story length than the number or references to those concepts, it is also possible that calculating rate based on total words may differentially impact the diagnostic utility of IE and TE. To assess these possibilities, the ROC curve analysis described above was conducted again for each measure calculated in three ways.

Because it also codes appropriate introductions and appropriate referential ties (both nominal and pronominal), the TREIN analysis provides the data necessary to calculate rates based on opportunity for each discourse function in the narrative as is a common practice in discourse analysis (e.g. Wong and Johnston 2004). This practice is logically sound, but in a diagnostic context, only makes sense if it improves accuracy of classification. Thorne *et al.* (2007, see footnote) did not see a diagnostic advantage for calculating ANR/TW as a function of opportunities rather than as a function of TW, but given the disparate rates at which opportunities occur across the different reference functions of introducing and making referential ties, the diagnostic impact of using rates based on opportunity rather than story length needs to be directly studied for IE and TE. In a TREIN analysis, opportunities for referential ties (Opt) can be calculated based on the sum of codes for nominal reference tie + pronominal reference tie + nominal tie error + pronominal tie error, while opportunities for nominal introductions (Opi) are based on the sum of codes

for nominal introduction error + indefinite introduction + definite introduction + possessive introduction + pronominal introduction.

To compare the diagnostic consequences of each of these various methods for calculating the rate of IE and TE, a ROC curve analysis was carried out for the comparisons described above with several different metrics of each discourse function. In addition to the rNRE, the following measures were available from the TREIN analysis:

- Total number of TE.
- Rate of TE calculated as a function of total words (rTE).
- Rate of TE calculated as a function of opportunities (TE/Opt).
- Total number of IE
- Rate of IE calculated as a function of total words (rIE).
- Rate of IE calculated as a function of opportunities (IE/Opi).

The rate of NRE calculated as a function of opportunity (Op) was also calculated to allow for a complete set of comparisons. An opportunity to use a nominal reference was considered to obtain for all occasions in a narrative where the child used a noun or anaphoric pronoun. Op, then, consisted of the sum of Opi + Opt. As in Thorne *et al.* (2007), a measure was considered to have diagnostic potential only if the lower bound of its confidence interval for AUC was 0.7 or greater.

Results

Question 1: intercoder agreement

Omnibus agreement on the 2090 coding decisions made between coders in the reliability sample reached a kappa (κ) of 0.90 with a 95% confidence interval (CI) between 0.87 and 0.93. When calculated for the six nominal categories in the TREIN, agreement was excellent with a κ=0.96 (95% CI=0.95–0.98). When codes for each rater were independently combined to indicate referential opportunities, interrater agreement was substantial for introduction opportunities (Opi: κ=0.96, 95% CI=0.94–0.98), referential tie opportunities (Opt: κ=0.97, 95% CI=0.96–0.99), and overall referential opportunities (Op: κ=0.99, 95% CI=0.98–1.00). Agreement between coders was also high when IE and TE were combined to create the summary code NRE reaching κ of 0.81 (95% CI=0.72–0.91). This is an improvement over the κ of 0.76 reported for ANR in Thorne *et al.* (2007).

Question 2: Comparison of the rNRE with ANR/TW

The rNRE compares favourably with the ANR/TW in all domains examined. For the 11,058 word-by-word categorizations made across the 32 narratives, the two measures achieve a substantially high agreement for categorization of nominals into equivalent bins reaching a kappa (κ) of 0.88 (95% CI=0.86–0.89). As shown in figure 1, the two measures have a reasonable concordance correlation coefficient (ρc=0.86) indicating reasonable precision (Pearson ρ=0.89) and accuracy (Cb=0.97). As can be seen in figure 1, the rNRE is greater than the ANR/TW (i.e. falls above the 45° reference line) for 17 narratives, lower for three, and equivalent for the remaining 12. Importantly, of the 17 where rNRE>ANR/TW,

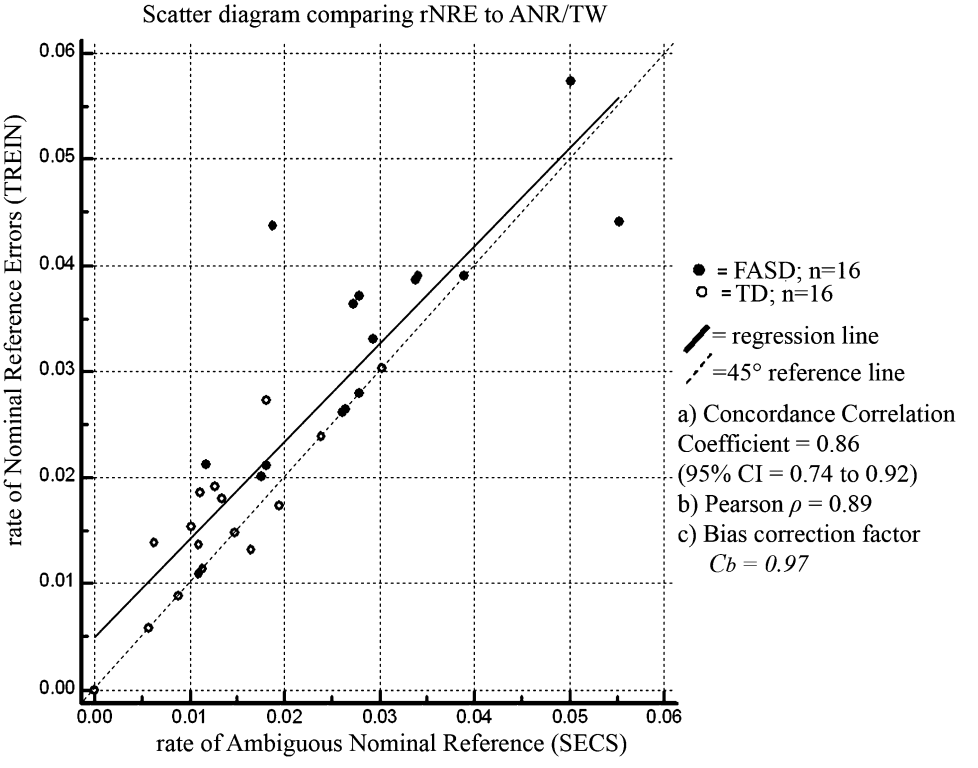

Figure 1. Scatter diagram with regression line comparing rNRE (TREIN) with ANR/TW (SECS) using ANR/TW as the reference standard with (a) concordance correlation coefficient, (b) precision ρ, and (c) accuracy C_b indicated.

ten are from children with a diagnosed FASD while two of the three with rNRE<ANR/TW are from the TD group. Taken together these trends are in the direction that should favour the diagnostic utility of rNRE over ANR/TW.

