Tallying Reference Errors in Narratives: Integrative Language Function, Impairment, and Fetal Alcohol Spectrum Disorders

John C. Thorne

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2010

Program Authorized to Offer Degree:

Department of Speech and Hearing Sciences

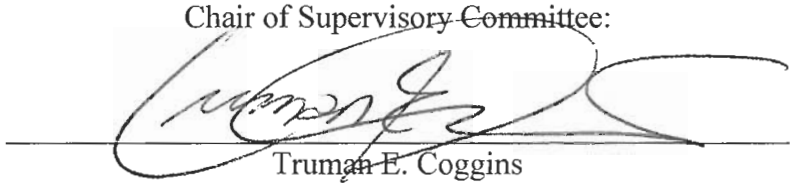
University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

John C. Thorne

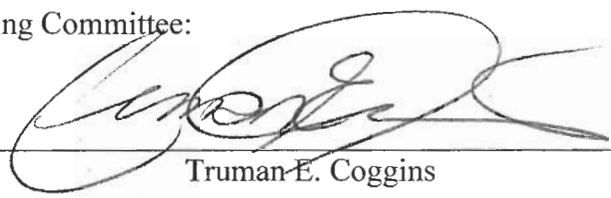
and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of Supervisory Committee:

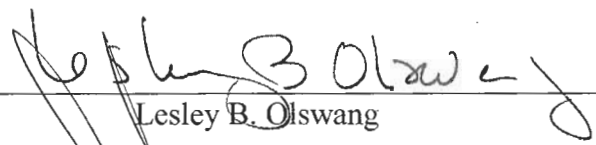


Truman E. Coggins

Reading Committee:



Truman E. Coggins



Lesley B. Olswang

Susan J. Astley

Date: 12/15/2010

University of Washington

Abstract

Tallying Reference Errors in Narratives:
Integrative Language Function,
Impairment, and Fetal Alcohol Spectrum Disorders

John C. Thorne

Chair of the Supervisory Committee:
Professor Truman E. Coggins
Department of Speech and Hearing Sciences

This study investigates the construct validity of a new measure of Integrative Language functioning, *Tallying Reference Errors In Narratives* (TREIN), by examining the association between previously existing CNS impairment and Expressive Language functioning and elevated outcomes on the TREIN measure "rate of Nominal Reference Errors" (rNRE). The rNRE is a measure of referential cohesion errors in noun phrases. Study participants included 155 elementary school aged children, 75 of whom had been identified with CNS impairments during a clinical assessment of suspected Fetal Alcohol Spectrum Disorders (FASD). Data came from existing clinical and research data including oral narratives elicited using a wordless picture book. Referential cohesion in the narratives was analyzed blind to participants' previous assessment results, diagnoses, age or gender. Statistical analysis of group level performance in terms of means, correlations, and performance distribution were conducted to reveal any existing relationships between narrative performance and other clinical measures of CNS impairment and language functioning. Results support the validity of the rNRE as a measure of Integrative Language functioning by demonstrating that an elevated rNRE, 1)

is associated with previously identified CNS impairment, 2) is more common in children with FASD than their typically developing peers, and 3) may be found in children whether or not impairments are apparent on clinical assessments of Expressive Language function. Exploration of the clinical utility of TREIN measures based on Nominal Reference Errors and those based on cohesive errors in pronoun phrases was also conducted and indicated that tallies of Nominal Reference Errors like rNRE have more clinical potential in this age range than measures of pronoun errors. A strong developmental trend in mastery of nominal reference seen in the typically developing participants was absent in the FASD participant group, indicating an increasing clinical utility for identifying impaired children in the upper elementary school years. Result support further development of the TREIN and point to a need to better understand its performance in populations of typically developing children and children with a variety of CNS impairments, including those associated with prenatal alcohol exposure.

Table of Contents

List of Figures

List of Tables

Preface:

The long term goal of the research program begun here is to better understand how neurocognitive impairments commonly associated with prenatal alcohol exposure (PAE) disrupt the development of "integrative language capacities" during the school years. These capacities are central to effective communication and involve the ability to integrate contextual cues in a communicative/social context with communicative behavior to produce or interpret a message. They are frequently reported to be impaired in children with PAE that are diagnosed with Fetal Alcohol Syndrome (FAS) as well as those diagnosed with other Fetal Alcohol Spectrum Disorders (FASD). To understand the neurocognitive roots of these impairments in FASD we will need tools for measuring both the damage caused by PAE in the central nervous system (CNS) and for measuring integrative language functioning with precision. Currently, there are a number of options for measuring the impact of PAE on the CNS, but there are few well validated tools available for measuring integrative language functioning in school-age children with PAE.

In the present research we take the first step in the process of validating just such a tool for measuring one aspect of integrative language functioning by examining the performance of a group of school-aged children, some with typical development and some from a clinical FASD population.

Acknowledgements

I would like to acknowledge the support and sage advice provided to me during this process by my supervisory committee including Dr. Stacy Betz, Dr. Lesley Olswang, Dr. Susan Astley, and (especially) Dr. Truman Coggins. I would also like to acknowledge the support of Dr. Heather Carmichael Olson who shared her carefully gathered data with me. Without her generosity and support this project would not have been possible. Likewise, data generously shared with me by Dr. Astley, Dr. Olswang, and Dr. Coggins, and exhaustive labor shared by the various members of the University of Washington Child Language Laboratory and the University of Washington Fetal Alcohol Syndrome Diagnostic and Prevention Network made this project possible. Lastly and most importantly, I would like to acknowledge the support provided to me by my wife during this endeavor. Without her, none of this, or any of the even more important things in my life would have happened.

Dedication

This work is dedicated to my parents: Dr. Billy Joe Thorne, and Linda Kay Thorne who have provided me with inspiration, encouragement, guidance, and unconditional love throughout my life. I have never been an easy project to manage, but they have succeeded with unflappable grace and humor. No small feat by any standard.

CHAPTER ONE

INTRODUCTION

Fetal Alcohol Spectrum Disorders:

Fetal Alcohol Syndrome (FAS) is a permanent birth defect syndrome resulting from prenatal alcohol exposure (PAE) that is characterized by growth deficiency, a unique cluster of 3 minor facial anomalies, and evidence of CNS abnormalities [1, 2]. At a prevalence of 1 to 3 cases per 1000 live births, FAS is the leading known preventable cause of developmental and intellectual disabilities [3], and places a significant social and financial burden on communities. Lifetime costs for FAS are estimated at 2 million dollars per case [4], with estimated annual cost in the United States in excess of 5 billion dollars [5]. However, FAS represents only the most readily recognized of the Fetal Alcohol Spectrum Disorders (FASD). This is largely because its distinctive facial phenotype provides a specific diagnostic marker of prenatal alcohol exposure [6, 7]. Disorders on the fetal alcohol spectrum that lack the facial features of FAS are many times more prevalent than FAS (approaching 1% of all children), but share a similar range and severity of impairments and social costs [1, 8-10]. Neurocognitive and behavioral impairments that have been associated with FASD occur across a variety of domains. These have included attention/impulsivity [11, 12], motor and choice reaction time [5], response conditioning [13], fine/gross motor control and balance [14], visio-spatial learning [5], executive function and working memory [15, 16], mathematical reasoning [17], and non-verbal inductive reasoning [18]. Deficits in language and communication (broadly defined) are among the most frequently reported in the literature [17-27]. Peripheral and central hearing impairments appear to be common in children with the FAS facial phenotype [27, 28]. Children with FASD are frequently described as having particular difficulty with the processing of complex information [see e.g., 29] and responding to dynamically changing social situations [30].

Integrative Language Impairments In FASD:

As suggested by its very name, successful Integrative Language functioning results from the dynamic integration of a number of basic cognitive and linguistic functions within a changing communication context. The World Health Organization (WHO) defines *Integrative Language Functions* as "mental functions that organize semantic and symbolic meaning, grammatical structure and ideas for the production of messages" and contrasts them with *Expressive* and *Receptive* language functions [31]. Integrative Language functioning, therefore, involves the coordination between production or interpretation of an intended message and the changing social, linguistic, and situational context in which that message is communicated. Integrative Language capacities allow individuals to use their knowledge of a shared cultural tool (their native language) to choose communication behaviors that will meet communicative and social goals in the face of ever changing contextual demands; similarly they should also help individuals to understand the communicative choices of others in light of the current communicative context.

Theoretically [32-34], since Integrative Language functions involve coordination between local language structures and global features of an evolving communicative context, they would be expected to be impaired in parallel with similar cognitive control functions in other domains. Indeed, as pointed out by Hikosaka & Isoda [35], the ability to proactively switch between behaviors in light of changing cues from the environment "might be particularly important in social contexts: an animal (or human) is surrounded by many animals (or humans) that have different behavioral traits. It is then crucial to switch behaviors in anticipation of (rather than in response to) the other individual's behavior. Facial expressions, gestures, vocalization and gaze direction can provide many cues for switching, which the animal might need to learn to enable proactive switching."

Current research suggests that cognitive control capacities used to match local responses to evolving global/contextual needs are associated with a diverse network of fronto-striatal brain structures [e.g., basal ganglia including the caudate, frontal lobes: 36, 37-50], and that they have

developmental trajectories that extend throughout the school years [51-58]. For instance, according to a recent review by Hikosaka & Isoda [35], activation in the striatum is associated with the switching of response strategies when switching is based on abstract rules, and the striatum may be involved along with other frontal structures in the execution of behavioral switching—facilitating the selection of contextual appropriate behaviors while suppressing inappropriate behaviors. So, significant structural impairments to these fronto-striatal systems should theoretically result in important functional impairments in cognitive control capacities and difficulty matching behavior to contextual requirements. Given that these structures seem vulnerable to PAE [59], we would expect to see these types of cognitive control impairments more often in school-age children with FASD than in their peers. A growing body of research is coherent with this expectation in non-communicative domains [e.g., 59, 60, 61-67]. If these cognitive control deficits impair performance during communicative behavior, reduced ability to incorporate contextual demands into communicative behavior would be predicted to be the language capacity most vulnerable to PAE induced damage to these fronto-striatal systems. In other words, damage caused by prenatal alcohol exposure to the fronto-striatal structures that support proactive behavioral switching and cognitive control would be predicted to leave a trace in the Integrative Language behavior of these children.

Indeed, clinical reports frequently describe these kinds of concerns for children with FASD, who are often characterized as having poor social communication, poor social skills, and limited communicative success even when they perform within the average range on standardized language tests measuring Receptive and Expressive Language functioning [see, e.g., 68]. Since there is a dizzying array of surface level language behaviors that may be considered part of Integrative Language performance, and since impaired performance of any number of these behaviors may be at the root of these clinical impressions, the clinical challenge is to discover language behaviors that can be reliably and efficiently measured that can serve as behavioral markers of underlying CNS abnormalities.

Meeting The Challenge:

A proposal for validating behavioral measures of neurocognitive functions. Determining with certainty if there is a relationship between degraded performance for a specific language behavior and specific neurocognitive impairments is a non-trivial scientific problem (and certainly goes beyond the scope of the current project). The challenge comes from a basic problem shared by the study of any complex adaptive system: the fact that in a redundantly structured, interactive system, there are many possible ways for the system to produce a specific output (in this case communicative output). The more potential paths there are branching between the output of interest and the subsystems that support that output, the more difficult it is to create an inferential path from root cause to final outcome when damage is suspected [c.f., 69, 70]. When the surface level is complex, as is certainly the case with Integrative Language functioning, it is even more difficult to determine which overt behaviors depend upon which covert neurocognitive functions. Complicating the problem further is the fact that functional relationships that hold between surface level language behaviors may not be neurocognitively based, but may instead result from the fact that the language itself is a complex adaptive system that responds to the dynamic functional needs of the larger language community over time; a complication which calls into question the utility of analyzing surface level behaviors in isolation without reference to underlying neurocognitive differences.

Since scientific progress depends upon the study of substantially equivalent objects or processes, if there are a number of potential solutions available to the neurocognitive system for producing a specific language output, it may be inappropriate to equate individuals based solely on that output if the goal is to understand underlying neurocognitive sources of that output. The resulting heterogeneity can impede attempts to understand the underlying processes and dilute our ability to understand relationships between CNS damage and behavior.

This problem is exacerbated by the common clinical practice of functionally grouping individuals based on combinations of behavioral outcomes using "OR" rules. In this approach, since behaviors **a**, **b**, **c** and **d** are all thought to represent a more global functional construct (e.g., Expressive Language Functioning), then cases are said to present with a diagnostic phenotype (e.g., "Expressive Language Impairment") if they demonstrate deficits in behaviors **a** OR **b** OR **c** OR **d**. These "OR" rules, which are used to maximize diagnostic sensitivity for the global diagnostic phenotype, are also at the root of most standardized language test batteries, which sample a wide variety of language behaviors and treat errors on any particular item of the battery as equivalent to any other when calculating a standard score. While having clear clinical utility, these "OR" rules increase the heterogeneity of any group identified making their use of limited utility in understanding specific neurocognitive impairments. This is true even if groups are based on similar scores on a single language battery or subtest that uses the "OR" rule approach unless the global construct the test purports to measure is well understood and unitary in nature; a condition few language measures can claim to meet even if targeted at domains such as "semantics" or "syntax," that are clearly complex constructs. For example, verb-tense agreement and plurality marking may conceivably be neurocognitively distinct functions even though both are syntactic functions. If this is true, the "OR" rule approach combining these two types of errors in a single measure of syntax will increase sensitivity for identifying the broad category—people with neurocognitive impairments related to impairments of syntax—at the cost of conflating information about which specific neurocognitive impairment is present in a particular individual. This increases the heterogeneity of the group (in terms of both neurocognitive profiles and syntactic performance) when such a tool is used for case ascertainment.

Discussing this issue, Gottesman and Gould [71, see also 72] point out that the readily observable "phenotypic output of the brain" is not optimized for revealing the biological or genetic etiology of brain-based impairments. Coherent with the discussion above, they argue that—because "behavioral macros" or "exophenotypes" used to define groups (e.g., "Depression"

or “Expressive Language Impairment” or “Impaired Syntax”) end up being based on heterogeneous combinations of symptoms—they are too global and too far removed from any biologically based CNS abnormality to be scientifically useful. This remove exists because the pattern of behaviors that define exophenotypes result from the complex interaction of the elementary functions of a variety of CNS structures. There are too many potential causal paths to a particular behavioral “macro” to make an exophenotype useful in revealing a covert etiology, because an exophenotype conflates information about several, potentially distinct, impairments in the CNS. They recommend instead that researchers should, to be most productive, base clinical diagnostic groups on “endophenotypes” –which consist of “*putatively* more elementary,” very narrowly defined, (perhaps latent/covert) biological and behavioral markers—to increase the chances of demonstrating associations with specific root causes. Given that language functions are wholly brain-based, Gottesman and Gould’s discussion is directly relevant to research on language-based communication disorders. But, of course, not just any marker will do; for, as pointed out by Gazzaniga [73], “the trick for any level of analysis is to find the effective variables that contain all the information from below that are required to generate all the behavior of interest above.” An appropriately chosen endophenotype will, in theory, achieve this trick for a particular impairment.

Strictly, “endophenotypes” as defined by Gottesman and Gould are relevant only to disorders that are thought to have a genetic basis and according to them would need to meet several criteria: 1) associated with disorder (at the phenotype level) in the population; 2) heritable; 3) primarily state-independent, (i.e., can be measured in individuals for whom the disorder is not “active” with clinically significant severity; meaning that differences in the measure may be associated with a risk of developing the disorder in the future and/or a history of previous disorder status, now remediated – perhaps due to intervention/treatment or, in the case of developmental delays, additional development); 4) co-segregate with disorder in families; and 5)

must be found in non-affected family members at a higher rate than in the general population. These criteria, of course, are specific to the search for genetic markers of disorder.

When looking for behavioral markers of underlying CNS impairment in the context of an environmental exposure such as PAE, this list of criteria would need to be altered. For instance, despite the fact that genetic disposition certainly impacts vulnerability to certain exposures, "heritability" and "co-segregation within families" could drop from the list and condition 5 could refer to non-affected "individuals with similar exposures," rather than "non-affected family members." In the context of PAE, "non-affected" could be defined in a variety of ways including "those without full FAS," or "those without structurally identifiable brain damage," or even "those without an FASD." For our purposes, we will define "non-affected" as "those without full FAS."

With these alterations to the concept of "endophenotype" in place, however, it seems that the broader argument holds; valid language-based behavioral markers of underlying CNS impairments associated with PAE will meet these three conditions:

(i) are associated with disorder (i.e., are associated with some diagnosed impairment),

(ii) are found in individuals with PAE that do not have full FAS at a higher rate than in the general population, and

(iii) are state-independent (i.e. are found even when the individual's current performance would not lead to a diagnosis of a more general "Language/Communication Impairment" – but may predict either future or past difficulties of clinical significance).

The current research represents an early step in this process that attempts to clarify the degree to which a specific behavioral outcome is a valid behavioral marker of underlying CNS impairment by examining performance in children with previously diagnosed CNS impairments found during clinical assessment of suspected FASD.

Narrative Analysis And Impairment:

By definition, Integrative Language capacities manifest themselves during communicative discourse embedded in a particular context and allow individuals to adjust their communicative behavior to meet the needs of that context. Following guidance from the WHO [31], the measurement of any "capacity" requires that the context of measurement be standardized, so that performance across individuals can be compared. This creates a challenge for measuring Integrative Language capacities because communicative contexts are both dynamic and self-organizing—in other words, the language performance being measured alters the nature of the context dynamically over time with the current context being largely shaped by the specific discourse choices that occurred earlier in the discourse. Indeed, the capacity being measured is, largely, the capacity to respond effectively to these dynamic changes by recognizing how specific past behavioral choices have changed the current context, and to predict how current choices will impact the future context. This is complicated by the collaborative nature of most communicative discourse whereby the discourse choices of an individual are influenced not only by their own earlier discourse choices, but also in response to the discourse choices of their communication partner. Measuring performance during monologic tasks (i.e., tasks where only one individual in the discourse provides language input) reduces these complications somewhat, but even for monologic discourse tasks defining a standardized context can be challenging. For school-aged children, one common solution to this challenge has been to analyze performance during standardized narrative discourse tasks [e.g., 74, 75-77]. Narratives are chosen because they are a structured discourse genre which places relatively stable behavioral expectations on speakers when compared to other monologic discourse genres. These expectations manifest in a variety of structural features in the narrative which can be analyzed using a variety of techniques specific to the particular narrative task.

These stable structures of narrative discourse are frequently discussed in terms of either narrative microstructure—the organization of internal linguistic elements—or narrative

macrostructure—the organization of story elements to meet communicative goals [see e.g., 75]. For this discussion, we will also include an intermediate level of structural analysis and assume that narratives are organized around a three-tier system (see Figure 1.1).