The trend towards greater diagnostic accuracy for the rNRE can be seen in the ROC curve comparison presented in figure 2 and table 1, which compare the two measures on their ability to classify correctly children into FASD versus TD groups. As seen in table 1, the AUC for rNRE is 'highly accurate' (Swets and Pickett 1982) with a point value of 0.90 and a 95% CI from 0.73 to 0.97 bridging the moderate to highly accurate ranges. This point value is 0.04 higher than the AUC point value for ANR/TW of 0.86, which has a 95% CI from 0.70 to 0.96.[1]

As is apparent from their overlapping confidence intervals, there is no statistically significant difference between the AUCs for the two measures (p=0.51). The ROC curve for rNRE is presented with the ROC curve for ANR/TW in figure 2 and shows graphically the near equivalence of the two measures' accuracy across values, with rNRE only slightly outperforming ANR/TW by more closely approaching the perfect-test point at the top left of the ROC plot (for guidelines on visual interpretation of ROC curves, see Kraemer 1992). Figures 3 and 4 illustrate the differential classification distributions of the two measures for this prediction at

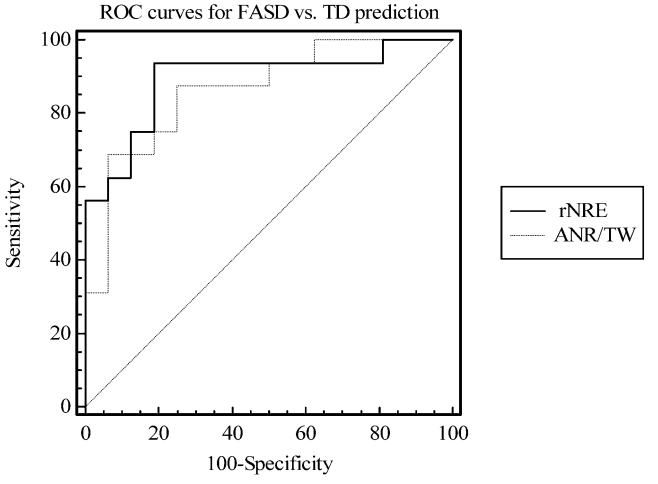

Figure 2. A 45° line indicates a random test reference line. The point on the curve closest to the top-left corner represents the best cut-point (maximizing overall accuracy). rNRE at the best cut-point outperforms ANR/TW despite the similarity of the ROC curves.

Table 1. Comparison of AUC for rNRE and ANR/TW for FASD versus TD classification

Measure used to predict FASD versus TD	AUC	SE	95% CI	Difference from rNRE	Significance level (p)	Percentage accuracy at the best cut-point (false-positive, false-negative)
rNRE	0.90	0.059	0.73–0.97	n.a.	n.a.	88% (+3, −1)
ANR/TW	0.86	0.067	0.70–0.96	−0.031	0.51	81% (+4, −2)

rNRE, rate of nominal reference errors; ANR/TW, ambiguous nominal reference rate.
The best cut-point is defined as the best overall accuracy (figures 3 and 4). If two points shared overall accuracy, that with the least false-negatives was chosen. In parentheses: +n, false-positives; −n, false-negative. Positive group=FASD, n=16; negative group=TD, n=16.

their best cut-point (rNRE at 2% and ANR/TW at 1.7%). For this group of children, there is a clinically important difference between the performance of rNRE and ANR/TW at their best cut-points, with rNRE reducing the number of false-positives from four to three, while reducing the number of false-negatives from two to one. This is an increase in overall accuracy from 81% for ANR/TW to 88% for rNRE.

As can be seen in table 2 and figure 5, when the two measures are used to predict which narratives were from children with a pre-existing diagnosis of full or partial FAS, both rNRE and ANR/TW are extremely accurate. For this classification, the two measures are, again, statistically equivalent (p=0.64), with rNRE making a single false-positive classification (97% overall accuracy) at its best cut-point of 3.8% (figure 6) and achieving an AUC of 0.98 (95% CI=0.85–0.99), while ANR/TW of 3.2% (figure 7) achieves perfect classification accuracy with AUC of 1.0 (95% CI=0.89–1.0). Given that any diagnosis of a disorder on the FASD continuum would be made only after considering information regarding all primary factors

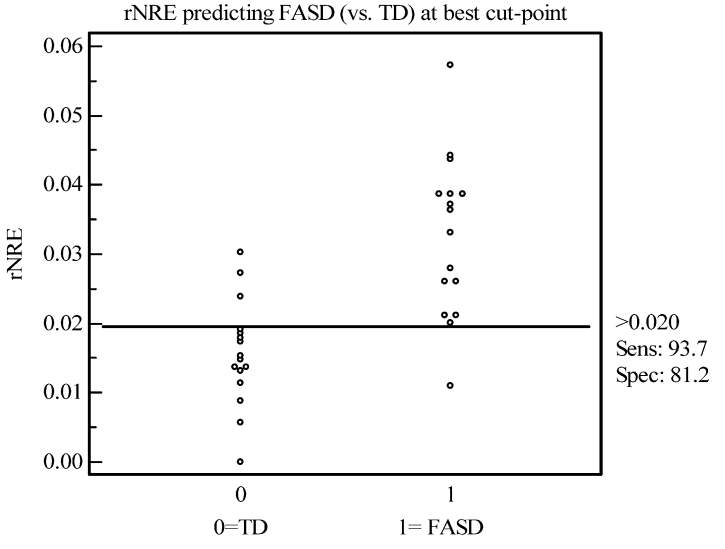

Figure 3. One=previously diagnosed with an FASD, while zero=children from the TD group. A child below the cut-off line in the column labelled '1' is a false-negative classification; those above the cut-off line in the column labelled 'zero' are false-positive classifications. The best cut-point is rNRE value >2%.

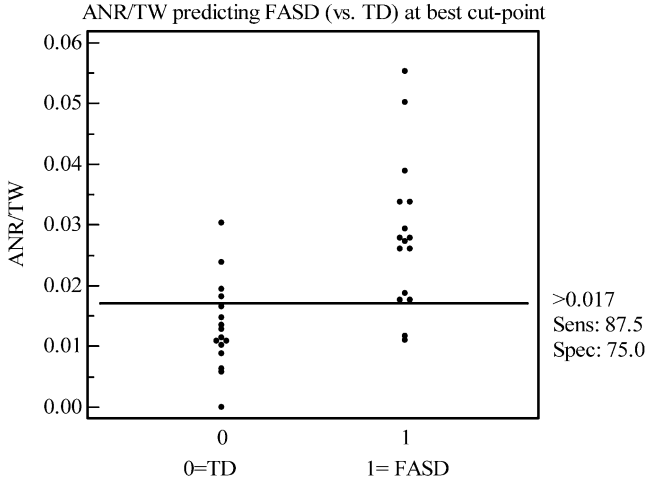

Figure 4. One=previously diagnosed with an FASD, while zero=children from the TD group. Children below the cut-off line in the column labelled '1' are false-negative classifications; those above the cut-off line in the column labelled 'zero' are false-positive classifications. The best cut-point is ANR/TW value >1.7%.

(growth, face, CNS, and prenatal alcohol exposure); a false-positive result is not as problematic as a false-negative result. For example, the single child misclassified here based on rNRE was diagnosed with static encephalopathy and high prenatal alcohol exposure, but failed to reach criteria for growth deficiency or facial morphology needed for a FAS/pFAS diagnosis.

Table 2. AUC comparison for rNRE and ANR/TW making an FAS/pFAS prediction

Measure used to predict FAS/pFAS	AUC	SE	95% CI	Difference from rNRE	Significance level (*p*)	Percentage accuracy at the best cut-point (false-positive, false-negative)
rNRE	0.98	0.047	0.85–0.99	n.a.	n.a.	97% (+1, −0)
ANR/TW	1.00	0.000	0.89–1.00	+0.02	0.64	100% (+0, −0)

rNRE, rate of nominal reference errors; ANR/TW, ambiguous nominal reference rate.
The best cut-point is defined as the best overall accuracy (figures 6 and 7). If two points shared an overall accuracy, that with the least false-negatives was chosen. In parentheses: +*n*, false-positives, −*n*, false-negative. Positive group=diagnosis of FAS or pFAS, *n*=5; negative group=TD and other FASD, *n*=27.

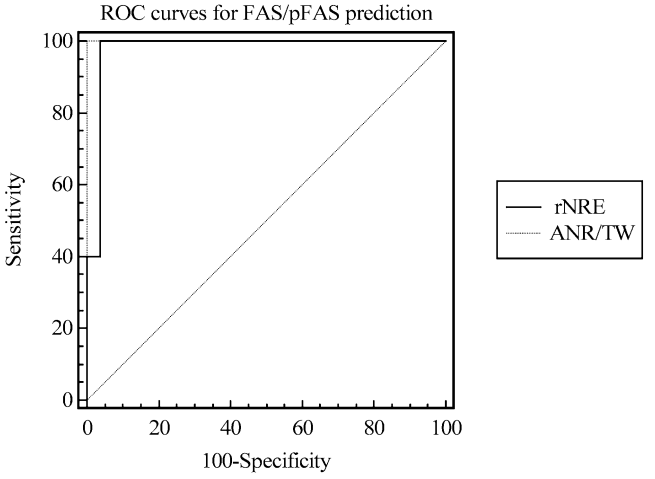

Figure 5. A 45° line indicates a random test reference line. the point on the curve closest to the top-left corner represents the best cut-point (maximizing overall accuracy). ANR/TW at the best cut-point outperforms rNRE despite the near equivalence of ROC curves.

Question 3: Diagnostic consequence of using various measures

Across the 32 narratives, the TREIN analysis identified 202 nominals as nominal introduction errors (IE) while only 16 nominal tie errors (TE) were identified. A total of 22 children (nine FASD and 13 TD) did not have any TE identified in their narratives. The diagnostic consequences of treating them separately are shown in tables 3 and 4. Table 3 compares the AUC for rNRE with all other available measures for making the FASD versus TD prediction and includes rates calculated both as a function of total words and opportunities. For this prediction task, the rNRE outperforms all other measures with none of the alternate measures based on IE or TE in isolation reaching criteria for a promising measure. The rate of introduction errors calculated as a function of total words (rIE) performed the best of the alternate measures with an AUC of 0.84 (95% CI=0.66–0.94). As shown by the pairwise comparison of rNRE to rIE, there is a statistically significant improvement in accuracy of this classification gained by combining the tally of inappropriate reference ties with the tally of inappropriate nominal introductions.

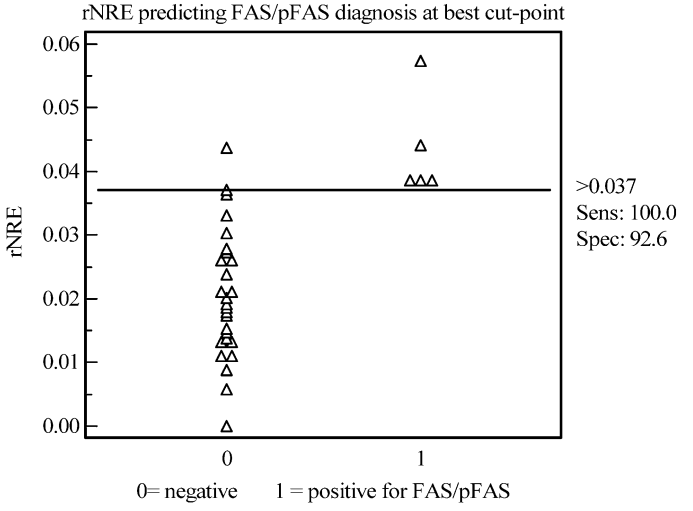

Figure 6. One=previously diagnosed with either FAS or pFAS; while zero=no diagnosis of FAS or pFAS and includes children from both FASD and TD groups. A child above the cut-off line in the column labelled 'zero' is a false-positive classification (diagnosed with static encephalopathy and confirmed alcohol exposure). The best cut-point is rNRE value>3.7%.