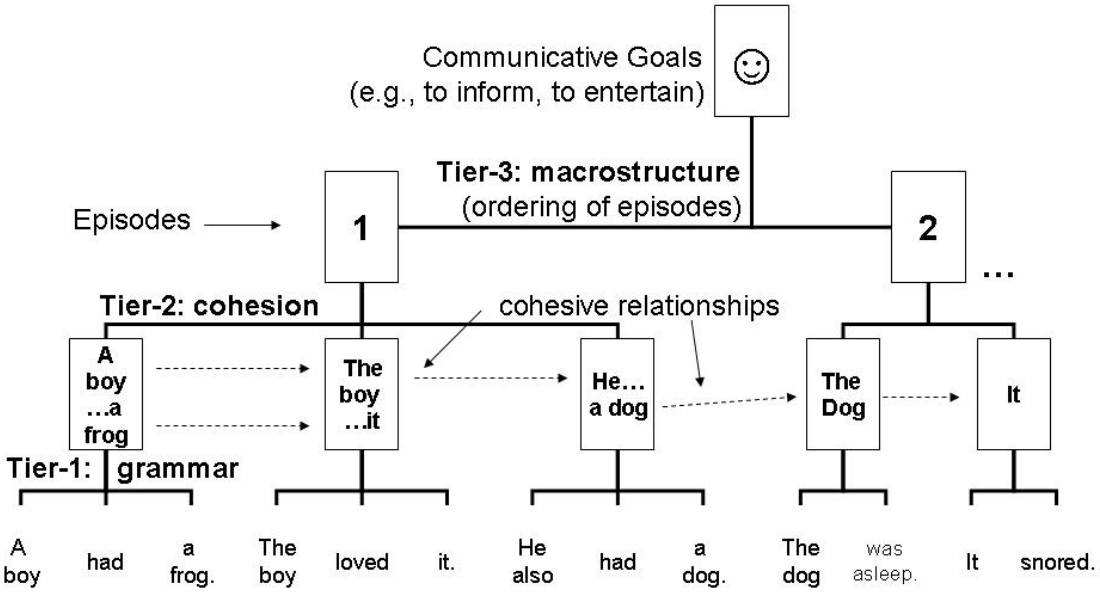


Figure 1.1: a Three-tiered structure of narrative: Tier-1 involves grammatical organization; Tier-2 includes cross-sentence cohesive relationships such as that between a pronoun and its antecedent ("a frog" to "it") or cross-sentence semantic relationships (e.g., between "a boy had" and "he also had"); Tier-3 involves the macro-structural ordering of a story's events.

To illustrate, consider the following: "*A boy had a frog. The boy loved it. He also had a dog. The dog was asleep. It snored.*" The microstructure of this chunk of narrative is built on the formal structure of the individual utterances as they conform to the grammatical conventions of the language being used. So, we get "*The dog was asleep*" rather than "*Dog the asleep was.*" Meanwhile, the semantic/pragmatic relationships that exist between/across these utterances constitute a second "Tier" in the structural organization of the narrative (i.e, "cohesion," following Halliday and Hasan [78]; c.f., Ariel [79]). Tier-2 cohesive relationships dynamically

respond to the needs of the listener as a series of individual utterances are organized into events or "episodes" that make up the story. So, in this case, we get "*The dog was asleep. It snored*" rather than "*They snored*," which would indicate more than one individual snoring. Similarly, we get "*He **also** had a dog*" to signal a similarity of status (i.e., "possessed by the boy") between the frog introduced in the first sentence and the dog introduced in the third. Finally, there is the macrostructural organization of these episodes into a narrative designed to meet specific communicative goals (i.e., "coherence"), which we will refer to as "Tier-3." For instance, a story teller may choose to withhold specific episodes in the narrative to heighten tension, or, conversely, introduce information early in the narrative to foreshadow important developments that will occur later in the discourse.

Hickmann [80] proposes that the cognitive capacities necessary for narrative production exist along three broad dimensions that parallel these structural elements of the narrative: 1) capacities supporting linguistic organization at or below the level of the sentence (Tier-1) which involves representation of propositional content according to grammatical conventions of the language, and 2) capacities supporting cohesive linguistic organization (Tier-2) at the level of discourse which integrate relationships across/between sentence-level propositions based on higher-order semantic and pragmatic demands, and 3) general all purpose "cognitive and/or communicative capacities" that support the strategic planning and goal setting needed to organize episodes (Tier-3) into a coherent narrative macrostructure [also see 79]. To put these capacities in the alignment with WHO terminology, Tier-1 depends primarily upon Receptive and Expressive Language capacities, while Tier-2 and Tier-3 depend upon Integrative Language capacities.

Hickmann's hierarchy of capacities is conceptually analogous to hierarchical models used to discuss cognitive control capacities that support non-linguistic cognitive tasks [see e.g., 38, 81, 82, 83]. These models recognize three levels of control above the most basic sensory-motor control level with higher-level executive capacities used in planning and goal setting sitting on the top of the hierarchy. In these models there is a distinction between those higher-level

executive capacities that organize action according to a relatively greater, more temporally dispersed set of information (referred to as “branching control”), and response control capacities that support the implementation of response plans in the current temporal frame.

Response control includes higher-level “episodic” control capacities and lower-level local control capacities (termed “contextual” control) that are thought to be supported by distinguishable neurocognitive networks in the frontal lobes of the brain, with higher-order control recruiting a wider and more anterior network of the frontal CNS structures [see e.g., 82, 84]. In this framing, local control allows individuals to match their responses to the immediate environmental demands of a particular situation based on currently available cues (perhaps processed according to hierarchical rather than temporal relationships in Broca’s area and it’s right hemisphere homolog [82]). So, for example, when crossing a busy intersection, local control capacities help us to process the traffic at the intersection to signal when it is safe to drive through the intersection (e.g., “no cars/pedestrians impeding my way, I can proceed safely”). Episodic control, on the other hand, allows us to choose our response to the local cues in light of additional information from “temporally distal” cues from the past that set the conditions for response during the current episode. This episodic control, therefore, helps us to dynamically shape a series of responses over a more extended period of time (e.g., “The light turned red before I arrived at the intersection. This means I should wait until the light turns green before crossing—even if it would be safe to do so—in order to avoid a traffic ticket”). Although they certainly interact, if episodic and local control systems are indeed independent, impairment of episodic control systems would produce a different class of performance errors than impairment of local controls systems whether the task was linguistic or non-linguistic. So, in our traffic example, impairment of episodic control systems may lead to safe-yet-illegal crossing of the intersection; while impairment of local control systems may lead to unsafe-whether-legal-or-not crossing of the intersection (see Figure 1.2). Impairment of systems for branching control, on the other hand, would make it difficult to insert a detour in the route to the store when a road-block or other

unexpected obstacle forces a change of plans (e.g., “In order to stay on my route, I can turn right here at this intersection, go left at the next, and after two-blocks get back on this road with another left and a right”).

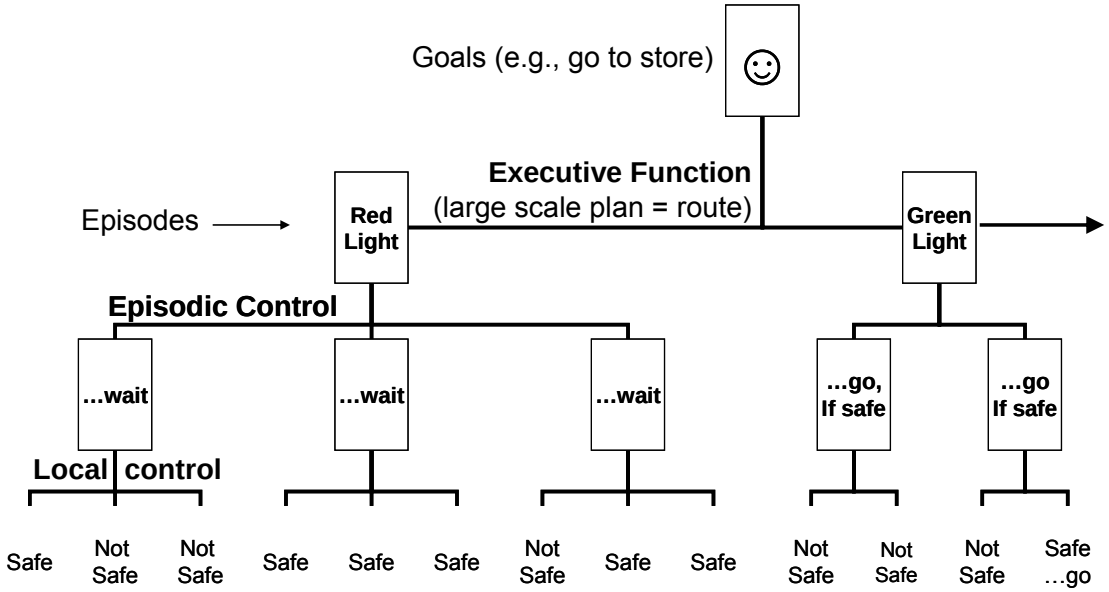


Figure 1.2: Cognitive control: Episodic versus local control of behavior.

Applying this to Hickmann’s hierarchy of narrative capacities, impairment of local control capacities would impact production of grammatically correct propositions, while impairment of episodic control capacities would degrade a child’s ability to respond to discourse-level factors (e.g., what has already been said) to integrate local sentence-level responses into a cohesive developing whole (see Figure 1.1). As such, episodic control in narratives is a prototypical example of an Integrative Language capacity that helps to maintain narrative cohesion between individual utterances. Episodic control, therefore, will primarily reveal itself in narrative cohesion (Tier-2). The Tier-3 macrostructural organization of the narrative (i.e., the narrative plan) will, meanwhile, reflect executive or branching control abilities that are also important for Integrative Language functioning.

Communicative Impairments and FASD. Given the heterogeneity of CNS abnormalities resulting from PAE, not all children with an FASD will have damage to the fronto-striatal systems thought to be involved in cognitive response control and executive functioning. As a result, even if these systems can be shown to be involved in narrative production, we would not expect all children with an FASD to exhibit Integrative Language deficits or have difficulty with appropriate use of Tier-2 or Tier-3 structures needed to produce cohesive and coherent narratives. However, as already mentioned, many children with FASD are reported to have deficits in social interaction and/or "social communication" [5, 18, 30, 66, 85-88], and as group seem to have the most difficulty with complex, later-developing aspects of language [89]. Indeed, many of these children do not show early language delay and perform within normal limits on standardized language instruments that focus on earlier developing Tier-1 aspects of language. However, because these tools measure language capacity using simple responses in Tier-1 (e.g., sentences, clauses, phrases, or words) they may not be sufficient for the evaluation of this population. They do not tap Integrative Language capacities (such as episodic control for maintaining cohesion) that may depend upon the fronto-striatal systems that seem particularly vulnerable to PAE. This may result in many children with meaningful communicative impairments involving Tier-2 and Tier-3 behaviors going unidentified.

The potential clinical advantage of using narrative analysis to tap into Tier-2 and Tier-3 behaviors as an addition to other standardized measures when assessing school-age children with FASD can be seen in the results from a retrospective survey of clinical records conducted by Coggins et al. [90]. They found that 149 (38%) of 393 school-aged children with FASD performed two or more standard deviations below the mean on standardized language measures—the level needed to demonstrate evidence of severe CNS impairment according to FASD diagnostic guidelines [1]. However, in a subset of these children assessed using narrative analysis, 201 out of 313 (64%, an additional 52 children) failed to produced extended narratives that were judged to be unambiguous and informative to their listeners or integrated structurally at age-

appropriate levels. Impairment in this domain of communication was more common in older school-aged children (8–13 years old). During a clinical narrative generation task, 145 of the 198 older children with FASD (73%) were judged to be performing below age-appropriate levels. Problems were seen in "both the referential aspects (i.e., representation of main story elements) and pragmatic aspects (i.e., ability to determine and convey relevant information) of narrative production" (c.f., Thorne et al.[91]; also, Thorne and Coggins [92]). These findings suggest that narratives are a language task that challenges many school-aged children with FASD and that narrative performance may indeed be sensitive to the type of CNS impairments commonly reported in these children. If impairment markers found with narrative analysis techniques can be shown to be (i) associated with disorder (ii) found in individuals with PAE that do not have full FAS at a higher rate than in the general population, and (iii) state-independent, they will have direct clinical application [1, pg. 38]. For reasons that will be discussed below, we have focused our search for such a narrative analysis measure in Tier-2 of the narrative structure.

Inferring impaired capacity from narrative microstructure – the case for measuring error. We can make some predictions about what kinds of narrative performance differences are likely to reveal underlying impairments. Impairments are likely to impede performance in two ways: (a) by reducing the rate at which certain desirable but optional features of the narrative occur (e.g., less complex utterances, smaller lexical diversity) and/or (b) by increasing the rate of errors in the narrative. Our initial research looking at both types of analyses [91] indicates that error rates may have greater potential for discriminating between typical and impaired populations [c.f., 77, 93]. This is likely due to the fact that desirable features of narrative are vulnerable to factors including motivation and creative choice (e.g., an unmotivated child may choose not to include a desirable feature they are capable of producing). For this reason, we have concentrated our efforts on the development of methods for measuring errors, which are less likely to appear in the narrative as the result of a conscious choice made by the storyteller.

In principle, errors will have a predictable relationship to more effective communication (with fewer errors in better narratives). Measurement of these errors will be most productive for features that are (a) highly frequent –providing a reasonable chance that sufficient numbers of tokens can be observed during a clinical sampling—and (b) obligatory—allowing for reliable judgment of the accuracy of performance. Highly frequent forms not only increase sample size, they also minimize the likelihood that lack of experience or exposure is a confounding element in performance, since children can be expected to have had massive exposure to these forms, and ample opportunity to gain knowledge about them from competent language users. If this assumption is true of a particular feature, one would predict that it would be produced with few if any errors in the age-range of interest, indicating that a typically developing neurocognitive system can easily meet the demands that this aspect of narrative places on it; both in terms of the demands required to learn the conventions for using the form, and the processing demands for deploying that knowledge during communication.

In a diagnostic context, this places an emphasis on the specificity (i.e., the "true-negative rate") of the behavior in relation to underlying neurocognitive impairment (broadly defined) and reduces the number of false-positives that result from particular performance criteria (because each false-positive reduces the number of true-negatives by one and vice-versa). Measures with high specificity provide a high degree of certainty that a measured positive is a true-positive (i.e., a measure with high specificity helps to confirm the existence of neurocognitive impairment). Of course, among features that demonstrate sufficient specificity for a particular use, those with the most sensitivity in relation to underlying neurocognitive impairment will provide for the greatest clinical and/or research utility by improving overall accuracy or "efficiency" (i.e., the probability that a particular criterion accurately classifies cases into diagnostic categories, thereby optimizing true positives/benefits relative to false positives/costs; [see 94 , pgs. 114-130 for discussion]).

Sufficient specificity is particularly important when multiple measures are combined using "OR" rules as a method of increasing sensitivity because false-positives across measures

will be additive in this situation—with some coming from one measure while others come from the additional measures. Once appropriate specificity has been established, determining what levels of sensitivity and efficiency are relevant is highly dependent upon the clinical or research context. When the target is a broadly defined and/or heterogeneous "exophenotype," such as "FASD" or even "FAS," even relatively modest levels of sensitivity and/or efficiency may have important clinical or research application. For any particular level of specificity, the measure with the most efficiency will be the one for which the distribution of performance in the impaired group overlaps the least with the distribution of the unimpaired group (i.e., it will have the smallest number of false-negatives in relation to the number of true-negatives). I would argue that the smaller this overlap between distributions is, the stronger the causal inference—increasing the likelihood that a causal relationship between a specific neurocognitive impairment and the surface behavior can eventually be demonstrated.

Choosing which errors to measure depends on your purpose. There are many highly-frequent and obligatory grammatical forms in Tier-1 of a narrative that could be used to reveal language impairments involving the Expressive and Receptive language capacities that support production of grammatical sentences. Of course, since grammatical constraints operate at or below the level of the utterance, they are not sensitive to the fact that a narrative is an extended discourse. Indeed, it may be more clinically efficient to measure Tier-1 grammatical errors using narrowly designed tasks that elicit word, phrase, or sentence-level responses containing a particular grammatical form, a strategy commonly used in standardized language batteries. If, on the other hand, the goal is to measure Integrative Language capacities [c.f., discussion in 95], it will be necessary to identify obligatory and highly frequent features of the discourse that function in Tier-2—at the level of the episode—or Tier-3—at the level of narrative macro-structure. Because they are defined by larger chunks of language (i.e., episodes), Tier-3 structures are not as frequent in an individual narrative as Tier-2 structures (see Figure 1.1), making Tier-2 structures a more attractive target for the purpose of revealing Integrative Language capacities. Of course,

these Tier-2 structures will need to be measured in a standardized context designed to elicit cohesively organized multi-utterance responses. Narrative tasks can be designed to provide such a standardized context.

Discourse obligations, and reference as a sign of Integrative Language capacity. While Tier-2 includes many highly frequent forms (e.g., conjunctives used to mark semantic relationships between sentences), there is not universal agreement as to what constitutes an obligatory Tier-2 feature of a narrative. Among the more well-studied candidates are unambiguous referring expressions [32, 78, 80, 96]. Of particular interest for school-age children is the development of capacities for achieving "endophoric" reference, whereby a linguistic form refers cohesively to a discourse-internal referent rather than one available in the environment [78]. Hickmann [80] notes that achieving cohesive endophoric reference requires attention to whether or not a chosen form provides the appropriate marking of informational status (e.g., marking whether the concept is *new* in the discourse, or *previously mentioned*). Hickmann identifies this feature of discourse as both obligatory—information status of referents is marked in all languages, in all instances—and variable in the way it is marked formally across languages [see also, 34, 97]. In English, this obligatory marking occurs in the nominal phrase (noun or pronoun), making it both frequent and readily identifiable. It is important to re-emphasize that forms which are inappropriate due to Tier-2 features of the context may be perfectly grammatical when considered at the sentence level. For instance, the sentence "It was in there" is grammatical, but may be inappropriate in an episode where the listener can identify neither what "it" is, nor where "in there" is. Without supporting information in previous discourse, this grammatically correct sentence is uninterruptable or ambiguous; there is a Tier-2 obligation controlling which forms are needed to achieve endophoric reference. Meeting this Tier-2 obligation in a dynamic communication context requires episodic response control that takes past information into account and may be particularly challenging to master [98].

Wong and Johnson [96] emphasize two types of information used by storytellers to determine information status in order to achieve endophoric reference: 1) their listeners' current knowledge of the intended referent; and 2) their listeners' relative attention state towards that referent [see also 97, 99, 100]. When integrated into the larger context, these two pieces of information help to define for the storyteller what is referred to as the "common ground" shared by participants in a particular communicative context. This common ground is made up of the shared memory of the speaker and the listeners about their on-going discourse in addition to general world knowledge (which is relatively stable across contexts and most communication partners). To achieve cohesive endophoric reference, storytellers must choose referring expressions that are "anchored" [101] to this dynamic and ever-changing common ground shared with their listeners. As endophoric referencing requires the storyteller to choose the current response based on information from past discourse events, anchoring the current statement to the common ground requires episodic control. With that control in place, a properly anchored referring expression can be chosen that will key listeners' attention to the concept the speaker deems necessary to advance the larger narrative plan. This process of anchoring the current utterance to the common ground is pervasive in communication and central to discourse cohesion. Indeed, even children two-years old or younger seem to use assumptions about common ground to guide their interpretation of linguistic information during communicative discourse [see, e.g., 102, 103, also, 104].