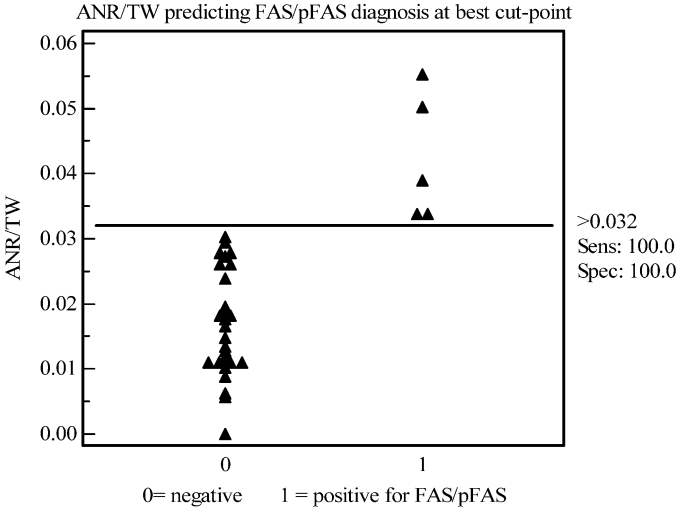

Figure 7. One=previously diagnosed with either FAS or pFAS; while zero=no diagnosis of FAS or pFAS and includes children from both FASD and TD groups. The best cut-point is ANR/TW value>3.2%.

As can be seen in table 4, the results for the FAS/pFAS prediction are similar. For this prediction, rates based on IE (rIE, IE/Opi), however, reached criteria for reasonable measures with AUC values that were not statistically different than rNRE (p=0.64 and 0.63, respectively) despite additional false-positive classifications. Overall accuracy for IE/Opi reached 84%. Overall accuracy for rIE reached 91%.

Table 3. Pairwise comparison with AUC of rNRE for predicting FASD versus TD

Measure used to predict FASD versus TD	AUC	SE	95% CI	Difference from rNRE	Significance level (*p*)	Percentage accuracy at the best cut-point (false-positive, false-negative)
rNRE	0.90	0.059	0.73–0.97	n.a.	n.a.	88% (+3, −1)
rIE*	0.84	0.073	0.66–0.94	−0.06	0.04**	78% (+3, −4)
IE*	0.73	0.091	0.54–0.87	−0.17	0.03**	78% (+5, −2)
TE*	0.63	0.099	0.44–0.79	−0.27	0.02**	63% (+3, −9)
rTE*	0.62	0.100	0.44–0.79	−0.28	0.02**	63% (+3, −9)
Rates calculated as a function of opportunity						
NRE/Op	0.89	0.061	0.73–0.97	−0.006	0.84	84% (+4, −1)
IE/Opi*	0.81	0.077	0.64–0.93	−0.08	0.06	78% (+4, −3)
TE/Opt*	0.62	0.100	0.43–0.78	−0.28	0.02**	63% (+3, −9)

rNRE, rate of nominal reference errors; rIE, rate of introduction errors calculated as a function of total words; I.E., total introduction errors; TE, total referential tie errors; Op, total reference opportunities; Opi, introduction opportunities; Opt, referential tie opportunities.
The best cut-point is defined as the best overall accuracy. If two points shared overall accuracy, that with the least false-negatives was chosen. In parentheses, +*n*, false-positives, −*n*, false-negative. Positive group=FASD, *n*=16; negative group=TD, *n*=16.
*Measures without diagnostic utility (lower bound of AUC CI<0.70).
**Measures with AUC significantly different than rNRE (α=0.05).

Table 4. Pairwise comparison with AUC of rNRE for predicting FAS or pFAS

Measure used to predict FAS/pFAS	AUC	SE	95% CI	Difference from rNRE	Significance level (*p*)	Percentage accuracy at the best cut-point (false-positive, false-negative)
rNRE	0.98	0.047	0.85–0.99	n.a.	n.a.	97% (+1, −0)
rIE	0.96	0.066	0.82–0.99	−0.02	0.64	91% (+3, −0)
IE*	0.70	0.140	0.52–0.85	−0.28	0.02**	88% (+0, −4)
TE*	0.66	0.143	0.47–0.82	−0.32	0.02**	84% (+0, −5)
rTE*	0.70	0.140	0.52–0.85	−0.28	0.02**	84% (+3, −2)
Rates calculated as a function of opportunity						
NRE/Op	0.95	0.071	0.81–0.99	−0.03	0.54	88% (+4, −0)
IE/Opi	0.96	0.063	0.82–1.00	−0.02	0.63	84% (+5, −0)
TE/Opt*	0.70	0.141	0.51–0.85	−0.28	0.05**	84% (+3, −2)

rNRE, rate of nominal reference errors; rIE, rate of introduction errors calculated as a function of total words; IE, total introduction errors; TE, total referential tie errors; Op, total reference opportunities; Opi, introduction opportunities; Opt, referential tie opportunities. The best cut-point is defined as the best overall accuracy. If two points shared overall accuracy, that with the least false-negatives was chosen. In parentheses: +*n*, false-positives; −*n*, false-negative. Positive group=diagnosis of FAS or pFAS, *n*=5; negative group=TD and other FASD, *n*=27.
*Measures without diagnostic utility (lower bound of AUC CI<0.70).
**Measures with AUC significantly different than rNRE (α=0.05).

Diagnostic consequences of calculation based on opportunity

As can be seen in tables 3 and 4, when the rates for NRE, IE or TE are calculated as a function of opportunities rather than total words in the story for either prediction task, they result in more incorrect predictions. For the FASD versus TD prediction, the overall accuracy of rNRE is 88% while its counterpart based on opportunity (NRE/Op) only achieves an overall accuracy of 84%. For the FAS/pFAS prediction, rIE outperforms its counterpart based on opportunity (IE/Opi) 91% to 84% with two fewer false-positive predictions. For the FAS/pFAS prediction, rNRE makes a single false-positive prediction giving it an overall accuracy of 97%, which is 9 percentage points above the accuracy of NRE/Op which makes four false-positive predictions.

As shown in table 3 and figure 8, when the performance of rNRE is compared with that of NRE/Op for the FASD versus TD prediction, there is virtually no difference in the ROC curve or the AUC of the two measures ($p=0.84$). Similarly, as seen in table 4 and figure 9, when the ability of rNRE to predict which narratives were produced by children with a diagnosis of FAS/pFAS is compared with that of NRE/Op, their ROC and AUC are very similar ($p=0.54$). The ROC curves of rNRE for both predictions, however, approach more closely the perfect-test point on the top left of the ROC plot (representing an AUC of 1.0). This reflects the fact that the measure rNRE obtains a higher point-value for AUC while producing fewer false-positive classifications than NRE/Op for both classifications.

Non-verbal ability, standardized language scores and rNRE

Non-verbal ability is available for the FASD group in the form of *Matrices* subtest of the KBIT (M-KBIT). These standardized test scores allow for descriptive examination of how this aspect of non-verbal ability is related to rNRE in the FASD group (the sample size lacks power for statistical analysis). Figure 3 shows that only one child from the FASD group falls below the rNRE cut-off score of 2%, while three children from the TD group fall above that cut-off. For the seven children with FASD who fall within the range of rNRE seen in the TD group (0–3%), their M-KBIT scores range from 92 to 130, with only one falling below the rNRE cut-off score of 2%. This child has an M-KBIT score of 122 — the third highest in the FASD group, 1.5 SDs above the mean. The two children with higher M-KBIT scores (128 and 130) have rNRE above the cut-off score (2.6 and 2.8%, respectively). For those FASD children with rNRE falling above the cut-off of 2%, M-KBIT scores fall in a range from 79 to 130, with only three falling more than –1.0 SDs from the mean. The child with the lowest M-KBIT score has the fifth highest rNRE (3.9%).

The FASD group is dichotomized based on their performance on the available standardized language measure, the RS-TLC. The majority of the children in the FASD group (nine of 16) had RS-TLC scores in the average range. Only one child out of nine who fell in the average range on the RS-TLC fell below the rNRE cut-off of 2% with the group ranging from a low rNRE of 1.1% to a high of 3.9%. The two highest rNRE (3.9 and 3.7%) from this group had RS-TLC scores of seven and nine, respectively. The child with the highest RS-TLC score (of ten) fell just above the 2% rNRE cut-off. Six of the eight highest rNRE (range=3.6–5.7%) come from children with RS-TLC scores of three.

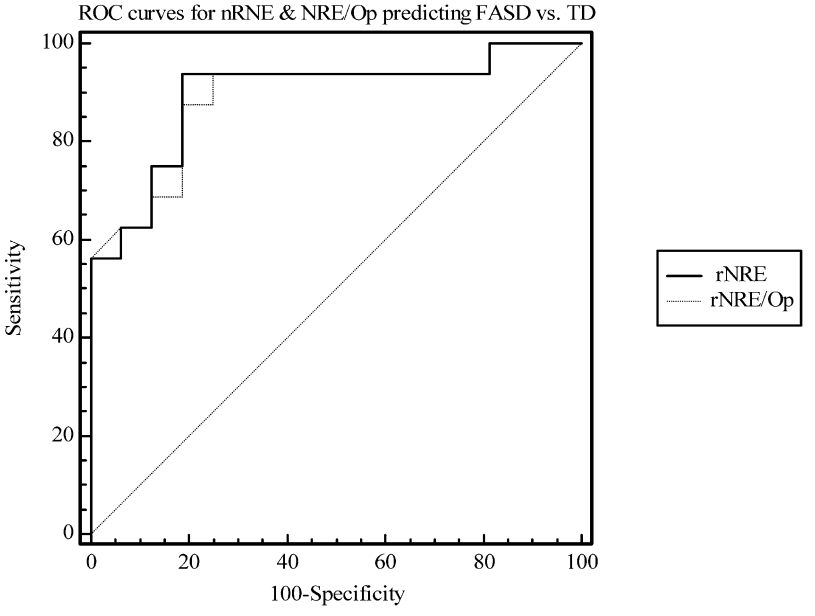

Figure 8. A diagonal line on the lower left of the NRE/Op ROC curve indicates tied values. A 45° line indicates a random test reference line. The point on the curve closest to the top-left corner represents the best cut-point (maximizing overall accuracy). rNRE at the best cut-point outperforms NRE/Op despite the near equivalence of ROC curves.

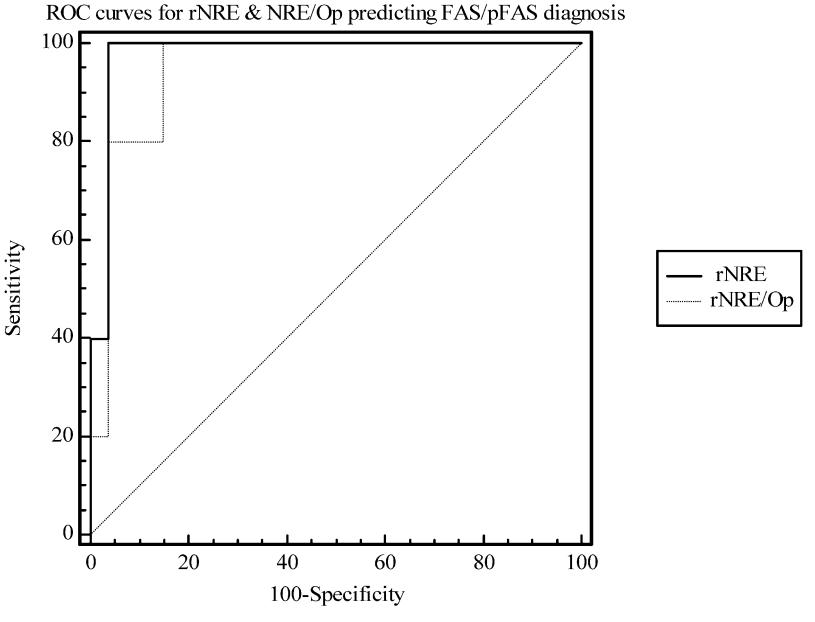

Figure 9. A 45° line indicates a random test reference line. The point on the curve closest to the top-left corner represents the best cut-point (maximizing overall accuracy). rNRE at the best cut-point outperforms NRE/Op despite the similarity of ROC curves.

Discussion

The evidence presented in this study reveals that the proposed method for calculating nominal reference errors is reliable and has sufficient diagnostic utility to merit further study with children suspected of prenatal alcohol exposure. The rNRE from the TREIN demonstrated an improvement in interrater agreement over ANR/TW from κ=0.76 to 0.81 and correctly classified FASD versus TD for 88% of the sample compared with 81% for ANR/TW. When used to predict which narratives were from children with previous diagnoses of FAS or partial-FAS, the rNRE performed with equivalent sensitivity compared with ANR/TW; despite making a single false-positive classification for a child confirmed to have high levels of prenatal alcohol exposure and a diagnosis of static encephalopathy.

No advantage was seen in treating errors of introduction separately from unsuccessful attempts to use nominals to make referential ties. While the metric rNRE is clearly dominated by errors of introduction (which occurred at over 12 times the rate in these narratives), the additional accuracy provided by including unsuccessful referential ties gives rNRE a greater potential for development as a diagnostic tool for use with children suspected of having a disorder on the foetal alcohol spectrum. Also, no improvements to the measure are gained by calculating the rate of nominal reference errors as a function of total opportunities rather than as a function of total words in a narrative. Given the ease with which total words can be calculated compared with a tally of reference opportunities, it is clearly the preferred metric.