Among the conventions that English speaking school-age children learn in order to cohesively anchor their references to the common ground are the proper uses of definite (e.g., *the*) and indefinite (e.g., *a/an*) articles in noun phrases [80]. These articles are among the most frequent words in English and help to mark both the specificity and the informational status of concepts. For example, (whenever a non-verbal cue such as pointing or eye-gaze is not available) indefinite articles are used to mark the *introduction* of new concepts into the discourse via an anchor to the common ground available in general knowledge. Tier-1 grammatical constraints

(e.g., subject versus object position) will determine whether that new concept is specific or non-specific [105, 106]. For example, in the sentence "He wants *a dog*," the indefinite noun phrase refers to a non-specific member of the general class of dogs (i.e., any dog will supposedly fulfill his desire for the purpose of the current discourse). Because the concept is non-specific, it can not be specifically referred to later in the narrative (otherwise desire would be more specifically characterized by mentioning, for instance, a specific individual dog—"he wants a dog that he saw at the pet store yesterday"). However, in "*A dog* was stolen from the pet-store" the indefinite noun phrase introduces a new and specific dog (i.e., $DOG_{specific}$) into the common ground of the discourse and that $DOG_{specific}$ can be referred to later in the narrative. Meanwhile, definite articles mark a *reference to* a specific concept that is already anchored to the common ground based on a previous mention in the discourse or a logical inference that can be assumed to flow from previous discourse. So, while the sentence "*The dog* was stolen" would be grammatically correct in isolation, its proper use is constrained by its informational status—its use assumes that the concept is specific and is already part of the common ground. Children may also use a pronoun form to refer to the concept $DOG_{specific}$ (e.g., "it was stolen" where "it" = $DOG_{specific}$). The pronoun will again be grammatically correct, but only appropriate in specific discourse contexts; in this case, contexts where the concept $DOG_{specific}$ is not only *available* to the listeners in the common ground, but is also the current focus of their attention [see e.g., 107].

Herein lies the challenge for younger narrators: to use noun phrases and pronouns cohesively for endophoric reference, narrators must know (at least) whether a nominal concept is specific/non-specific (which comes with Tier-1 grammatical constraints on form) *and* must be able to continually take into account the dynamic common ground they share with their listeners as new concepts are introduced and move in and out of primary focus (with Tier-2 constraints on form). Because they entail different assumptions regarding their specificity and information status, proper nouns, mass nouns, and generic forms introduce additional complications into Tier-2 organization and require children to learn additional conventions for how articles are used in

nominal phrases to anchor these concepts to the common ground. It is apparent that different aspects of this task present greater or lesser challenge to children as they learn the necessary conventions and attempt to implement them to create a cohesive narrative: with accurate Tier-1 grammatical marking using articles mastered by 3 or 4 years of age [105]; endophoric marking of a “discourse new/previously mentioned” distinction in Tier-2 largely mastered by (perhaps) entry into elementary school [80, 108]; and Tier-2 marking of the more subtle aspects of mutual knowledge (e.g., assumed focus) not typically mastered until age 10 or 11 [80].

CHAPTER TWO

DEVELOPING A NARRATIVE ANALYSIS SYSTEM

Tallying Reference Errors In Narrative:

At this writing, there is not a well-validated system for isolating and quantifying the adequacy of school-age children's use of cohesive endophoric reference in narratives, despite the availability of several useful tools for quantifying other aspects of narratives (see, e.g., Justice et al. [75]). Indeed, Justice et al. specifically did not include referential cohesion in their *Index of Narrative Microstructure* due to concerns that this kind of analysis would not be "amenable to reliable field-based use by clinical professionals" despite recognition that it would "provide significant information concerning children's narrative performance." This concern may come from the complex nature of previous systems proposed for quantifying cohesion (including endophoric references) in narratives [see, most notably, 109].

The current work is aimed at development of an efficient, valid, and clinically useful method for tallying these errors of endophoric reference in the discourse of school-age children. This will be done keeping an eye on the three criteria proposed above for valid behavioral markers of underlying CNS impairments:

(i) are associated with disorder (i.e., are associated with some diagnosed impairment),

(ii) are found in individuals with PAE that do not have full FAS at a higher rate than in the general population, and

(iii) are state-independent (i.e. are found even when the individual's current performance would not lead to a diagnosis of a more general "Language/Communication Impairment" – but may predict either future or past difficulties of clinical significance).

The system, *Tallying Reference Errors in Narratives* [TREIN; 110], is designed to identify Tier-2 performance errors during narrative production (for complete details, see the

TREIN manual, available on-line at http://johncthorne.wordpress.com/tallying-reference-errors-in-narrative-trein/). The TREIN concentrates on forms which mark informational status of referents (*discourse new* versus *previously mentioned*) as they are introduced, maintained, and reintroduced across the developing narrative and provides, along with other measures, a tally of "Nominal Reference Errors" (NRE; i.e., failures of endophoric reference in noun phrases). Previous research suggests that these errors should be rare in the narratives of typically developing children who have reached elementary school age increasing the chances that they will provide a specific marker of underlying neurocognitive impairment. It is also designed to be both efficient enough and reliable enough for use by clinicians in the field.

The TREIN protocol obligates coders to use Tier-2 features of the discourse to exhaustively categorize each nominal phrase or pronoun produced for a "naïve" listener in a structured narrative task into one of nine categories: five categories for "appropriate" reference strategies; and four categories for "reference errors" (two for pronominal phrases and two for nominal phrases). An overview of TREIN codes is presented in Table 2.1.

Table 2.1: Summary of TREIN codes.

Codes for Introduction	
+ [indefintro]	*Indefinite introduction* of concept (e.g., "**A boy** was looking…")
+ [defintro]	*Definite introduction* of concept with supporting contextual factors. (e.g., "**The moon** was out…")
+ [possintro]	*Possessive introduction* of concept. (e.g., "**His dog** was with him.")
- [ambigintro]	*Ambiguous introduction* of concept using a definite form not supported by contextual factors. Also used for an inappropriate 2nd use of an indefinite form.
- [pnintro]	*Pronominal introduction* of concept. (e.g., "**It** was in there" on first mention of "it.")
Codes for referential maintenance (i.e., "reference ties," see Halliday & Hasan [78])	
+ [ntie]	Clear *referential tie* using nominal form. (e.g., "**A boy** had a frog. **The boy** liked it.")
- [ambigntie]	*Ambiguous referential tie* using nominal form. (e.g., "He saw **two frogs**. **The frog** was…"; also mislabeling available concepts: "dog" for FROG)
+ [pntie]	Clear *referential tie* using pronominal form. (e.g., "**A boy** had a frog. **He** …")
- [ambigpntie]	*Ambiguous referential tie* using pronominal form. (e.g., "**The boy and the dog** were looking for it. **He** found it in the woods.")
"+" indicates an appropriate strategy while "-" indicates an inappropriate strategy.	

The tally of each of these errors can then be used to calculate the rate at which they appear in the narrative. For instance, in order to determine "rate of nominal reference errors" (rNRE), two of the nine TREIN codes are used: 1) codes for "*ambiguous introductions*"; and, 2) codes for "*ambiguous nominal ties*." First a tally, Total Nominal Reference Errors (NRE) = Ambiguous Introductions + Ambiguous Nominal Ties is determined. In order to control for variability in the length of narratives, NRE is then used to calculate the *Rate of Nominal Reference Errors* (rNRE). The rNRE = NRE / total words (TW) in the narrative with TW calculated using the *Systematic Analysis of Language Transcripts* [SALT; 111] function, "number of words in analysis set." Of course additional outcome measures can be generated by a TREIN analysis. As we can expect narrative length to impact the value of all raw error counts, only error rate measures which adjust for length (like rNRE) would be considered logical candidates for measuring Integrative Language performance.

Preliminary Studies:

Four studies previously completed in the development of the TREIN are briefly summarized below. Since the first three of these studies share common participants [91, 92, 112], materials, narrative collection/ transcription procedures, and analyses, these are presented first, followed by the main findings from each. This will be followed by summary of an additional study involving an additional 21 subjects [113]. Across studies, results demonstrate high interrater agreement when using the TREIN protocol to identify NRE with a reasonable amount of training (approximately 10 hours).

MATERIALS:

All three of our initial studies were conducted in a retrospective manner using the same 32 narratives. These child-generated narratives were selected based on age of storyteller from two independent sources: an intervention study involving children diagnosed with a FASD

experiencing behavioral problems [114]; and a normative study of typically developing (TD) school-aged children [115].

PARTICIPANTS:

Thirty-two participants (16 FASD; 16 TD) were matched on age which ranged from 8;5 years to 11;7 years (mean: 9;11 years). The family income, gender, and ethnicity of the groups were similar to each other and representative of the Seattle metropolitan area (see Thorne et al. [91]). Participants with FASD were diagnosed by an interdisciplinary team using the *FASD 4-Digit Code* [1] at the University of Washington Fetal Alcohol Syndrome Diagnostic & Prevention Network (FASDPN). The first three digits in the *4-Digit Code* are used to characterize the degree of impairment found across 3 domains important for identifying FAS. These domains (in order from left to right) are growth, FAS facial features, and degree of CNS impairment. The fourth digit in the code documents the degree of prenatal alcohol exposure. The higher the number indicated in each category, the greater the severity in that domain. So, for instance, a code of 4444 would indicate full FAS based on severe growth deficiency, the full facial of FAS, definitive evidence of CNS impairment, and confirmed heavy prenatal alcohol exposure.

NARRATIVE COLLECTION & TRANSCRIPTION:

All 32 narratives were elicited using the wordless picture book, *Frog Where Are You* [116]. After the child previewed the storybook, the examiner asked the participant to tell the best story possible while using the picture book as a visual prompt. Examiners were always seated across the room from the child to make it clear that they were unable to see the storybook pictures. Narratives were recorded on audiocassette and orthographically transcribed by trained graduate students. Analytical coding of narrative transcripts in all studies was conducted blind to any characteristics of the story teller.

ANALYSIS OF CLINICAL UTILITY:

For the first two studies, signal detection methodology was used [94, 117]. Visual inspection of Receiver Operation Characteristic Curves (ROC) identified narrative outcome measures with diagnostic potential, while Area under the ROC (AUC) was used as a summary effect-size measure of potency for discriminating between FASD and TD narratives [118, 119]. A measure was considered to have sufficient potential to warrant further development if the lower bound of the 95% confidence interval for AUC fell above 0.70 (i.e., 25% overlap or less) and the ROC curve did not cross the "random test" line.

Study 1: Elaboration & Ambiguity in Narratives; Thorne et al. [91]

Primary question: which of our elaboration and ambiguity measures accurately predict which narratives were produced by children with an FASD and which were produced by children with typical development (TD)?

Results: This study used ROC curve analysis to test a set of 26 narrative outcome measures (e.g., number of verb/nominal modifiers, number of specific verbs/nominals, see [120] for details). The most important finding was that while elaboration measures (which quantify desirable features) predicted performance on standard language measures for the FASD group, a focused measure of ambiguity (a measure of error) could accurately predict which narratives were produced by children from the FASD group and which from the TD group. The *rate* of *ambiguous nominals* (calculated as a percentage of total words) was able to correctly classify the narratives of 26 of the 32 children (81%, AUC = 0.86; 95% CI 0.70-0.96; 88% sensitivity, 75% specificity, $p < 0.0001$).

Study 2: *Tallying Reference Errors in Narrative* (TREIN); Thorne & Coggins [92] .

Since only a single ambiguity code from our original system met criteria for clinical potential, a refined system designed to capture this aspect of narratives was developed. The system discussed

above, *Tallying Reference Errors in Narrative* [TREIN; 110], focuses exclusively on reference strategies within a narrative and allows narratives to be classified based on the *rate of Nominal Reference Errors* (rNRE). While the rate of ambiguous nominals used in our initial study relied on a reader's gross judgment that a nominal form was "ambiguous", the TREIN carefully defines when a nominal phrase should or should not be considered ambiguous (i.e., a "*Nominal Reference Error*"; see discussion above).

Primary questions:

1) Does rNRE differentiate FASD from TD narratives as accurately as the *rate of ambiguous nominals*?

2) Among children with FASD, can rNRE discriminate between those with/without FAS facial features?

Results: The rNRE was more accurate and more reliable *(Kappa = 0.90)* than the rate of ambiguous nominals used in Thorne et al. [91] correctly categorizing 28 of the 32 narratives (88%). Typically developing children obtained rNRE that stayed at or below 3% of words in the narrative, while some children with impairments had rates approaching twice that figure (5.7%; see Figure 2.1). Compared to the rate of ambiguous nominals, sensitivity for FASD improved by 6% reaching 94% and specificity improved 6% reaching 81% (AUC = 0.90, 95% CI from 0.73 to 0.97, $p < 0.0001$). The rNRE accurately categorized all children with FAS facial features (AUC= 0.98, sensitivity = 100%, specificity = 93%, 95% CI from 0.85 to 0.99 $p<0.0001$) who had the highest rNRE of any group (mean 4.36%, SD 0.8).

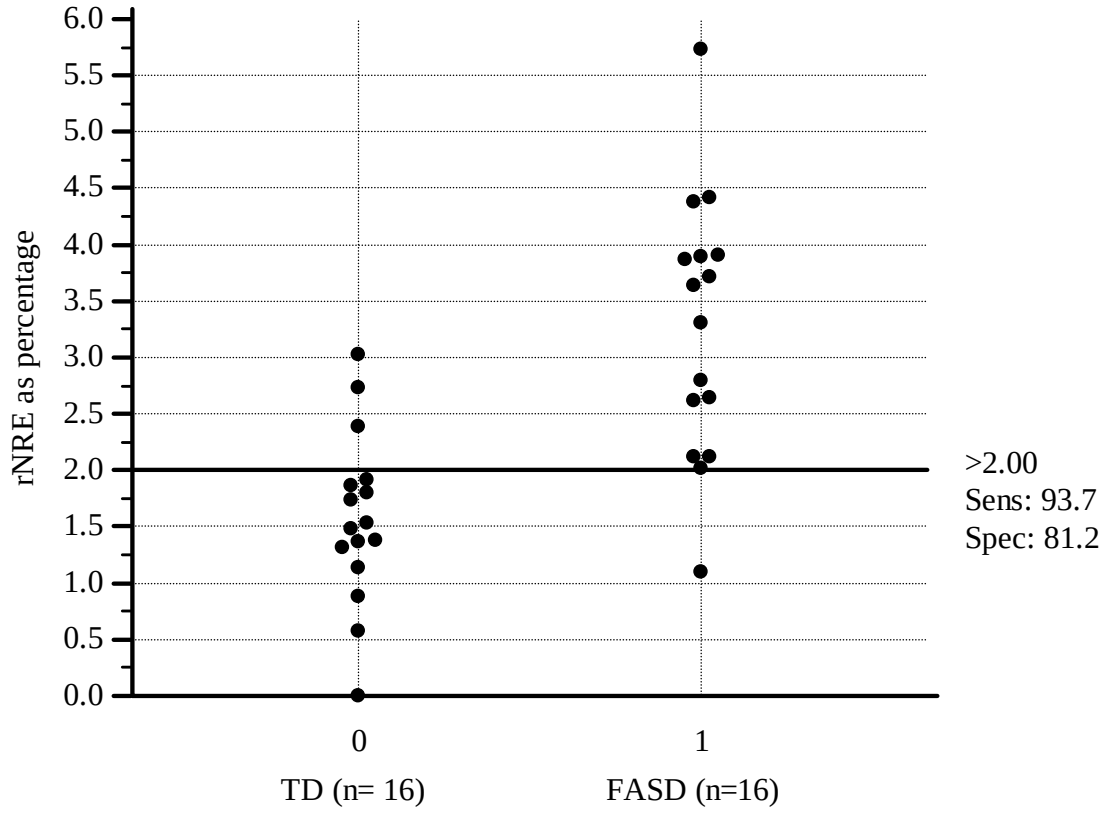


Figure 2.1: rNRE for TD and FASD groups in original 32 narratives with sensitivity and specificity for predicting FASD status at best cut-point

THE TREIN AND CNS ABNORMALITY

Study 3: Thorne & Coggins [112]

This study explored the relationship between narrative performance and clinical evidence of structural CNS abnormality in the same sample of 32 children used in the previous studies. In addition to the nine cohesion measures generated by TREIN analysis used to address the primary questions, 26 measures (e.g., Type-Token Ratio; Number of Different Words; Total Utterances, standard word types) were generated with *Systematic Analysis of Language Transcripts* [111] software and were used to explore narrative performance more broadly.

Primary Questions:

1) Does performance on *Frog* narratives (as measured by a TREIN analysis) become more impaired with increasing evidence of structural CNS abnormality (as measured by the FASD CNS RANK 1-4)?

2) Do important differences (i.e., > 2 SD) in narrative performance discriminate between children with FASD that have *Static Encephalopathy* (i.e., significant CNS abnormality) and their TD peers in this sample of 32 children?

Secondary Question: What is the relationship between severity of FASD diagnosis and narrative performance in the 32 school-age children in this sample when additional narrative measures are also included in analysis?

To answer the primary questions, children were grouped into four CNS RANKS using criteria from the FASD *4-Digit Diagnostic Code* [1]. FASD CNS RANKS are based on evidence of underlying structural CNS abnormality found during a comprehensive interdisciplinary diagnostic evaluation. CNS RANK 4 indicates "definite" structural CNS abnormality based on direct structural/neurological evidence (e.g., microcephaly, seizures). CNS RANK 3 is assigned when functional evidence indicates "significant" impairment in three or more domains of brain function (i.e., a performance deficit equivalent to 2 SD or more from the mean on a standardized test). A child with CNS RANK of 3 or 4 is diagnosed with "*Static Encephalopathy*." CNS RANK 2 is assigned when the child exhibits at least mild to moderate delay or impairment in some domain of functioning, but does not qualify for a CNS RANK of 3. CNS RANK 1 indicates that the individual does not meet criteria for RANKS 2 through 4. Although this system is an indirect indicator or CNS impairment, Astley et al. [61] confirms that as CNS RANK moves from 1 to 4 there is increasing probability of underlying structurally identifiable CNS abnormality. All children in the TD group were assigned CNS RANK 1, while children in the FASD group were CNS RANK 2-4.

Selected Results:

Primary Question 1: One-way ANOVA revealed two of the nine TREIN measures to have significant group contrasts. Groups defined on CNS RANK differed in their mean rNRE (F= 15.8, p= 0.0001) and their mean number of Indefinite Nominal Introductions (a category of appropriate nominal reference, essentially the opposite of a Nominal Reference Error; F=5.9, p= 0.003; see Table 2.2).