Validity of the rNRE in clinical application

To illustrate the potential clinical value of rNRE, let us consider the impact it might have on the diagnosis given to a child from our clinical database, a 9-year-old boy diagnosed with partial-FAS. Information available in our database for this child allow us to confirm high-levels of prenatal alcohol exposure, and that he has growth and facial features consistent with a diagnosis of pFAS. With evidence of CNS impairment, this child would meet all the criteria needed for a diagnosis of pFAS (Astley 2004).

The child's non-verbal IQ, as represented by his score of 101 on the Matrices subtest of The Kaufman Brief Intelligence Test (M-KBIT), places him near the mean for his age, and does not provide any evidence of functional CNS impairment. Records also included three standardized measures related to language. On the Vocabulary subtest of the KBIT this child scored 106, again near the mean and not indicative of CNS impairment. On two subtests of the Test of Language Competence (TLC), Recreating Sentences — scaled score of seven, and Figurative Language — scaled score of nine; the child's scores are within 1 SD of the mean and, likewise, do not provide evidence of CNS impairment. A language specialist that relied on this child's performance on these standardized measure would not be able to provide any evidence to the interdisciplinary team that this child's language performance was indicative of the CNS impairment associated with pFAS.

It is possible that another standardized language measure would give different results, but available research suggests that children with FAS or pFAS often perform within an acceptable range on standardized tests despite parental reports of poor communication skills (Timler *et al.* 2005, Kodituwakku 2007). The most likely

reason for this apparent discrepancy is that these children are exhibiting impairment of higher-level integrative language functions not tapped by the discrete responses typical of standardized language tests (cf. Kalberg and Buckley 2007). The production of a narrative requires these higher-level integrative language functions and, if the current results can be validated, may provide a more sensitive measure of CNS impairment in this complex clinical population.

If the sensitivity and overall accuracy of the rNRE found in the current study can be validated for the clinical FASD population in general, language specialists would have a sensitive and efficient tool for identifying the impact of CNS damage on communicative discourse. For the case presented above, then, if the rNRE were added to the interdisciplinary discussion, this child's rNRE of 3.9% would place him in the performance range of the children with FAS/pFAS and could provide the team with evidence to support a diagnosis of pFAS. The language specialist on the interdisciplinary team would have contributed a valuable piece of information to the diagnostic decision making process using a measure which is easy to administer and score, and appears to be highly sensitive to the type of CNS impairment associated with FASD.

Limitations of the current study

Because this feasibility study utilized a retrospective design, we were limited to available data to answer our research questions. TD participants were identified as typically developing based on a review of school records indicating unremarkable behaviour and adequate yet unexceptional school achievement. They did not undergo the same clinical assessment as the children diagnosed with FASD. The lack of objective language and cognitive measures on these children increases the chances of type I errors in the unlikely event that as a group their ability was significantly above average along the parameters of interest. In clinical practice, however, an academic and behavioural profile like that presented by the TD group would not typically trigger a diagnostic assessment and, therefore, provides a reasonable basis for contrasting the two groups in the context of this feasibility study.

The lack of important objective data for the TD group has the potential to *enhance* the appearance of diagnostic accuracy for rNRE if, for instance, a relationship between rNRE and some other cognitive ability were important and the TD group were skewed towards the upper end of the scale on that ability. For instance, the 13 children from the TD group that produced rNRE below the 2% cut-off might all have performed in the above average range on the RS-TLC if given the chance. If this were the case, our rNRE cut-off score would be identifying a difference not between FASD and typical development, but between FASD and above average language ability. The range and distribution of M-KBIT scores seen in the FASD group make variation in the abilities that underpin performance on the M-KBIT a less than satisfying explanation for variation seen for rNRE in our narrative task — even if the TD group were assumed to have abilities that would result in high M-KBIT scores. There may be, of course, other measures (e.g. executive functioning or attention) that would also explain higher versus lower rNRE resulting in a type I error. Future research will be needed to examine these possibilities.

It must be kept in mind that the lack of objective measures confirming that these children were indeed 'typically developing,' and not subject to a prenatal alcohol exposure, also increases the risk of type II errors. This has the potential for *masking* the diagnostic accuracy of rNRE if, for instance, the three TD children with rNRE above the cut-off score of 2% (figure 3) experienced significant prenatal alcohol exposure and would, therefore, be more appropriately included in the FASD group.

The risk of both type I and type II errors resulting from our use of retrospective data enhance the need for the results of this initial feasibility research to be confirmed with validation research.

Conclusion

We conclude by emphasizing that this study documents a behavioural tendency in a small sample of children with previously evaluated CNS impairment (required for their FASD diagnoses). Future research will be required to determine to what degree this behaviour is a reliable marker of CNS impairments associated with prenatal alcohol exposure. The results presented here will need to be replicated with a similar group of children as a first step towards developing rNRE as a diagnostically useful tool for use with this population. Additional validation research with larger clinical populations will also be required. Given that it is unlikely that the behaviour being measured is one that is specific to children with a diagnosed FASD, despite its apparent sensitivity for this group, exploration of the rNRE in narratives by children from other clinical groups is also sensible. Of particular interest will be groups, such as those diagnosed with Attention Deficit Hyperactivity Disorder and those with Pragmatic Language Impairment that, even in the absence of a confirmed alcohol exposure, share many of the communication challenges seen in FASD.

As noted by Riley *et al.* (2004):

> if we could identify a profile of behavioral and brain markers indicating prenatal alcohol exposure, individuals without the facial characteristics of FAS could more easily obtain the educational and social support that is so necessary. (p. 40)

FASD, however, are more properly conceptualized as a collection of distinct disorder categories (Bertrand *et al.* 2005). This means it is unlikely that a single profile of behaviour/brain markers exists across the range of FASD. The range of profiles, however, is likely to be limited (Riley and McGee 2005, Suzuki 2007), and unevenly distributed across the population (assuming that certain patterns of prenatal alcohol exposure are more common than others).

The current study indicates that rNRE is a measure that has the potential to help define one highly prevalent profile in children with FASD. If it can be shown that the rNRE is robust at predicting FAS/pFAS, and, more importantly, that it can predict the CNS damage in a substantial subset of the clinical population of children suspected of having FASD, the rNRE could also become an valuable tool to help researchers understand the behavioural and neuro-cognitive consequences of prenatal alcohol exposure. Careful study of contrasting disability groups in the context of modern neuroimaging technologies also provides the opportunity to examine the possibility that children with similar impairments of narrative cohesion

may share similar underlying CNS damage — irrespective of the source of that damage. This would have important implications for diagnosis and intervention planning in treating FASD and other disorders associated with impairment of higher-level integrative language functioning.

Note

1. The use of Medcalc to calculate the AUC resulted in slight differences (0.04) in confidence interval values for ANR/TW from those published by Thorne *et al.* (2007) based on different methods for obtaining confidence intervals between MedCalc and SPSS (1998).

References

Abel, E. L. and Sokol, R. J., 1987, Incidence of fetal alcohol syndrome and economic impact of FAS-related anomalies. *Drug and Alcohol Dependence*, **19**, 51–70.

Astley, S. J., 2004, *Diagnostic Guide for Fetal Alcohol Spectrum Disorders: The Four-Digit Diagnostic Code* (Seattle, WA: FAS Diagnostic and Prevention Network, University of Washington) (available at: http://fasdpn.org).

Barr, D. J. and Keysar, B., 2002, Anchoring comprehension in linguistic precedents. *Journal of Memory and Language*, **46**, 391–418.

Bertrand, J., Floyd, L. L. and Weber, M. K., 2005, Guidelines for identifying and referring persons with fetal alcohol syndrome. *Morbidity and Mortality Weekly Report: Recommendations and Reports*, **54**, 1–14.

Brennan, S. E. and Clark, H. H., 1996, Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **22**, 1482–1493.

Burd, L., Cotsonas-Hassler, T. M., Martsolf, J. T. and Kerbeshian, J., 2003, Recognition and management of fetal alcohol syndrome. *Neurotoxicology and Teratology*, **25**, 681–688.

Carmichael-Olson, H. and Astley, S. J., 2005, Intervening with children adolescents with FAS/ARND. Unpublished raw data (Seattle, WA: University of Washington).

Chudley, A. E., Conry, J., Cook, J. L., Loock, C., Rosales, T. and Leblanc, N., 2005, Fetal alcohol spectrum disorder: Canadian guidelines for diagnosis. *Canadian Medical Association Journal*, **172**, S1–S21.

Church, M. W. and Kaltenbach, J. A., 1997, Hearing, speech, language, and vestibular disorders in the fetal alcohol syndrome: a literature review. *Alcoholism, Clinical and Experimental Research*, **21**, 495–512.

Coggins, T. E., 1995, Narratives produced by typically developing school-aged children. Unpublished raw data (Seattle, WA: University of Washington).

Coggins, T. E., Friet, T. and Morgan, T., 1998, Analysing narrative productions in older school-age children and adolescents with fetal alcohol syndrome: an experimental tool for clinical applications. *Clinical Linguistics and Phonetics*, **12**, 221–236.

Coggins, T. E., Olswang, L. B., Carmichael-Olsen, H. and Timler, G., 2003, On becoming socially competent communicators: the challenge for children with fetal alcohol exposure. *International Review of Research in Mental Retardation*, **27**, 121–150.

Coggins, T. E., Timler, G. R. and Olswang, L. B., 2007, A state of double jeopardy: Impact of prenatal alcohol exposure and adverse environments on the social communicative abilities of school-age children with Fetal Alcohol Spectrum Disorder. *Language, Speech, and Hearing Services in Schools*, **38**, 117–127.

Cone-Wesson, B., 2005, Prenatal alcohol and cocaine exposure: influences on cognition, speech, language, and hearing. *Journal of Communication Disorders*, **38**, 279–302.

Cortese, B. M., Moore, G. J., Bailey, B. A., Jacobson, S. W., Delaney-Black, V. and Hannigan, J. H., 2006, Magnetic resonance and spectroscopic imaging in prenatal alcohol-exposed children: preliminary findings in the caudate nucleus. *Neurotoxicology and Teratology*, **28**, 597–606.

De Villiers, P., 2004, Assessing pragmatic skills in elicited production. *Seminars in Speech and Language*, **25**, 57–71.

Gillam, R. B. and Pearson, N. A., 2004, *Test of Narrative Language* (Austin, TX: Pro-Ed).

Greenbaum, R., Nulman, I., Rovet, J. and Koren, G., 2002, The Toronto experience in diagnosing alcohol-related neurodevelopmental disorder: a unique profile of deficits and assets. *Canadian Journal of Clinical Pharmacology*, **9**, 215–225.

Greene, T., Emhart, C. B., Martier, S., Sokol, R. and Ager, J., 1990, Prenatal alcohol exposure and language development. *Alcoholism: Clinical and Experimental Research*, **14**, 937–945.

Guerri, C., 2002, Mechanisms involved in central nervous system dysfunctions induced by prenatal ethanol exposure. *Neurotoxicity Research*, **4**, 327–335.

Halliday, M. A. K. and Hasan, R., 1976, *Cohesion in English* (London: Longman).

Hamilton, M., 1981, Linguistic abilities of children with fetal alcohol syndrome. Unpublished doctoral dissertation, University of Washington, Seattle.

Hanley, J. A. and McNeil, B. J., 1983, A method of comparing the areas under receiver-operating characteristic curves derived from the same cases. *Radiology*, **148**, 839–843.

Heusinger, K. V., 2003, The double dynamics of definite descriptions. In J. Peregrin (ed.), *Meaning: The Dynamic Turn* (New York, NY: Elsevier), pp. 149–168.

Jones, K. and Smith, D., 1973, Recognition of the fetal alcohol syndrome in early infancy. *Lancet*, **ii**, 999–1001.

Justice, L. M., Bowles, R. P., Kaderavek, J. N., Ukrainetz, T. A., Eisenberg, S. L. and Gillam, R. B., 2006, The Index of Narrative Microstructure: a clinical tool for analyzing school-age children's narrative performances. *American Journal of Speech–Language Pathology*, **15**, 177–191.

Kalberg, W. O. and Buckley, D., 2007, FASD: what types of intervention and rehabilitation are useful? *Neuroscience and Biobehavioral Reviews*, **31**, 278–285.

Kaufman, A. S. and Kaufman, N. L., 1990, *The Kaufman Brief Intelligence Test* (Circle Pines, MN: American Guidance Service).