Table 2.2: Mean (SD) & planned contrasts (Scheffe's Multiple Comparison Test) by CNS RANK for INI and rNRE (Preliminary Study 3).

	Indefinite Nominal Introductions		**rNRE (rate of Nominal Reference Errors)**	
CNS RANK	**mean** (SD)	**Significant Contrasts**	**rNRE** (SD)	**Significant Contrasts**
1-*unlikely* (n=16)	16.7 (3.6)	1>4 (p=0.003)	1.57% (0.7)	1<4 (p<0.0001); 1<3 (p=0.006)
2-*possible* (n=8)	15.4 (3.5)	2>4 (p=0.03)	2.59% (0.9)	2<4 (p=0.01)
3-*probable* (n=4)	14.0 (2.4)	ns	3.40% (1.2)	3>1 (p=0.006)
4-*definite* (n=4)	8.8 (2.8)	4<2(p=0.03); 4<1(p=0.003)	4.48% (0.9)	4>1 (p<0.0001); 4>2 (p=0.01)

Primary Question 2: Using a cut-point of 2 SD from the mean of the TD group, the best measure for discriminating between children with FASD that have "*Static Encephalopathy*" and their TD peers was rNRE (sensitivity=88%; specificity=92%). This was the only measure to show clinical potential using Plante & Vance's criteria (i.e., 80% sensitivity & specificity;[121]). All false-positives came from the FASD group with CNS RANK of 2. A total of 56% of all children with FASD had rNRE above the 2 SD cut-point.

Secondary Question: Children diagnosed with FAS had the highest rNRE (4.36%, SD 0.8), children with TD had the lowest (1.57%, SD 0.7), and children with FASD without FAS facial features (i.e., other FASD) fell in between (2.77%, SD 0.9) with little overlap between groups (F= 23.0, p< 0.0001; Scheffe's contrasts: FAS>TD, p<0.0001, FAS>other FASD, p=0.005, other FASD>TD, p=0.004). All 5 children with FAS and an additional 4 with other

FASD had rNRE more than 2 SD from the mean of the TD group (1.57%, SD 0.7). The group "*Static Encephalopathy*" had significantly smaller Number of Different words (NDW; mean 92.1, SD 25.8) than the group "*CNS RANK 1 or 2*" (mean 125.8, SD 38.6; t-test -2.29, p=0.03). They also produced fewer cohesive pronoun ties (16, SD 8.6) than the "*CNS RANK 1 or 2*" group (mean 24.4, SD 11.6; t-test -2.18, p=0.04), but no cases produced NDW or cohesive pronoun ties more than 2 SD below the mean of the TD group (NDW=130, SD 43.1; cohesive pronoun ties =27.4, SD 12). No other measures produced important contrast between groups.

KEY FINDINGS OF THE THREE PRELIMINARY STUDIES:

In the 32 *Frog Where Are You* narratives initially examined (16 FASD; 16 TD)—

1) The group with FASD had significantly impaired performance compared to the TD group.
2) Of the measures examined, the TREIN measure rNRE appears to be the most sensitive and efficient indicator of impaired performance for use in differentiating FAS from other FASD, and all FASD from typically developing peers.
3) The rNRE increased significantly with increased risk of structural CNS abnormality as reflected in both the FASD CNS RANK (RANK 1-4) and severity of FASD diagnosis (FAS>other FASD>TD).

Expanding The Sample Of Narrative Examined:

As is apparent from Figure 2.1, our initial research on the TREIN [92] indicated very little overlap in the distribution of rNRE for the children with CNS impairments and those with typical development. While this was an encouraging result, the within-group distribution of rNRE raised several questions stemming both from inherent limitations of available data in this retrospective study, and from theoretical predictions about the expected range of performance in

the children with FASD diagnoses. The first question could not be directly addressed in our original studies due to a limitation in data available for 16 children in the original TD group, which did not include any standardized measures of language or cognition. Because it makes it impossible for us to confirm the developmental level of these children—this raises the possibility that an unexpected proportion of these children are generally high performers. There is some evidence to support this idea in the within-group distribution of scores as only 3 of the 16 in the TD group produce rNRE above 2%, with a noticeably sparse distribution between 2% and the maximum rNRE in the group of 3.03%. Theoretically, if 3.03% rNRE is on the high-end of the true population distribution (e.g., +2 standard deviations from the TD mean), we would expect to see this sparse region of the distribution begin to populate as more narratives are examined. This would, of course, increase the percent overlap between the TD and impaired groups. If, however, the original 16 children are indeed a high-performing group, we may also expect to see significant numbers of TD children with scores substantially above their 3.03% maximum. A range extending substantially higher than this would call into question the clinical potential of the measure for this age range, because this potential is premised on the idea that these errors are rare in typically developing children who have reached the elementary school years.

A similar gap seen in the distribution of scores in the 16 children in the original FASD group also raises questions. This gap, however, is on the lower end of the rNRE distribution with only 1 of the 16 children from the original FASD group performing below rNRE of 2%. Given the heterogeneous nature of CNS impairments seen in children with FASD, however, some children would be expected to have impairments unrelated to the episodic control capacities needed to maintain cohesive narrative reference. In essence, more overlap between the TD and FASD groups at the bottom of the rNRE distribution would be predicted as more narratives are examined (i.e., we expected reduced sensitivity at a +2SD cut-point). Therefore, we predicted that as more children with FASD were added to the bottom of the rNRE distribution, we would see groups means moderating somewhat, but clinical utility remaining high. For this to happen, the

distribution of rNRE in children with CNS impairment needed to include substantial numbers outside the typical range, and greater proportions of children having high rNRE found in groups having increased risk or severity of impairment. This would help us to establish the strength of the association between rNRE and underlying impairment and begin to demonstrate the potential validity of the measure for indentify that impairment in elementary school aged children.

Study 4 Thorne, Coggins, Grittner, Olswang [113] :

Study 4 explored these issues and tested three specific predictions related to the distribution of rNRE across expanded TD and FASD groups. A brief description is provided below. See the Appendix for additional descriptive information related to this study.

Methods:

New Participants:

To extend our initial corpus, we used a convenience sample of narrative productions from a previous case-control matched-pair study of two groups of school-age children: one group with identified CNS impairments associated with FASD; and one group enrolled in regular education classroom with no cognitive, academic, or behavioral concerns. Because this sample includes identical measures for both groups of children, it allows us to probe individual differences between children and to confirm the developmental status of each child. The group expanded our pool of narratives to a total of 53 by adding an additional twenty-one participants, all elementary school children between the ages 7;5 to 11;8 years. Narratives were collected during the source investigation using the same procedures used in our original research. Eleven of the new participants had a CNS impairment associated a diagnosed FASD. These eleven children were also identified as having clinically meaningful social problems, but were, nonetheless, enrolled in regular education classrooms. The remaining ten children were considered typically developing

(TD) and were matched to their FASD counterpart on several key characteristics (described below).

Fetal Alcohol Spectrum Disorders Group. As with the participants in our initial research [91, 92], the 11 participants with FASD came from the University of Washington's existing data-base of children with a confirmed prenatal alcohol exposure. All were diagnosed by an experienced interdisciplinary clinical assessment team using the *4-Digit Diagnostic Code* [1]. Participants in the FASD group were selected using the following *4-Digit Code* criteria: Confirmed prenatal alcohol exposure = RANK 3 or RANK 4; CNS RANK indicating impairment = RANK 2 or higher, which could potentially include impairments ranging from deficits between one and two standard deviations below a normative mean in a single area of functioning up to clear structural abnormalities and/or severe dysfunction across multiple domains [1]. Growth and facial feature rank were left open. Based on functional severity scores, all 11 children exhibited mild to moderate CNS impairment despite 2 children receiving a CNS RANK of 4 due to microcephaly.

Typically Developing Control Group. Typically developing controls were classmates of the FASD participants. In each case, the respective teachers were asked to choose a classroom peer who was "as close a cognitive match as possible" to the FASD participant [122]. Children in this group did not present with any academic concerns and caregivers did not *endorse* any of the following: attention deficits/hyperactivity, behavioral/emotional problems, learning problems, speech/language problems, or trouble making friends. Each TD participant was matched to an FASD participant based on gender and chronological age (mean difference $\pm$ 6.2 months, range $\pm$ 0-18 months). TD participants were excluded if they had a Composite IQ score more than -1.0 SD from the mean on the *Kaufman Brief Intelligence Test* [K-BIT; 123] and or receive scores on the SSRS-PB in the clinical range as rated by their teachers.

Narrative Data:

Inter-Coder Reliability. All narrative coding was conducted blind to group membership (TD or FASD) age or gender of the storyteller. Two coders independently coded all 21 narrative transcripts. These independent codes were used to compute inter-coder reliability. The coders reached a kappa of 0.89, differing on 238 of 1,457 coding decisions (16.3%). For all disagreements, the coders reached consensus on the appropriate code. These consensus codes were complied in master transcripts and were used in computing the rate of Nominal Reference Errors (rNRE).

Analysis and Results:

Analyses were conducted to test three specific predictions related to the expected rNRE distributions of TD and FASD groups. The predictions tested were as follows:

Prediction One: The maximum rNRE of 3.03% found in our original study will fall +2 standard deviations above the mean of the pooled TD group.

Results for Prediction One. The mean rNRE for the pooled TD group was 1.89% with a standard deviation of 0.953%. Therefore, the +2 standard deviation point for the pooled TD group fell at an rNRE of 3.81%. This result did not validate our first prediction. Four of the newly added TD participants had an rNRE that fell in the 2-3% range, and two had scores that exceeded the 3.03% maximum from the original study. The majority of TD children maintained rNRE below our predicted maximum of 3.03%, which ended up at the 90th percentile for the combined group (95% confidence interval 2.83% to 3.21%).

Prediction Two: More than 50% of children from the pooled FASD group will have rNRE above at least one of three common clinical cut-points including the 90th percentile, +1.5 and +2.0 standard deviations above the mean of the pooled TD group.

Results for Prediction Two. Table 2.3 presents three clinical rNRE cut-points used to test prediction two. More than 50% of children in the FASD group fall above one of our clinical cut-off scores. This finding supports prediction two. Indeed, if we use rNRE to predict which narratives were produced by the pooled FASD group, rNRE remains an accurate predictor that is significantly better than a random test (AUC 0.77; 95% CI 0.63 to 0.87; p<0.0001), despite the increased range of rNRE seen in the new larger TD group.

Table 2.3: Mean, Standard Deviation (SD) and classification measures for 53 children at 3 cut-points based on scores from pooled TD Group (TD, n = 26; Preliminary Study 4).

	TD Mean	TD SD	+2 SD cut	+1.5 SD cut	90th percentile[a] cut	Lowest cut with 100% specificity
rNRE	1.89%	0.953%	>3.81%	>3.32%	>3.03%	>3.89%
TD > cut = False Positives			1	2	2 - 4	0
FASD >cut =True Positives			11	13	14-16	9
Sensitivity			41%	48%	52%-59%	33%
Specificity			96%	92%	85% -92%	100%

[a] 3.03% is most accurate cut-point and falls at the 90th percentile of the TD group (75% accurate; AUC 0.77; 95% CI 0.63 to 0.87; p<0.0001). More than 50% of FASD group are above this cut-off. 95% CI for the 90th percentile of the TD group from 2.83 to 3.21.

Prediction Three: In groups with increasing risk of structurally identifiable CNS impairment as ranked using criteria published in the 4-Digit Code *(Astley, 2004), as CNS risk increases from No Risk (RANK 1) to Confirmed Impairment (RANK 4) the proportion of children with rNRE above the range expected in TD children will also increase. This prediction should hold for any clinically meaningful rNRE cut-off (e.g., above 90th percentile, +1.5 or +2.0 standard deviations from the TD mean).*

To test prediction three, all 53 cases were divided into groups based on CNS RANK with TD children assigned a default CNS RANK of 1 (no risk).

Results for Prediction Three. Prediction 3 is largely confirmed. Table 2.4 shows the proportion of each group (%) that has an rNRE falling above each of the three tested cut-points,

while Figure 1.4 shows graphically the proportion of each group above and below the strictest cut-off set at +2 standard deviations of the pooled TD group mean.

Table 2.4: Percentage (and number) of cases with rNRE falling above 3 select cut-points (Preliminary Study 4).

	+2.0 SD cut-point	+1.5 SD cut-point	90th percentile cut-point
CNS RANK 1, n= 26	4% (1)	7% (2)	7% (2)
CNS RANK 2, n= 17	29% (5)	41% (7)	47% (8)
CNS RANK 3, n=4	50% (2)	75% (3)	75% (3)
CNS RANK 4, n= 6	67% (4)	67% (4)	83% (5)

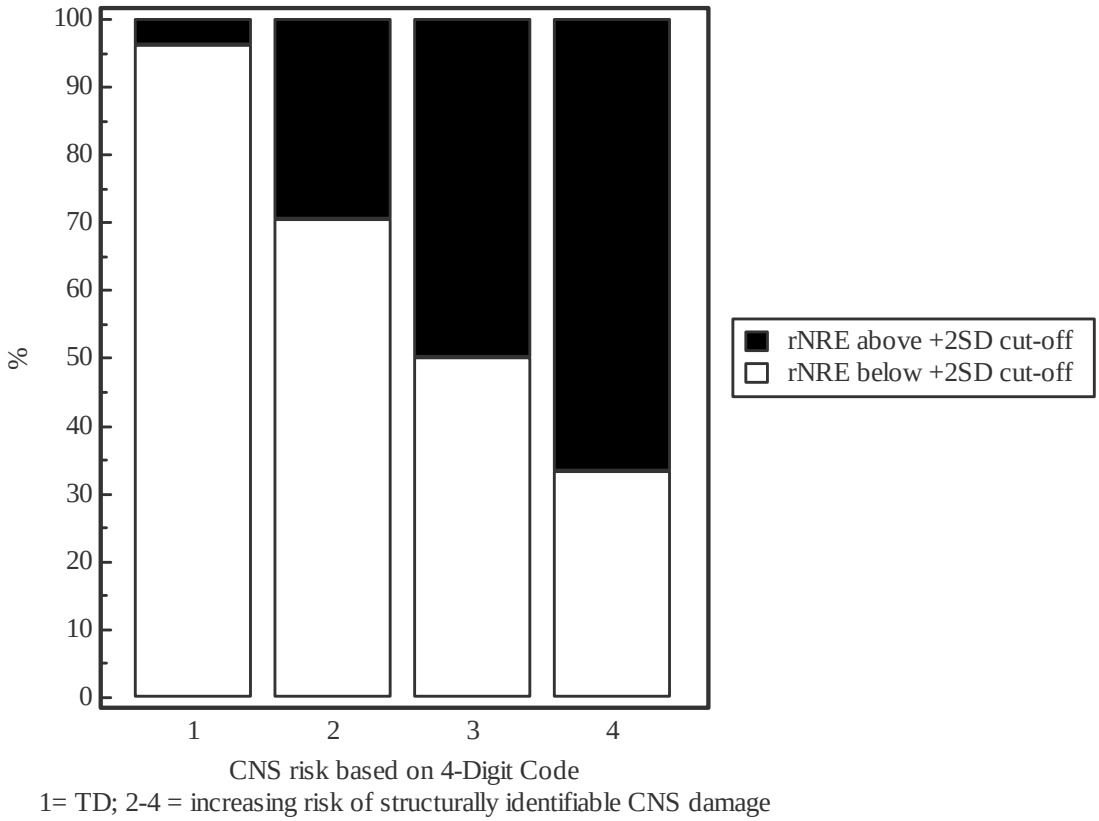


Figure 2.2: Cumulative percentage above and below a +2SD rNRE cut-off for each CNS RANK with the white portion of the bar indicating percentage of children below cut-off and the dark portion of the bar indicating those above that cut-off (each bar adds to 100% of children with that CNS RANK).

Post-hoc Analysis and Discussion. Because children with CNS impairments produced Nominal Reference Errors at rates exceeding those in children considered to have typical development, two of our predicted outcomes held up robustly as we increased the pool of narratives analyzed: (a) more than 50% of children from the pooled FASD group had rNRE above a clinically relevant cut-point based on the 90th percentile rNRE of the pooled TD group; and (b) when our participants were ranked according to increasing risk of structurally identifiable CNS impairment [1], as CNS risk increased from *No Risk* to *Confirmed Impairment* the proportion of children with rNRE +2SD above the mean of the pooled TD children also increased.

At the same time, the range of rNRE values measured in narratives produced by typically developing children increased slightly beyond our predicted maximum of 3% rNRE. This finding suggests that our original sample may have contained a disproportionate number of higher performing youngsters. Nevertheless, the fact that the maximum error rate seen in the original group fell at the 90th percentile for the pooled TD group leads to a prediction that the true population range for typically developing children will not extend significantly into the range seen in our FASD group. For example, if we choose to use the maximum rNRE of 3.03% from our original study as a cut-off value for predicting children at risk for CNS impairment, we would get only two false positives from the current group of 53 children.

Of course, the analysis above uses an "as treated" approach to defining the TD group. In other words, children were considered typically developing for our analysis if they were assigned to the TD group in the source studies. However, children from our newly added TD group have available standardized testing data which allow us to examine their developmental status (see Table A.1 in Appendix). When this is done, four of the newly added TD group would also receive a "false positive" result if we were to use their available standardized testing to classify them into either TD or FASD groups.

For this reason, we conducted post-hoc analysis of our data based on group assignments that took available standardized testing into consideration. For the post-hoc analysis, we used the *4-Digit Code* criteria for CNS RANK to reassign those children from the TD group to a CNS RANK of 2 as appropriate. In the *4-digit Code*, a CNS RANK of 2 can be assigned to any child that exhibits mild (between 1 and 2 SD from normative mean) to severe impairments (greater than 2 SD from the normative mean) in one or more areas of functioning while not meeting criteria for a CNS RANK of 3, which requires severe impairment in at least three domains of functioning [1]. This lead to 4 children being moved from the TD group (i.e., CNS RANK 1 group) into the CNS RANK 2, indicating they were "at-risk for CNS impairment" according to the *4-Digit Code* criteria. Group rNRE mean and standard deviation were then recalculated with this smaller "CNS RANK 1"group. As seen in Table 2.5, when the revised grouping is used, rNRE remains a highly accurate classifier of children "at risk for CNS impairment" (AUC = 0.84; 95% CI 0.71 to 0.93; $p<0.0001$) when "at-risk" is defined according to a CNS RANK 2 or higher using the *4-Digit Code*. Importantly, for this group of 53 children, 58% of children with a CNS RANK of 2 or higher (n= 18) have an rNRE outside the range seen in the 22 children with a CNS RANK of 1 and more than 2/3^{rds} fall beyond both the 90th percentile and the +1.5 standard deviation cut-off scores.