Kodituwakku, P. W., 2007, Defining the behavioral phenotype in children with fetal alcohol spectrum disorders: a review. *Neuroscience and Biobehavioral Reviews*, **31**, 192–201.

Kraemer, H. C., 1992, *Evaluating Medical Tests: Objective and Quantitative Guidelines* (London: Sage).

Kvigne, V. L., Leonardson, G. R., Neff-Smith, M., Brock, E., Borzelleca, J. and Welty, T. K., 2004, Characteristics of children who have full or incomplete fetal alcohol syndrome. *Journal of Pediatrics*, **145**, 635–640.

Liles, B. Z., 1993, Narrative discourse in children with language disorders and children with normal language: a critical review of the literature. *Journal of Speech and Hearing Research*, **36**, 868–882.

MacWhinney, B., 2000, *The CHILDES project: Tools for analyzing talk* (Mahwah, NJ: Lawrence Erlbaum Associates).

Mayer, M., 1969, *Frog, where are you?* (New York: Dial Press).

Riley, E. P. and McGee, C. L., 2005, Fetal alcohol spectrum disorders: an overview with emphasis on changes in brain and behavior. *Experimental Biology and Medicine*, **230**, 357–365.

Riley, E. P., McGee, C. L. and Sowell, E. R., 2004, Teratogenic effects of alcohol: a decade of brain imaging. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, **127**, 35–41.

Sampson, P. D., Streissguth, A. P., Bookstein, F. L., Little, R. E., Clarren, S. K., Dehaene, P., Hanson, J. W. and Graham, J. M., Jr., 1997, Incidence of fetal alcohol syndrome and prevalence of alcohol-related neurodevelopmental disorder. *Teratology*, **56**, 317–326.

Schoonjans, F., 2006, MedCalc, 8.2.0.1 edn (Mariakerke: MedCalc Software).

Sowell, E. R., Mattson, S. N., Thompson, P. M., Jernigan, T. L., Riley, E. P. and Toga, A. W., 2001, Mapping callosal morphology and cognitive correlates: effects of heavy prenatal alcohol exposure. *Neurology*, **57**, 235–244.

Streissguth, A. P., Barr, H. M., Olson, H. C., Sampson, P. D., Bookstein, F. L. and Burgess, D. M., 1994, Drinking during pregnancy decreases word attack and arithmetic scores on standardized tests: adolescent data from a population-based prospective study. *Alcoholism: Clinical and Experimental Research*, **18**, 248–254.

Streissguth, A. P. and O'Malley, K., 2000, Neuropsychiatric implications and long-term consequences of fetal alcohol spectrum disorders. *Seminars in Clinical Neuropsychiatry*, **5**, 177–190.

Suzuki, K., 2007, Neuropathology of developmental abnormalities. *Brain and Development*, **29**, 129–141.

Swets, J. A. and Pickett, R. M., 1982, *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory* (New York, NY: Academic Press).

Talmy, L., 2000, *Toward a Cognitive Semantics: Volume II- Typology and Process in Concept Structuring* (Cambridge, MA: MIT Press), pp. 417–482.

Thorne, J. C., 2004, *The Semantic Elaboration Coding System* (Seattle, WA: University of Washington) (available from the first author at: jct6@u.washington.edu).

Thorne, J. C., 2006, *Tallying Reference Errors in Narrative* (Seattle, WA: University of Washington) (available from the first author at: jct6@u.washington.edu).

Thorne, J. C., Coggins, T. E., Carmichael Olson, H. and Astley, S. J., 2007, Exploring the utility of narrative analysis in diagnostic decision making: picture-bound reference, elaboration, and Fetal Alcohol Spectrum Disorders. *Journal of Speech, Language, and Hearing Research*, **50**, 459–474.

Timler, G. R., Olswang, L. B. and Coggins, T. E., 2005, 'Do I know what I need to do?' A social communication intervention for children with complex clinical profiles. *Language, Speech, and Hearing Services in Schools*, **36**, 73–85.

Wiig, E. H. and Secord, W. A., 1989, *Test of language competence-expanded edition* (San Antonio, TX: Psychological Corporation).

Wong, A. M.-Y. and Johnston, J. R., 2004, The development of discourse referencing in Cantonese-speaking children. *Journal of Child Language*, **31**, 633–660.

Appendix

Table A.1. Tallying reference errors in narrative (TREIN): overview of codes

Introduction codes	
+ [indefintro]	Indefinite form used to introduce a concept [explicit addition of a concept to the common ground of the story (see example 1)]
+ [defintro]	Definite introduction of a concept already in the common ground. [endophoric, implicit concept made explicit: indicates situational salience, functional dependency to already introduced concept, unique or ubiquitous entity (cf. Heusinger 2003; see example 2)]
+ [possintro]	Possessive introduction of concept. [endophoric (see example 2)]
− [IntroError] IE*	Ambiguous introduction of concept using a definite form not supported by contextual factors. [exophoric (see example 3)] Also used for an inappropriate use of an indefinite form [aphoric, these are rare, (see example 4)]
− [pnintro]	Pronominal introduction of concept. [exophoric, (see example 5)]
Referential tie codes	
+ [ntie]	Clear referential tie to concept in story text utilizing nominal form. [endophoric, either anaphoric or cataphoric (see example 4)]
− [nTieError] TE*	Ambiguous referential tie using nominal form that breaks referential precedent. [bi/multiphoric (see example 5): also mislabelling available concepts, e.g. 'dog' for FROG]
+ [pntie]	Clear referential tie using pronominal form. [endophoric]
− [pntieError]	Ambiguous referential tie using pronominal form. [biphoric/multiphoric]

+, Appropriate strategy; −, inappropriate strategy in the TREIN narrative context.
*NRE is the combination of IE and TE.

Table A.2. Example sentences for nominal codes: [in brackets]

1. A little boy[indefintro] was looking at a frog[indefintro] that he kept in a jar[indefintro]. *(This is the first utterance in the story)*
2. A boy[indefintro] was sitting on the floor[defintro] looking at his new frog[possintro]. *(This is the first utterance in the story)*
3. The little boy[IntroError] was looking at the frog[IntroError] that he kept in the jar[IntroError]. *(This is the first utterance in the story)*
4. A boy[indefintro] was sitting on the floor[defintro] looking at his new frog[possintro]. A boy[IntroError] thought the frog[ntie] was happy.
5. The boy[ntie] saw two frogs[indefintro], and that frog[nTieError] was on a log[indefintro]. *(Which frog is 'that frog'?)*