Table 2.5: Mean, Standard Deviation (SD) and classification measures at 3 cut-points based on rNRE of group with NO EVIDENCE OF CNS RISK (*4-Digit* CNS RANK 1, n = 22; Preliminary Study 4).

	RANK 1 Mean	RANK 1 SD	+2 SD cut	+1.5 SD cut	90th percentile[a] cut	100% specificity cut
rNRE	1.63%	0.75%	>3.13%	>2.76%	>2.60%	>3.03%
CNS RANK 1> cut = False Positives			0	1	2 (1 - 4)	0
"at–risk" > cut = True Positives			16	21	21-23	18
Sensitivity			52%	68%	68% -74%	58%
Specificity			100%	95%	91% (85 -95%)	100%

[a]2.60% is most accurate cut-point and falls at the 90th percentile of the CNS RANK 1 group (81% accurate; AUC = 0.84; 95% CI 0.71 to 0.93; $p<0.0001$). 95% CI for the 90th percentile of the CNS RANK 1 group from 2.44 to 2.76.

Conclusions. Results from Study 4 provide evidence of the strong potential that rNRE from a TREIN analysis has as a clinical tool for use in indentifying underlying CNS impairments in children suspected of having FASD. Our results confirm that rNRE is a measure that can be reliably obtained with a reasonable amount of training. We were also able to demonstrate that, for this sample of 53 children: (a) rNRE significantly above 3% was rare in the typically developing children and when present was associated with deficits on other clinical language measures; (b) that a significant proportion of children with clinically identified CNS impairments produced rNRE outside the range seen in the typically developing group; and (c) that increased risk of structurally identifiable CNS impairment resulted in increased chances of having rNRE outside the range seen in children free from evidence of CNS impairment.

Summary Of Results From All 4 Preliminary Studies:

The accumulated evidence gathered in our preliminary research points to a strong potential for the TREIN to have both clinical and research utility as a tool for measuring Integrative Language abilities in school-aged children with FASD. The system was shown to have strong inter-rater reliability (kappa = 0.90—see details in [92]and kappa = 0.89—details in description of Study 4 above) with minimal training consisting of less than 10 hours of instruction and practice. One outcome measure from the system, rNRE, was shown to have strong potential to discriminate between typically developing and impaired populations with increasing accuracy as severity of underlying impairment increased suggesting an association between underlying CNS impairments and difficulty with the Integrative Language capacities needed to maintain a low rNRE.

However, even with 53 narratives analyzed, there were significant questions left unanswered. The first comes from the fact that the children in the FASD groups examine so far were chosen for previous research due to their significant behavioral and social deficits. While

these impairments are common in children with FASD, many children with FASD do not have these deficits. In addition, the groups of FASD children examine so far were chosen strategically to include a heavier representation of children with full FAS than would be expected in a clinical population of children with FASD. These facts speak to the need to confirm these results in a larger clinical population of children who are more representative of the true range of clinical outcomes found across all FASD.

Perhaps more importantly, we have hypothesized that the skills measured with the rNRE should be relatively stable by elementary school age. However, while the age range of the children examined so far included a few children in the later half of their 7th year, the majority of narratives came from children in the 9 and up age range. This speaks to the need to examine a much larger set of narratives produced by typically developing children in order to provide a more stable estimate of the true range of variability of performance found in children during the elementary school years. The research presented below was designed to meet these challenges using a much larger, more representative sample.

CHAPTER THREE:

THE CURRENT RESEARCH

Introduction:

The primary aim of the current research was to provide evidence of the degree to which integrative language impairment identified using a TREIN analysis can serve as a behavioral marker of underlying CNS impairments. Our focus was on demonstrating the validity and clinical utility of TREIN measures that quantify errors of nominal reference (e.g., the rate of Nominal Reference Errors, rNRE) by demonstrating an association between deficits in this skill and underlying CNS impairment identified in previous clinical evaluations.

In this dissertation I have proposed three validity criteria for the TREIN analysis [adapted from 71]. If outcome measures from a TREIN analysis have potential as behavioral markers of underlying CNS impairments associated with prenatal alcohol exposure (PAE), they should:

(i) be associated with disorder (i.e., with clinically diagnosed impairment),

(ii) be found in individuals with PAE that do not have full FAS at a higher rate than in the general population,

(iii) be state-independent (i.e. are found whether or not the individual's current performance would lead to a diagnosis of a more general "Language/Communication Impairment").

While each of the validity criteria are ultimately important to demonstrate, the current dissertation will place emphasis on demonstrating the association to impairment needed for rNRE to meet criterion (i), while exploring potential for meeting criteria (ii) & (iii). Criterion (i) serves as the minimal bar for a valid measure of impairment that would indicate utility for a number of purposes, so emphasis is placed on demonstrating that it is met. Criteria (ii) and (iii) are required for validation of measures that would be used in the endophenotyping approach discussed above. The research presented below will provide evidence that TREIN measures have potential to meet these criteria, but further research will be needed to firmly establish that these two criteria can be

met. In addition to exploring these questions of validity, the relative clinical utility of the various TREIN measures will also be explored.

The degree of association between Integrative Language impairment as measured by the TREIN and underlying CNS impairment was examined in a retrospective analysis of a large data set that included children with CNS impairment identified during a diagnostic evaluation for suspected Fetal Alcohol Spectrum Disorders (an "FASD" group) and typically developing peers (a "TD" group). We conducted a serious of tests of the strength of this association. For a TREIN outcome measure to be considered a potential candidate for use in the endophenotyping approach advocated above, there needed to be not only a difference in performance on average between the TD and FASD groups, but that difference needed to result in substantial numbers of children in the FASD group performing outside the range seen in the TD group. In addition, elevated numbers of errors found in a TREIN analysis should be correlated with other indicators of underlying CNS abnormality (including severity of disorder and severity of other risk factors).

In addition to establishing the potential validity of the TREIN, another important aim of this dissertation is to better understand the relative clinical utility of each of the various TREIN outcome measures for use with elementary school aged children. This information will help to inform refinements of the tool as it is moved from a purely research tool to a viable clinical instrument. For instance, while it is clear that some control for length is needed in the outcome measures for the TREIN, it is unclear whether controls based on total story length are empirically superior to those based on referential opportunities. Likewise, while the studies described below were designed based on an assumption that measures incorporating nominal reference errors will have greater clinical utility in this age-range than those based on pronominal reference errors, there was a need to directly test this assumption. A study designed to test specific hypotheses related to these questions was included in the research described below.

Research Design And Methods:

We conducted a series of five studies to demonstrate the degree to which TREIN measures could meet each of the validity criteria outlined above and to explore the relative clinical utility of the various TREIN measures. All studies were carried out through retrospective analysis of existing clinical and research data available from either the University of Washington Fetal Alcohol Syndrome Diagnostic and Prevention Network (FASPDN) or the University of Washington Child Language Laboratory (CLL). As all studies shared participants, we first provide a detailed description of the study sample. This includes descriptive statistics related to the participants themselves, followed by a report summarizing the narratives they produced. This chapter concludes with a detailed description of each of our specific aims including the hypotheses tested, study design, methods, and results relevant to each aim. There were four specific aims for the dissertation, three related to the validity criteria described above, and one related to the clinical utility of the TREIN for use with elementary school-aged children.

Study Populations:

Participants included 155 children. A total of 53 participants included in the current study population were participants in the preliminary studies described above. Participants constituted two groups: a group of children following a typical developmental course, the "TD" group; and a clinical sample of children with CNS impairment previously diagnosed during assessments for suspected FASD, the "FASD" group. Details related to each group are presented below.

TD Group (n = 80):

Inclusion Criteria:

1. Typically developing based on review of school records
2. Participated in a *Frog Where Are You* narrative assessment at the University of Washington

3. Age at time of narrative assessment: 6 thru 14 years of age
4. Transcript or audio recording of narrative assessment available
5. All races, English as primary home language
6. Normal hearing

Exclusion Criteria: None

Source and representativeness of TD subjects:

The TD group was recruited for previous research [115] from elementary schools representing two school districts in the greater metropolitan Seattle area. These districts were chosen because median family incomes and socio-demographic characteristics were similar across school districts and are representative of the area. No intelligence or standardized language measures are available for TD participants. However, school psychologists familiar with the children and FASD screened school records for each child with respect to school performance, social ability, and general behavior. Based on this review of available records, each was judged to be following a typical developmental course due to their unremarkable behavior and adequate yet unexceptional school achievement. Age at time of narrative, gender, and school location are the only demographic data available for the TD group. Prenatal alcohol exposure was not directly assessed in the TD group, but the screening process used was designed to identify a group of children considered at low-risk of having had significant PAE.

<u>FASD Group (n = 75):</u>

Inclusion Criteria:

1. Received one of the following diagnoses from the University of Washington Fetal Alcohol Syndrome Diagnosis and Prevention Network (FASDPN) clinic:
    a. Full or partial FAS (i.e., severe CNS impairment, FAS facial features)
    b. Static Encephalopathy (i.e., severe CNS impairment, no FAS facial features)
    c. Neurobehavioral Disorder (i.e., mild CNS impairment)

2. Participated in a "*Frog Where Are You*" narrative assessment during the FASD exam
3. Age at time of narrative assessment: 6 to 14 years of age.
4. Transcript or audio recording of narrative assessment available
5. All races, English as primary home language
6. Normal hearing

Exclusion Criteria: None

Source and representativeness of FASD subjects:

The FASD group is a subset of over 2,000 patients evaluated for suspected FASD by an interdisciplinary team using the FASD *4-Digit Code* [1] at the University of Washington Fetal Alcohol Syndrome Diagnostic & Prevention Network (FASDPN). This unique dataset, the world's largest FASD clinical data base, provides a wealth of relevant data. The 75 narratives available from this dataset were collected as part of clinical or research assessments. These 75 subjects are typical of the population of children between 6-14 years of age seen in the FASDPN clinic in terms of FASD diagnosis, race, gender, and socio-economic status.

All but 4 cases had confirmed prenatal alcohol exposure (PAE). These 4 children had suspected, but not confirmed, prenatal alcohol exposure. One had the full face of FAS (*4-Digit Code* 3422). All had a diagnosis of "neurobehavioral disorder", indicating at least mild CNS impairment. Although 3 of these 4 children have PAE status that is highly questionable, all have identified CNS impairments similar to those found in the FASD cases in the sample. As the primary purpose of the current research was to establish a relationship between underlying CNS impairment and narrative performance, and not to directly assess the relationship between prenatal alcohol exposure and narrative performance, these cases were included in our analyses. This group of 75 children will be referred to below as the "FASD group," however, the reader should keep in mind that 4 of these 75 children do not meet clinical criteria for FASD as PAE is only suspected but not confirmed.

Descriptive Statistics – Participants:

Age at time of narrative & gender. A Welch-test (used due to unequal variance between the groups) revealed no significant difference in the age distribution across groups (see Table 3.1). Similarly, while the proportion of females in the TD group was larger than that in the FASD group, Fisher's exact test of proportions indicated no significant difference ($p<0.05$) in the gender distribution between groups (see Table 3.2). Other demographic information was not available for the TD group.

Table 3.1: Comparison (t-test) of age distribution between TD and FASD groups.

AGE distribution of groups	TD group	FASD* group
Sample size	80	75
Arithmetic mean (years)	10.21	9.81
95% CI for the mean	9.70 to 10.73	9.45 to 10.17
Variance	5.3287	2.3901
Standard deviation	2.3084	1.5460
Standard error of the mean	0.2581	0.1785
RANGE (years)	7.2 to 14.4	6.3 to 12.8
F-test for equal variances	p = 0.001: Reject equal variance	
Welch-test (assuming unequal variances)		
Difference		-0.4020
Standard Error		0.3138
95% CI of difference		-1.0225 to 0.2184
Test statistic t(d)		-1.281
Degrees of Freedom (DF)		138.8
Two-tailed probability		p = 0.2023

*N= 75: 71 with a FASD and 4 with neurobehavioral disorder, alcohol unknown.

Table 3.2: Comparison of gender distribution between TD and FASD groups.

Gender of groups	Gender		
Group	Male	Female	
TD	34	46	80 (52.3%)
FASD*	36	37	73 (47.7%)
	70 (45.8%)	83 (54.2%)	153
Fisher's exact test: p = 0.420615607			

*N= 73: 69 with a FASD and 4 with neurobehavioral disorder, alcohol unknown. Gender information was not available for two children in this group.

Available Data:

FASD Diagnostic Data: All patients in the FASD group received a comprehensive FASD diagnostic evaluation at the FASDPN clinic. As a standard of clinical care, a comprehensive set of clinical measures are obtained on all patients (over 500 outcome measures) including height, weight, head circumference, computerized measurement of FAS facial anomalies from digital photos, and a battery of standardized neuropsychological assessments administered by licensed professionals in the area of Speech-Language Pathology, Psychology, and Occupational Therapy. This dataset is maintained by Dr. Susan Astley with Institutional Review Board approval and patient consent (renewed annually since 1993).

The following data were obtained from the FASDPN database:

1) Gender, race, & age at time of diagnosis.
2) FASD Diagnosis and *4-Digit Diagnostic Code* (See Table 3.3 below).
3) CNS RANK, a 4-point ordinal rank of severity of CNS abnormality [1].
4) Occipital-Frontal Circumference (OFC) including the age-normed percentile.
5) 4-Point Severity RANK of Structural, Neurological, and Functional FASD CNS measures including a Language Impairment RANK.
6) Standardized measures of intellect, achievement, language, cognition, sensory motor integration, etc.
7) Prenatal Alcohol Exposure: Average number of drinks per occasion, maximum number of drinks per occasion, average number of drinking days per week, trimester(s) in which alcohol was consumed.
8) 4-point Likert Rank for other prenatal and postnatal adverse exposures/events.
9) Other issues that could explain CNS abnormalities (e.g., head injury).
10) Presence of a history of ear infections.

Table 3.3: 4-Digit Codes for *FASD group (n=75) organized by degree of CNS impairment.

Static Encephalopathy		Neurobehavioral Disorder		
2444 -FAS	4234	1124	1224	1123
2443 -FAS	4234	1223	1124	1224
3442 -FAS	4234	3223	1124	1224
1443 -FAS	3233	3123	1124	1224
4343 -FAS	1234	1424	1124	1223
4344 -FAS	1234	1424	1124	1223
1343 -FAS	1234	1423	1124	1223
3244	2234	1423	1124	1223
3243	1233	1423	1123	1223
1244	1234	1324	2123	1223
1343	1234	1324	1424	1223
1243	1233	1323	1224	**3422***
4244	1233	2223	1224	**1222***
3244	2134	2124	1224	**4222***
2244	1134	1224	1123	**1122***

FAS = full or partial FAS according to Astley, 2004 [1] following Astley et. al [59].
***N= 75: 71 with a FASD and 4 with neurobehavioral disorder, alcohol unknown.**

Narrative Data: Narratives were collected as part of clinical assessments at the FASDPN or as part of research at the University of Washington. Each FASD and TD subject had an audio recording of their *Frog Where Are You* narrative collected as described above (see preliminary studies above). After the child previewed the storybook, the examiner asked the participant to tell the best story possible while using the picture book as a visual prompt. Examiners were always seated across the room from the child to make it clear that they were unable to see the storybook pictures. Narratives were recorded on audiocassette and orthographically transcribed by trained graduate students.

Although approximately a third of the narratives included in our sample had been previously examined (as part of the preliminary studies described above), new transcripts for all narratives were prepared for the current study from the source audio using a more systematic transcription process designed to ensure the fidelity and uniformity of the transcripts used in coding. Transcripts for analysis were prepared in a multi-staged process. First, a team of two trained graduate students independently transcribed all narratives from audio sources according to

conventions from *Systematic Analysis of Language Transcripts* (SALT; [111]). Then, a team including these transcribers and two additional trained research assistants identified differences between these two sets of independent transcripts and came to agreement as to how these differences should be resolved based on additional review of the audio tapes. If the team could not resolve a conflict, the transcripts were presented to Dr. Truman Coggins, who reviewed both transcripts with the team to come to a resolution. All analysis was conducted on these consensus transcripts. A trained Speech-Language Pathologist (JCT) analyzed narratives from transcripts utilizing the TREIN protocol [110].

TREIN outcome measures: All analytical coding of narrative transcripts was conducted blind to any characteristics of the story teller including participant diagnosis/group membership, age, and gender. Prior to the TREIN analysis the length of the narrative was calculated with SALT. This provides the Number of Total Words (NTW) in the analysis set ("NTW.a" from a standard SALT analysis).

In the TREIN system, only error rate measures which adjust for length would be considered logical candidates for measuring integrative language performance. This can be done by controlling for either total words (NTW) or reference opportunities (opp). The primary outcome measures generated by the TREIN analysis were:

1. Total Nominal Reference Errors (NRE) = [ambigintro] + [ambigntie]
2. Rate of Nominal Reference Errors (rNRE) = NRE / NTW
3. Rate of Nominal Reference Errors by opportunity (rNREopp) = NRE/([ambigintro] + [ambigntie] + [defintro] + [indefintro] + [possintro] + [pntie] + [ntie] + [ambigpntie] + [pnintro])
4. Pronominal Reference Errors (PRE) = ([pnintro] + [ambigpntie])
5. Rate of Pronominal Reference Errors rPRE = PRE/ total words (NTW.a)

6. Rate of Pronominal Reference Errors by opportunity (rPREopp) = PRE/ ([ambigintro] + [ambigntie] + [defintro] + [indefintro] + [possintro] + [pntie] + [ntie] + [ambigpntie] + [pnintro])
7. Total Reference Errors (ALL) = [ambigintro] + [ambigntie] + [ambigpntie] + [pnintro]
8. Rate of All Reference Errors (rALL) = (NRE + PRE)/ NTW
9. Rate of All Reference Errors by Opportunity (rALLopp) = (NRE + PRE)/ ([ambigintro] + [ambigntie] + [defintro] + [indefintro] + [possintro] + [pntie] + [ntie] + [ambigpntie] + [pnintro])

Each of the rates defined above are converted to a percentage for reporting.

Because new transcripts were used for the 53 narratives previously analyzed, no direct intra-rater comparison between current and previous coding was possible. However, a comparison of approximately 50% of the 53 transcripts that had previously been analyzed was conducted after coding had been finalized. This examination revealed only minor differences between the older and newer transcripts. These differences primarily manifest as reductions in the number of unintelligible words. Because there is a subtle auditory distinction between “a” and “the”, particular attention was paid to the impact that the new transcriptions had on these forms, which are central to the identification of Nominal Reference Errors. In all cases where a change in these forms was identified, the change went from “the” to “a” or from “unintelligible” to one or the other form. These changes are most likely to reduce the number of Nominal Reference Errors. Indeed, in all cases where a change in the total number of Nominal Reference Errors in a narrative was identified, the change resulted in a reduction in errors.

Descriptive Statistics –Narrative Data:

Table 3.4: Group performance on each TREIN outcome measure.

TREIN measures based on NTW.a

	rNRE= NRE/NTW		rPRE = PRE/NTW		rAll = All/NTW	
Group	TD	*FASD	TD	*FASD	TD	*FASD
N	80	75	80	75	80	75
Mean	1.52%	2.84%	1.22%	1.81%	2.74%	4.65%
Variance	0.7477	2.6396	1.7200	3.7520	3.1109	9.3743
SD	0.8647	1.6247	1.3115	1.9370	1.7638	3.0617
SEM	0.09667	0.1876	0.1466	0.2237	0.1972	0.3535
Minimum	0.0%	0.0%	0.0%	0.0%	0.0%	1.25%
Maximum	3.89%	8.89%	7.14%	11.46%	8.04%	17.71%
Reject Normality	0.0359	0.0005	<0.0001	<0.0001	0.0078	<0.0001

TREIN measures based on number of reference opportunities (opp)

	rNREopp = NRE/opp		rPREopp =PRE/opp		rAllopp = All/Opp	
Group	TD	*FASD	TD	*FASD	TD	*FASD
N	80	75	80	75	80	75
Mean	5.57%	10.28%	4.41%	6.42%	9.98%	16.70%
Variance	10.3247	33.9655	21.7307	44.7459	40.3855	113.9969
SD	3.2132	5.8280	4.6616	6.6892	6.3550	10.6769
SEM	0.3592	0.6730	0.5212	0.7724	0.7105	1.2329
Minimum	0.0%	0.0%	0.0%	0.0%	0.0%	4.48%
Maximum	14.58%	30.77%	23.88%	37.93%	26.87%	58.62%
Reject Normality	0.0100	0.0024	<0.0001	<0.0001	0.0090	<0.0001

*N= 75: 71 with FASD and 4 with neurobehavioral disorder, alcohol unknown.

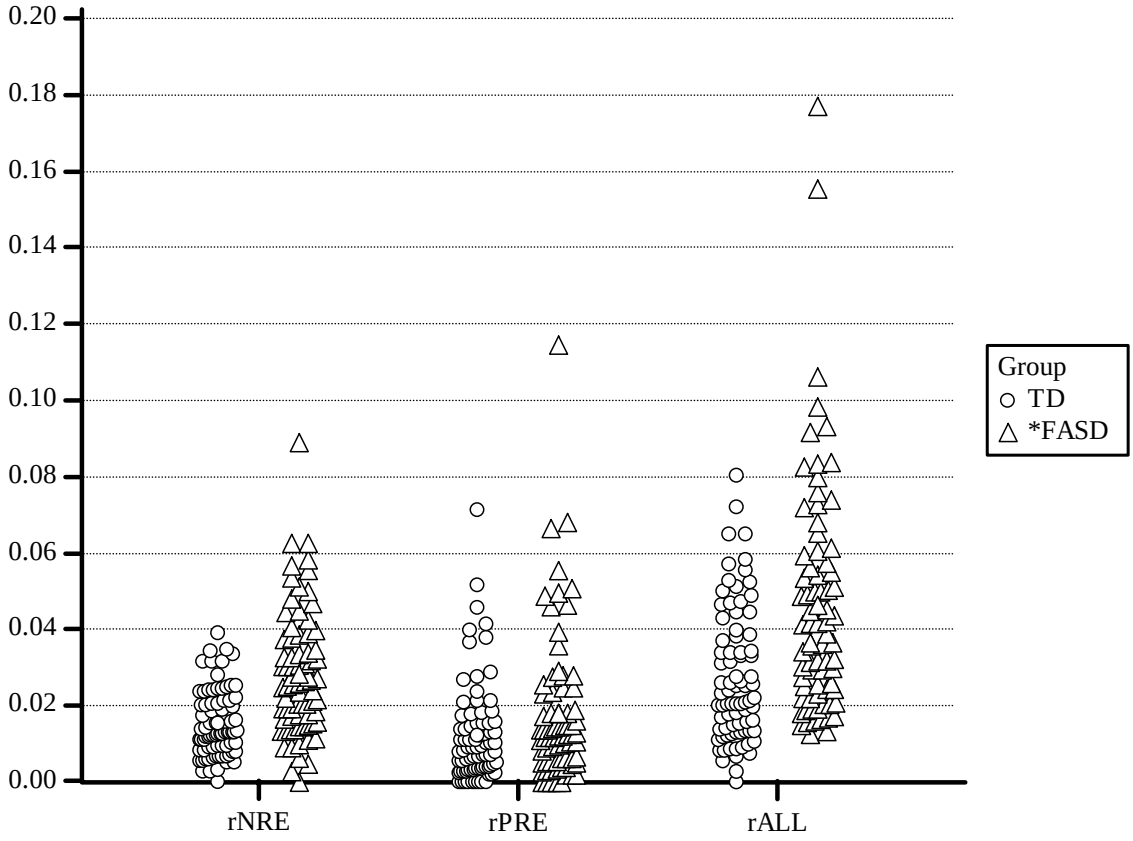


Figure 3.1: TREIN measures based on number of total words for both TD and *FASD groups. *N= 75: 71 with a FASD and 4 with neurobehavioral disorder, alcohol unknown.

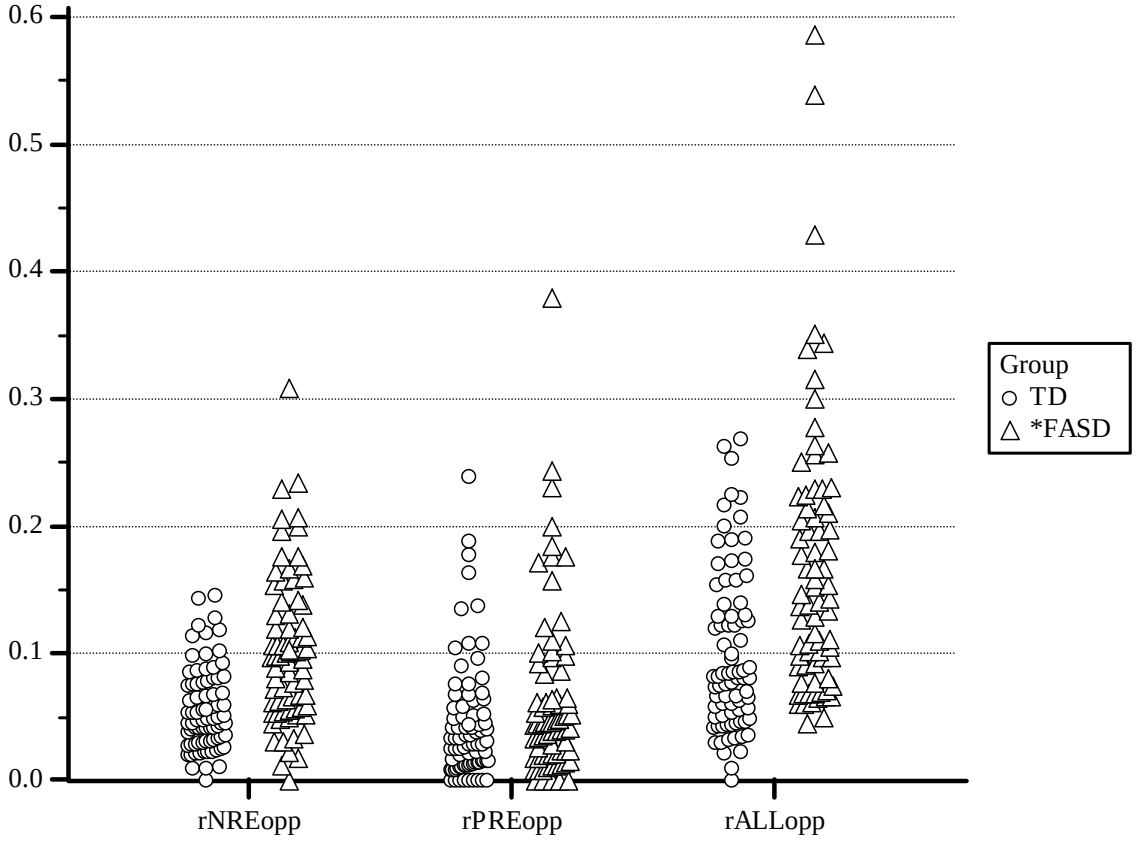


Figure 3.2: TREIN measures based on number of opportunities.
***N= 75: 71 with a FASD and 4 with neurobehavioral disorder, alcohol unknown.**

Age and gender are potentially important factors in narrative performance so the impact of each factor was examined. When performance was compared between genders (Welsch test for unequal variance; two-tailed $p<0.05$) no significant differences were found for any TREIN measure.

Age, was significantly correlated ($p<0.05$) with performance for all TREIN measures in the TD group, but not in the FASD group (see Table 3.5). Based on visual inspection of the distribution of performance across the TD group, this correlation is driven largely by a decrease in the variation of scores as age increases; with younger children in the TD group having a wider range of variation than older children in the TD group (e.g., see distribution of rNRE by age in

Figure 3.3). In contrast, the range of variation remained large as age increased in the FASD group.

Table 3.5: Correlation of TREIN measures to age at time of narrative.

GROUP	**TD**	***FASD**
Sample Size	80	75
rate of nominal reference errors (rNRE) to age		
Correlation coefficient r	-0.4462	-0.0795
Significance level	**p<0.0001**	**p=0.4977**
95% Confidence interval for r	-0.6065 to -0.2511	-0.3010 to 0.1502
rate of Nominal Reference Errors by opportunity (rNREopp) to age		
Correlation coefficient r	-0.4311	-0.1076
Significance level	**p=0.0001**	**p=0.3582**
95% Confidence interval for r	-0.5945 to -0.2335	-0.3266 to 0.1224
rate of Pronominal Reference Errors (rPRE) to age		
Correlation coefficient r	-0.2372	-0.1808
Significance level	**p=0.0341**	**p=0.1206**
95% Confidence interval for r	-0.4343 to -0.01845	-0.3917 to 0.04816
rate of Pronominal Reference Errors by opportunity (rPREopp) to age		
Correlation coefficient r	-0.2295	-0.1846
Significance level	**p=0.0406**	**p=0.1129**
95% Confidence interval for r	-0.4277 to -0.01030	-0.3950 to 0.04424
rate of All Reference Errors (rALL) to age		
Correlation coefficient r	-0.3951	-0.1566
Significance level	**p=0.0003**	**p=0.1798**
95% Confidence interval for r	-0.5657 to -0.1921	-0.3704 to 0.07300
rate of All Reference Errors by opportunity (rALLopp) to age		
Correlation coefficient r	-0.3863	-0.1744
Significance level	**p=0.0004**	**p=0.1346**
95% Confidence interval for r	-0.5586 to -0.1821	-0.3861 to 0.0548

*N= 75: 71 with a FASD and 4 with neurobehavioral disorder, alcohol unknown.

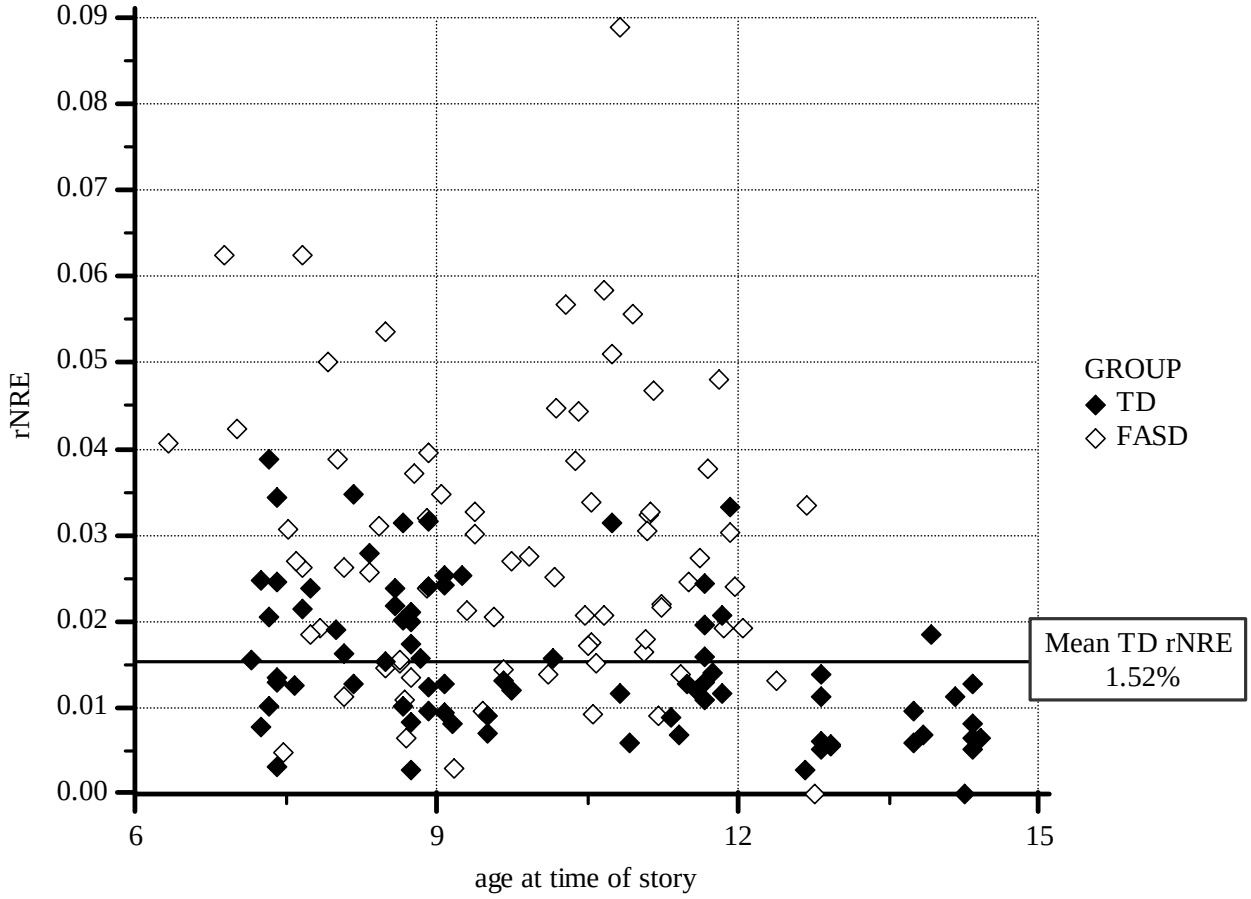


Figure 3.3: Distribution of rNRE by age with TD and *FASD groups indicated.
***N= 75: 71 with a FASD and 4 with neurobehavioral disorder, alcohol unknown.**

Because of the unexpected correlation between age and performance in the TD group, there was a need to further explore the impact of age on TREIN performance. Post-hoc analyses of this impact are included in the studies below as appropriate.

Specific Aims, Hypotheses, And Methods:

The first three specific aims of this dissertation were evaluated in a series of four studies addressing specific hypotheses related to each of the validity criteria proposed at the outset of Chapter 3. A fourth specific aim, related to clinical utility was evaluated in a fifth study. In the sections below, each specific aim is presented, followed by the hypotheses related to that aim.

Study design, Analysis, and Results for each hypothesis are also described. Post-hoc analyses related to the impact of age on results related to each specific aim are included as appropriate.

SPECIFIC AIM ONE:

Specific Aim One: To assess the degree of to which the rate of Nominal Reference Errors (rNRE) can meet validity criteria (i) described above by showing that rNRE is "associated with disorder (i.e., with clinically diagnosed impairment)."

If TREIN outcome measures are valid measures of Integrative Language impairment that reflect underlying CNS impairment, one would expect the TD group to make fewer errors on average as measured by a TREIN analysis when compared to the FASD group who have previously identify CNS impairments. In addition, one would expect a substantial number of children in the FASD group to have clinically significantly elevated rNRE. While any decision as to what constitutes "a substantial number" will be arbitrary to some degree, it was felt that finding a simple majority of children with elevated rNRE would provide strong support for the idea that the association between elevated rNRE and previously diagnosed impairment is not only real, but important. In addition, if rNRE is associated with underlying CNS abnormalities, one would also expect 1) that the more elevated the rNRE, the more severe the global risk of underlying CNS impairment; and 2) that more severe integrative language impairment will be associated with more severe presentation of other proxy measures of CNS impairment (both functional and structural). Meeting each of these expectations will strengthen the claim that rNRE is associated with disorder and can meet validity criterion (i) described above. Two studies testing 6 hypotheses were conducted.

Study One:

Hypotheses:

1. The mean rate of errors for all TREIN measures in the *Frog Where Are You* narratives will be greater for the FASD group than for TD group.
2. The proportion of children in the FASD group who generate a narrative with an rNRE in the impaired range will be 50% or greater with impairment defined by the following performance criteria: rNRE greater than 2.0 standard deviations above the mean of the TD group.

Design:

For our first study, group-level performance was compared between the two study groups: 1) children with impairment identified in an evaluation of suspected FASD (FASD; n = 75), and 2) typically developing peers (TD; n= 80).

Analyses:

Hypothesis 1: For each TREIN measure, a t-test for independent groups (TD versus FASD) assuming unequal variance was used (Welch-test with a 2-tailed alpha). A Bonferroni correction for multiple comparisons was used and alpha was set at 0.007.

Hypothesis 2: A one-sample test of proportions (2-tailed alpha = 0.05) was used comparing the proportion in the defined impairment range to a prediction of 50%.

Results for Study One:

Hypothesis 1: The mean performance for each TREIN measure was significantly different between the TD and FASD groups for all measures including nominal reference errors (rNRE, rNREopp, rALL, rALLopp) with FASD group means almost twice that of the TD group. As can be seen in Table 3.6, those measures based on pronominal errors were not significantly different at this alpha.

Table 3.6: Comparison of TD and FASD group performance on all TREIN outcome measures.

Independent samples t-test (Welch-test assuming unequal variances)		
Group	**TD**	***FASD**
Sample size	80	75
		rate of Nominal Reference Errors (rNRE)
Arithmetic mean	1.521%	2.843%
Difference		1.322
Standard Error		0.2110
Test statistic t(d)		6.263
Degrees of Freedom (DF)		111.2
Two-tailed probability		**p < 0.0001
		rate of Nominal Reference Errors by opportunity (rNREopp)
Arithmetic mean	5.569%	10.28%
Difference		4.710
Standard Error		0.7628
Test statistic t(d)		6.174
Degrees of Freedom (DF)		113.5
Two-tailed probability		**p< 0.0001
		rate of Pronominal Reference Errors (rPRE)
Arithmetic mean	1.222%	1.805%
Difference		0.5828
Standard Error		0.2674
Test statistic t(d)		2.179
Degrees of Freedom (DF)		129.0
Two-tailed probability		p= 0.0311: NS
		rate of Pronominal Reference Errors by opportunity (rPREopp)
Arithmetic mean	4.408%	6.423%
Difference		2.015
Standard Error		0.9318
Test statistic t(d)		2.162
Degrees of Freedom (DF)		131.2
Two-tailed probability		p= 0.0324: NS
		rate of All Reference Errors (rALL)
Arithmetic mean	2.743%	4.648%
Difference		1.905
Standard Error		0.4048
Test statistic t(d)		4.705
Degrees of Freedom (DF)		116.6
Two-tailed probability		**p< 0.0001
		rate of All Reference Errors by opportunity (rALLopp)
Arithmetic mean	9.977%	16.70%
Difference		6.725
Standard Error		1.423
Test statistic t(d)		4.726
Degrees of Freedom (DF)		119.0
Two-tailed probability		**p< 0.0001

** Significant at p< 0.007: NS = non-significant at p<0.007

*N= 75: 71 with a FASD and 4 with neurobehavioral disorder, alcohol unknown.

Hypothesis 2: The percentage of cases within the FASD group falling in the impaired range was compared to a prediction of 50%. With 25 of the 75 cases (33.33%) falling in the impaired range for rNRE hypothesis 3 is rejected (see Table 3.7).

Table 3.7: Percentage of *FASD cases falling above a +2 SD rNRE cut-off compared to prediction of 50% or greater (test for one proportion).

+2SD cut-off = rNRE>3.2504	
Observed proportion of 75 FASD cases	33.33% (25 out of 75 children)
95% CI of observed proportion	22.86% to 45.16%
z statistic	2.887
Significance level	p = 0.0039: **reject prediction**

*N= 75: 71 with a FASD and 4 with neurobehavioral disorder, alcohol unknown.

Study Two:

Hypotheses:

3. As severity of *4-Digit Code* CNS RANK increases in severity from RANK 1 (TD group) to RANK 2 (*possible* CNS damage) to RANK 3 (*probable* CNS damage) to RANK 4 (*definite* CNS damage) the proportion of children having an elevated rNRE (+2 SD from mean of TD group) will increase.
4. For those children in the FASD group for which both measures are available, rNRE will be inversely correlated with clinical estimates of Full-scale IQ (i.e., higher rNRE will be associated with lower IQ).
5. For those children in the FASD group for which both measures are available, increasing rNRE will be negatively correlated with head size as reflected in the clinically measured Occipital-Frontal Circumference percentile (OFC) for age.
6. In the FASD group, the proportion of children with an elevated rNRE (+ 2 SD of the TD mean) will be greater among those who have a *4-Digit code* FACE RANK of 4 than among those without this FACE RANK.

Design:

Study Two examined the association between the TREIN measure rNRE and CNS impairment taking advantage of the gradation in severity of CNS impairment risk available across our sample. Study Two involves a series of tests of association between rNRE and specific and increasingly narrow clinical measures of risk for CNS abnormality.

Specifically,

- severity of the *4-Digit Code* CNS RANK indicates the global risk of underlying CNS abnormality across domains (including e.g., sensory-motor functioning);
- Full-scale IQ indicates the degree of cognitive impairment (excluding e.g., sensory-motor deficits);
- OFC is a gross measure of brain size that can be impacted by abnormal hypo/hyperplasia in any brain region;
- *4-Digit Code* FACE RANK is used as a proxy measure of frontal lobe damage caused by prenatal alcohol exposure, as mid-line facial features and the frontal lobes share cellular origins during embryonic development.

All 155 participants were included in analysis for hypothesis 3. For hypotheses 4, 5, & 6, group size was determined by available clinical data in the FASD group.

Analysis:

Hypotheses 3: Along with visual inspection of the distribution of impaired versus unimpaired children across categories, a chi-square test for trend (with a two-tailed p-value of 0.05) was used to compare the proportion of children having an elevated rNRE (+2 SD above TD group mean) for each categorical rank. Children from the TD group were given a default CNS RANK of 1 indicating no CNS impairment while FASD cases ranged from CNS RANK 2 to 4.

Hypothesis 4: Pearson's correlation co-efficient (with a two-tailed p-value of 0.05) was run between rNRE and available clinical measures of IQ. For some children in the FASD group, more than one measure of IQ was available. In those cases the highest IQ score was used following the logic that IQ tests are less likely to over-estimate underlying ability than they are to under-estimate that ability. As a variety of measures were used to estimate IQ across subjects, standardized scores were converted to z-scores based on the normative mean and standard deviation of the available measure for each child. Correlations were run using these z-scores.

Hypotheses 5: Head size is a potential proxy of CNS impairment with microcephaly being the most common structural indicator of CNS impairment found in our clinical sample. The relationship between rNRE and deviation of head-size was determined using Kendall's Tau (with a two-tailed p-value of 0.05).

Hypothesis 6: The children with available 4-Digit FACE RANKS were dichotomized into two groups, those with a FACE RANK of 4 and those without. Chi-square test was used to compare the proportion of children in each group that met criteria for impairment based on rNRE> +2 SD above the TD group mean.

Results for Study Two:

Hypothesis 3: Based on a Chi-square test for trend, the proportion of children meeting criteria for impairment based on their rNRE during the narrative was significantly different between categories defined by *4-Digit Code* CNS RANK. There was a linear trend indicating increasing proportions of impairment in groups moving from CNS RANK 1 (TD group) to CNS RANK 2 (*possible* CNS damage) to CNS RANK 3 (*probable* CNS damage) to CNS RANK 4 (*definite* CNS damage; see Table 3.8 & Figure 3.4). Based on visual inspection of Figure 3.4 this trend

appears to be a 3-step rather than 4-step trend with proportions of children with impairment in the middle CNS ranks being similar.

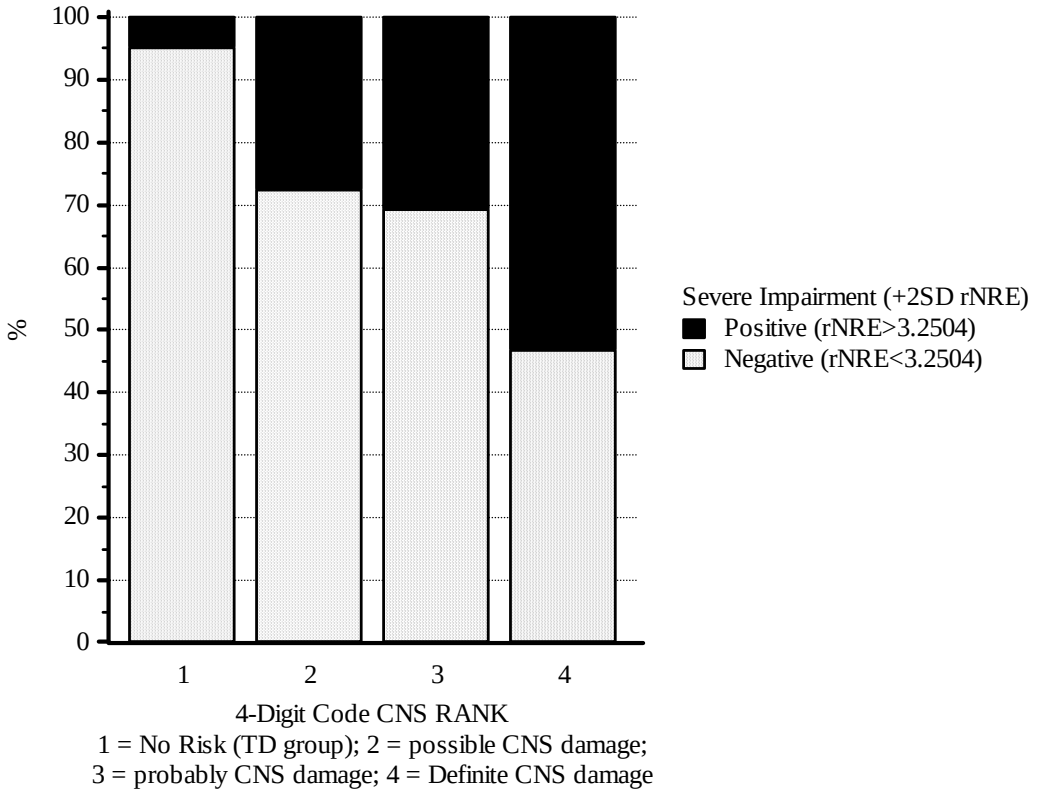


Figure 3.4: Proportion of children above and below cut-off of +2SD in *4-Digit Code* CNS RANKS.

Table 3.8: Chi-square test for trend of elevated rNRE across CNS RANKS (N= 155).

	***4-Digit Code* CNS RANK**				
+2SD = rNRE>3.2504	**1**	**2**	**3**	**4**	
Negative	76	34	9	7	126 (81.3%)
Positive	4	13	4	8	29 (18.7%)
	80 (51.6%)	47 (30.3%)	13 (8.4%)	15 (9.7%)	155
Chi-square test for trend					
Chi-square (trend)	23.964				
DF	1				
Significance level	$p < 0.0001$				

Hypothesis 4: There were clinical IQ estimates available for 66 of the 75 children in the FASD group with IQ estimate z-scores ranging from a low of -2.20 to a high of +2.00. Pearson's correlation co-efficient indicated a modest but significant negative correlation (-0.33, $p = 0.0066$; 95% Confidence interval for r from -0.53 to -0.097) between rNRE and the clinical IQ estimates. Figure 3.5 presents the distribution of rNRE by z-score for IQ estimate with age-strata indicated (filled triangle = age 6-8y; empty square = 9-11y; X= 12-14y).

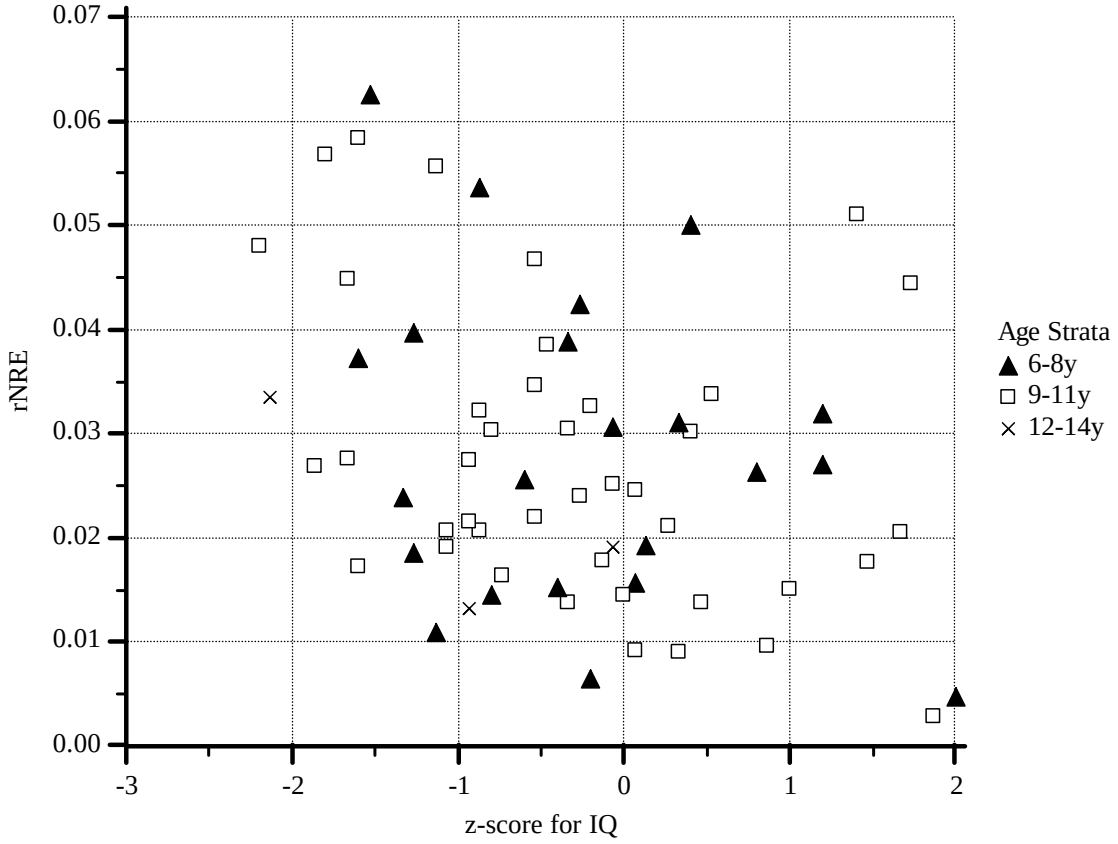


Figure 3.5: Distribution of scores: rNRE by z-score for IQ by age-strata (6-8y, 9-11y, 12-14y).

Hypothesis 5: There were clinical OFC measurements available for 65 of the 75 children in the FASD group with OFC percentile ranging from a low of 1st percentile to a high of 99th percentile. Kendall's Tau indicated a non-significant association (Tau = -0.109; p=0.1985; 95% Confidence Interval for Tau from -0.265 to 0.0576) between rNRE and OFC percentile. Figure 3.6 presents the distribution of rNRE by head-size.

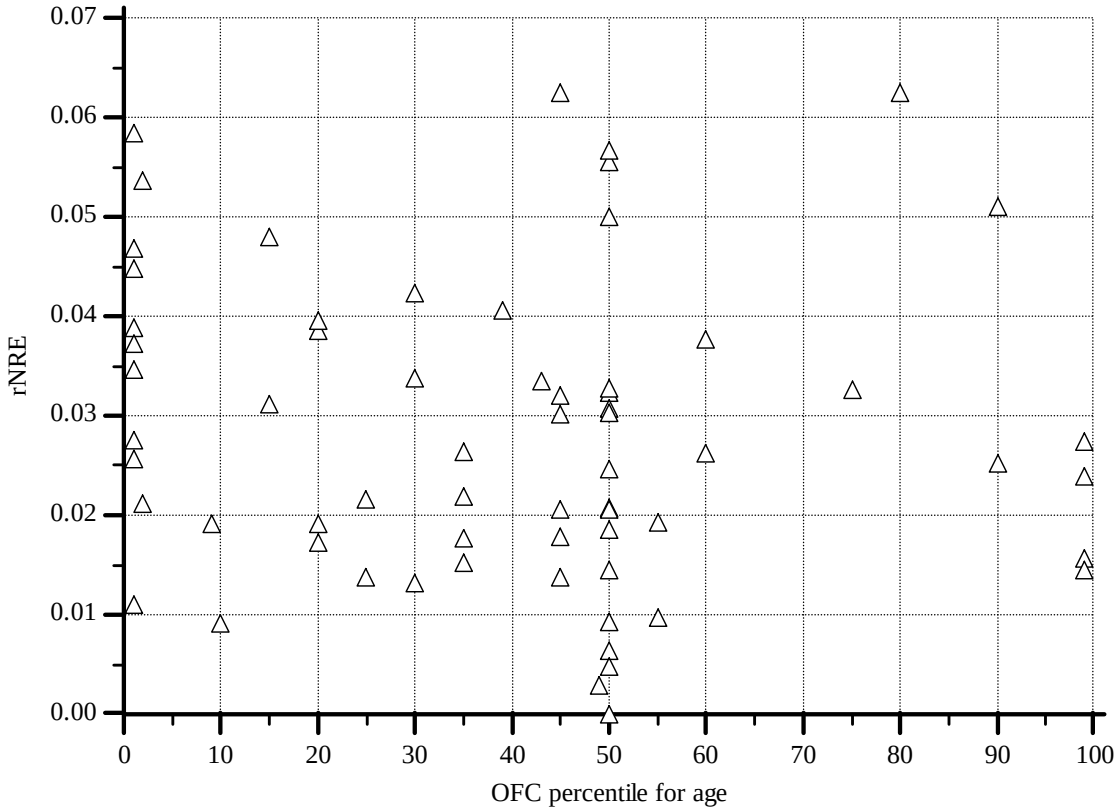


Figure 3.6: Distribution of scores: rNRE by OFC percentile for age.

Hypothesis 6: With 27.3% (3 of 11) of children with a FACE RANK of 4 and 34.4% (22 of 64) of children without a FACE RANK of 4 having significantly elevated rNRE (i.e., rNRE>3.2504), based on a Chi-square test, the proportion of children meeting that criterion was not significantly different between the children with a *4-Digit Code* FACE RANK of 4 and those without. Results are presented in Table 3.9.

Table 3.9: Chi-square test of the proportion of elevated rNRE compared between children with *4-Digit Code* FACE RANK of 4 and those without.

	***4-Digit Code* FACE RANK**		
+2SD = rNRE>3.2504	**1 or 2 or 3**	**4**	
Negative	42	8	50 (66.7%)
Positive	22	3	25 (33.3%)
	64 (85.3%)	11 (14.7%)	75
Chi-square	0.013		
DF	1		
Significance level	p = 0.9081		
Contingency coefficient	0.013		

Post-hoc Analyses Related to Specific Aim One:

To better understand the impact that age had on results from Study One and Study Two, post-hoc analyses were conducted.

Post-hoc Analysis for Study One:

Hypothesis 1: Additional exploratory analysis using t-tests were run on a subset of 51 age-matched pairs in the sample (+/- 6 months; range 6y 9m – 12y 9m). Age-adjusted results were redundant to those found in whole-group analysis having no impact on outcomes.

Hypothesis 2: When children from the TD group below the age of 9 were excluded 45 TD children were left in the 9-and-up TD group. Mean and standard deviation of rNRE for the older children in the TD group was calculated to provide a new set of cut-off scores. The resulting mean for rNRE was 1.238% (SD 0.7403). This resulted in an impairment cut-off (+2SD) of rNRE>2.7186%. In the 9-and-up age group, our prediction that at least 50% of the FASD group would have elevated rNRE falling in the impaired range would be accepted (see Table 3.10).

Table 3.10: Percentage of FASD cases falling above a +2 SD rNRE cut-off compared to prediction of 50% or greater (test for one proportion).

Cut-off scores based on TD group cases aged 9 and up (n = 45)	
+2SD cut-off = rNRE>2.7186	
Observed proportion	45.83% (22 of 45 children)
95% CI of observed proportion	31.37% to 60.82%
z statistic	0.578
Significance level	p = 0.5634: **accept prediction**

Post-hoc Analysis for Study Two:

Hypothesis 3: To explore the impact of age on these results, chi-square test for trend was run with the 6-8y age-stratum removed. When children from the TD group below the age of 9 were excluded 45 TD children were left in the 9-and-up TD group. Mean and standard deviation of rNRE for the older children in the TD group was calculated to provide a new set of cut-off scores. The resulting in mean for rNRE was 1.238% (SD 0.7403). This resulted in an impairment cut-off (+2SD) of rNRE>2.7186%. As seen in Table 3.11, a significant linear trend is found in the older age group. As can be seen in Figure 3.7 below, the 3-step trend identified in the whole-group analysis is more pronounced when analysis is restricted to the older age strata.

Table 3.11: Chi-square test for trend of Impairment across CNS RANKS (ages 9y and up).

	***4-Digit Code* CNS RANK**				
+2SD = rNRE>2.7186	**1**	**2**	**3**	4	
Negative	43	17	6	3	69 (74.2%)
Positive	2	12	4	6	24 (25.8%)
	45 (48.4%)	29 (31.2%)	10 (10.8%)	9 (9.7%)	93
Chi-square (test of trend)	20.090				
DF	1				
Significance level	p < 0.0001				

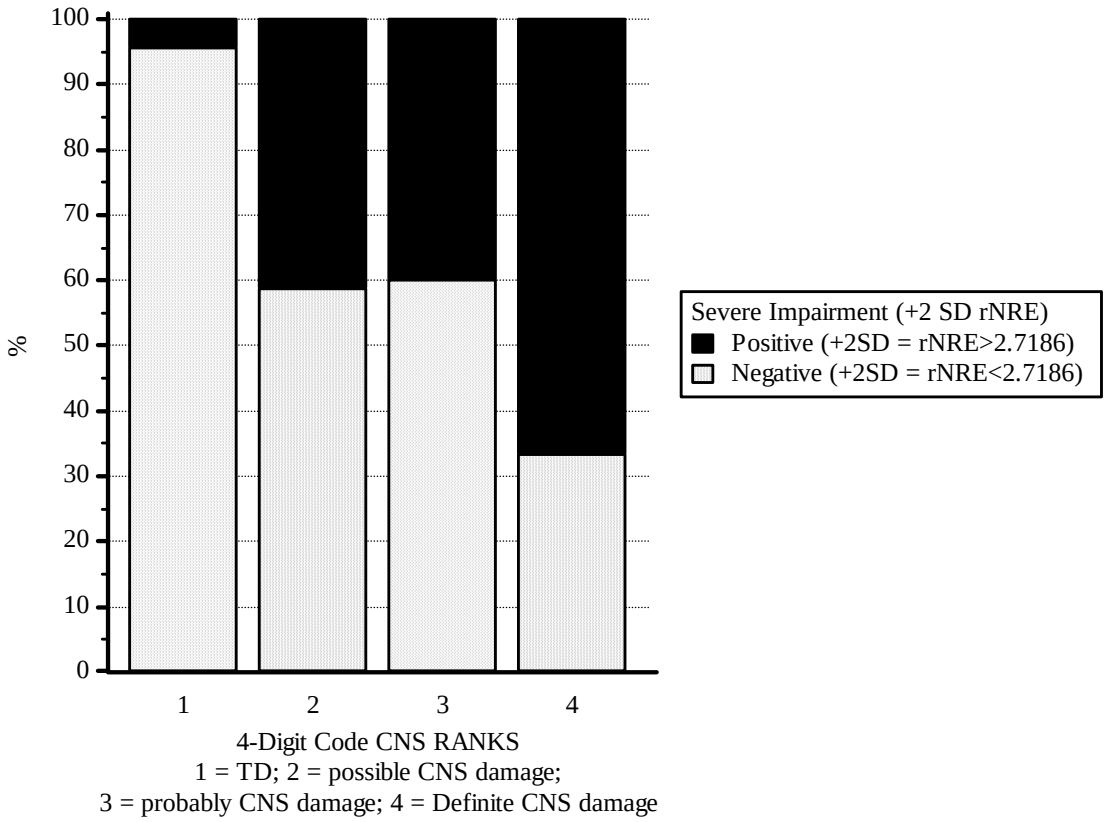


Figure 3.7: Proportion of children above and below impairment cut-off of +2SD in *4-Digit Code* CNS RANKS (ages 9y and up, N=93).

SPECIFIC AIM TWO:

An important long-term goal for establishing the construct validity of the TREIN for use with children suspected of having an FASD is to show that TREIN outcome measures are elevated in individuals with PAE at a higher rate than in the general population (i.e., criterion (ii) presented above). However, the retrospective nature of our clinical sample made it impossible to design studies that would allow us to come to firm conclusions on this question, even if positive results were found. This is primarily because of the fact that the available sample of children with confirmed prenatal alcohol exposure (PAE) comes from a clinical population. While some of these children were referred for assessment simply because PAE was either confirmed or strongly suspected, the vast majority were assessed in response to clinically important developmental

concerns in the context of a confirmed or suspected PAE. Children without these developmental concerns who might have a PAE are missing from our available sample (or, if present, may be unidentified in our TD group). A true test of criterion (ii) would require a prospective confirmation of prenatal alcohol exposure in a random sample of the general population to identify children with PAE across the full range of potential outcomes, and to confirm a lack of PAE in the control group. Of course, as discussed in Astley et. al [59, see page 1675], when this effort is made, it comes with its own set of complications as the group with confirmed lack of exposure may be atypical in other ways. As a result of this limitation, the study described below was designed to examine a weaker version of criterion (ii). The weaker version of criterion (ii) states that "markers of impairment from a TREIN analysis will be found in individuals with Fetal Alcohol Spectrum Disorders associated with PAE that do not have full FAS at a higher rate than in the general population."

Specific Aim Two: To estimate what proportion of elementary school-aged children present with significant integrative language impairment, when impairment is defined as rNRE greater than 2.0 standard deviations from the mean of the TD group, across four diagnostic groups with increasing severity of diagnostic outcomes.

Study Three

Hypothesis:

As severity of diagnosis increases from TD to Neurobehavioral Disorder to Static Encephalopathy to FAS, the proportion of children performing in the impaired range based on rNRE will increase (with impairment defined as +2 SD from the TD group mean).

Design:

Using the same design as Study Two, Study three examined the association between the TREIN measure rNRE and severity of diagnosis taking advantage of the gradation in severity of

diagnosis available across our sample, including both the TD and the FASD groups. Groups were defined according to criteria used by Astley et al. [59]. It should be noted that unlike the cases in the Astley et al. study, the TD group was not formally evaluated to confirm a lack of PAE. Since, as discussed above, four of the children in the FASD group had unknown alcohol exposure, they were excluded from this analysis as their impairments do not meet criteria for FASD.

Analysis:

Along with visual inspection of the distribution of impaired versus unimpaired children across categories, a chi-square test for trend (with a two-tailed p-value of 0.05) was used to compare the proportion of children performing in the impaired range for each categorical rank.

Results for Study Three:

Based on a Chi-square test for trend, the proportion of children meeting criteria for impairment based on elevated rNRE during a narrative was significantly different between diagnostic categories with a linear trend indicating larger proportions in groups with more severe diagnoses moving from TD to Neurobehavioral Disorder to Static Encephalopathy to FAS (see Table 3.12 & Figure 3.8). Based on visual inspection of Figures 3.8 this trend appears to be a 3-step rather than 4-step trend with proportions of children with impairment in the middle diagnostic categories (Neurobehavioral Disorder and Static Encephalopathy) being similar.

Table 3.12: Chi-square test for trend of elevated rNRE across Diagnostic Category.

	Diagnostic Category				
Impairment (+2SD = rNRE>3.2504)	**TD**	**Neurobehavioral Disorder**	**Static Encephalopathy**	**FAS**	
Negative	76	31	14	3	124 (82.1%)
Positive	4	11	8	4	27 (17.9%)
	80 (53.0%)	42 (27.8%)	22 (14.6%)	7 (4.6%)	151
Chi-square (test for trend)	22.941				
DF	1				
Significance level	P < 0.0001				

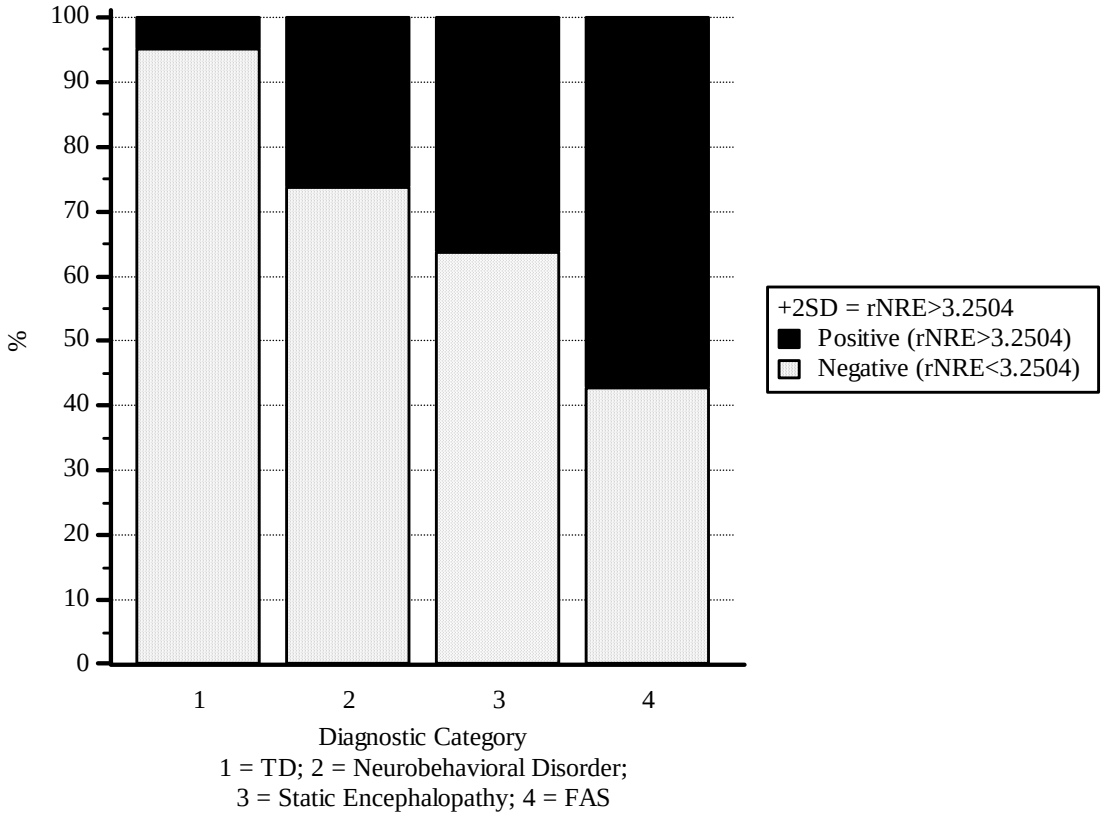


Figure 3.8: Proportion of children above and below impairment cut-off of +2SD in 4 diagnostic categories (N=151).

Post-hoc analysis related to Specific Aim Two:

To explore the impact of age, tests were run with the 6-8y age-stratum removed. There was a linear trend across diagnostic categories in the older group and the 3-step trend identified in the whole-group analysis is more pronounced in the upper age-strata (Table 3.13 and Figure 3.9).

Table 3.13: Chi-square test for trend of impairment across Diagnostic Category (ages 9y and up).

Impairment (+2SD = rNRE>2.7186)	Diagnostic Category				
	TD	Neurobehavioral Disorder	Static Encephalopathy	FAS	
Negative	43	14	10	0	67 (74.4%)
Positive	2	11	7	3	23 (25.6%)
	45 (50.0%)	25 (27.8%)	17 (18.9%)	3 (3.3%)	90
Chi-square (test for trend)	21.164				
DF	1				
Significance level	$p < 0.0001$				

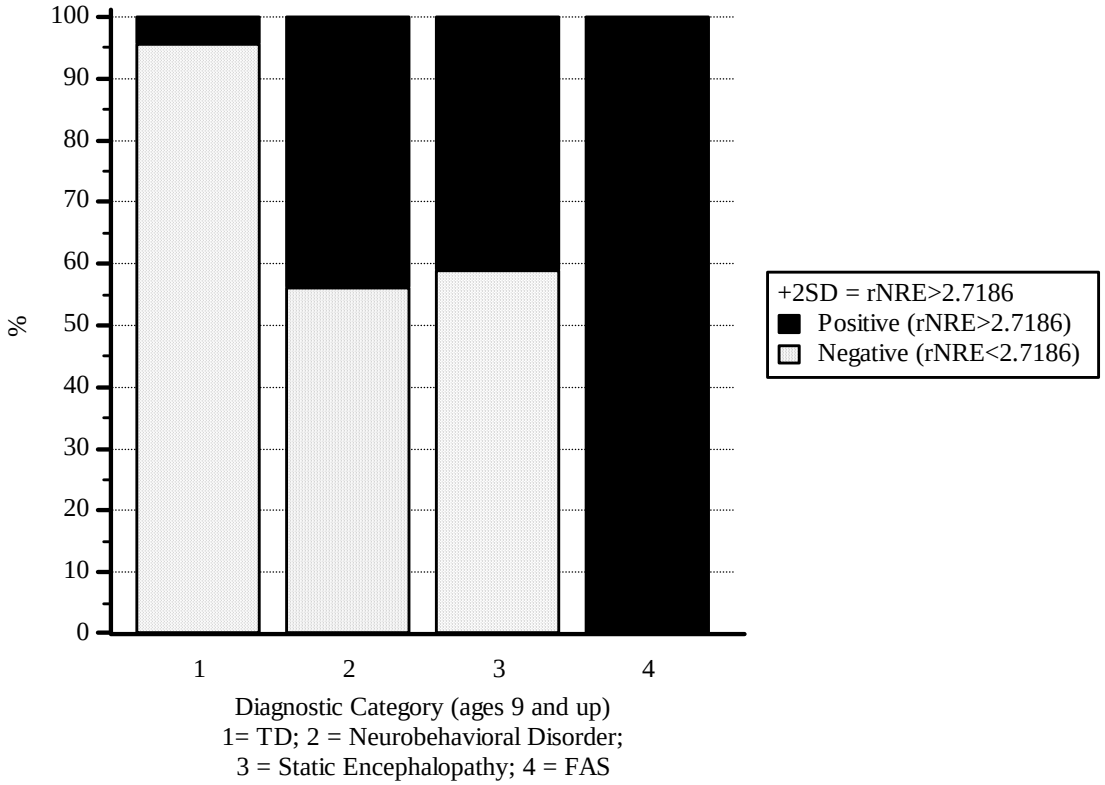


Figure 3.9: Proportion of children above and below impairment cut-off of +2SD in 4 diagnostic categories (ages 9y and up, N=90).

SPECIFIC AIM THREE

According to criterion (iii) described above, if rNRE is a valid measure of ability that is not redundant with existing measures of language ability, then rNRE should be "state-independent." In other words, impairment based on rNRE would be expected in some individuals whether or not the individual's current performance would lead to a diagnosis of a more general "Language/Communication Impairment." This notion of state-independence for rNRE is predicated on the presumption (implied by the WHO categorization of functioning [31]) that Integrative Language functioning is a separate domain of functioning from Expressive and Receptive Language functioning, generally. Since rNRE is designed as a measure of Integrative Language functioning, its validity, then, depends, in part, upon an ability to show independence from measures of Expressive and Receptive Language. Theoretically, it would be expected that some children would show impairments in rNRE in the absence of impairments identified using measures of Receptive/Expressive language, while others would show impairments on measures of Receptive/Expressive language in the absence of impairments identified using rNRE. Some children may, of course, show impairments on both types of measures. If rNRE is state-independent, the overlap between these two groups would need to be relatively minimal—shown by weak associations between rNRE and more traditional assessments of Receptive/Expressive language ability. As comprehensive assessments of Receptive/Expressive language abilities were not available for all children in our sample, exploratory analyses were conducted with the available measures and results are interpreted with caution.

Specific Aim Three: To explore the association between impairments of Integrative Language impairment indicated by an elevated rNRE and more general language impairment identified during clinical assessment of language ability.

Study Four:

Hypotheses:

1. Among the children in the FASD group for whom both measures are available, the rNRE will be negatively but weakly correlated (Tau<0.30) with scaled scores from a standardized measure of Expressive Language ability, "Recreating Sentences/Speech Acts" from the *Test of Language Competence*.
2. Among the children in the FASD group for whom both measures are available, the proportion of children performing in the impaired range based on rNRE will not increase reliably as severity of Language Impairment RANK (see below) increases in severity from RANK 1 (no impairment) to RANK 2 (Mild to Moderate Impairment) to RANK 3 (Severe Impairment).

Design:

All analyses were conducted on measures available within the FASD group. Comparisons were made between Integrative Language performance reflected in rNRE and Expressive language performance reflected in an available standardized measure of expressive language, and a clinical ranking of language impairment. Group size for each comparison was determined by the number of participants for whom both measures were available. Details of these measures are presented below.

Available Standardized Language Testing. The *Test of Language Competence* (TLC) subtests "Recreating Speech Acts" and "Recreating Sentences" were used frequently as part of the clinical assessment conducted at the FASDPN. These subtests are designed to be equivalent measures of a child's ability to produce complex/coordinating sentences, with each subtest aimed at different age-ranges [124]. For children aged 5-9, "Recreating Speech Acts" is used. For children ages 10-18, "Recreating Sentences" is used. The subtests require the child to create a grammatically correct phrase (younger children in "Recreating Speech Acts") or sentence (older children in "Recreating Sentences") that is appropriate to a series of discrete pictured scenes using specific vocabulary presented by the clinician. Since appropriate responses do not require

any information from previous discourse, Integrative Language abilities are not needed to produce an appropriate response on either subtest. These TLC subtests are measures of Expressive Language ability. Both subtests have a mean a scaled score of 10 and a standard deviation of 3. Scores below 5 are more than -2 SD from the mean and indicate clearly impaired performance. Since some children had multiple clinical and research assessments available in our data base, some had multiple results from both subtests. For the current study, if more than one subtest was available, the one that was administered closest to the date of the child's narrative was used.

Clinical Ranking of Impairment. Available clinical assessment data from the FASDPN included a Likert ranking of degree of language impairment that is based on standardized test scores available at the time of assessment (see the *4-Digit Code* manual for details [1]). This Likert Scale of language impairment was used as the Language Impairment RANK for each child when testing hypothesis 2. This RANK will identify children as: RANK 3 = Severe impairment; RANK 2 = Mild-Moderate impairment; and RANK 1 = No Impairment. At the time of clinical assessment, a child may have received a RANK of 0 indicating that there was insufficient information available at the time of clinical assessment to judge language impairment risk. Those with a RANK of 0 were not included in the analysis.

Analysis:

Hypothesis 1: Due to the large number of tied values expected among the TLC subtest scores, association between rNRE and the TLC scores was determine using Kendall's Tau (with a two-tailed p-value of 0.05). Visual inspection of a scatter plot between rNRE and TLC subtest scores provided descriptive information of the distribution across scores.

Hypothesis 2: Along with visual inspection of the distribution of children meeting criteria for impairment according to rNRE (rNRE>3.2504) versus unimpaired children across categories, a chi-squared test for trend (with a two-tailed p-value of 0.05) was used to compare the proportion

of children with an elevated rNRE for each Language Impairment RANK. There were 65 children with a Language Impairment RANK of 1, 2 or 3 available in the sample.

Results from Study Four:

Hypothesis 1: There were subtest scores available for a total of 61 children. As predicted, there was a statistically significant, but weak negative correlation between TLC subtest scores and rNRE (Tau = -0.192; p = 0.028; 95% Confidence Interval for Tau from -0.384 to 0.0277). Figure 3.10 shows that, 21 of 61 children (34.4%) had TLC subtest scores <5 (-2 SD from the normative mean) and 20 of 61 children (32.8%) had rNRE>3.2504, (+2 SD above TD group mean). 11 of the 61 children (18%) would be identified as impaired according to both measures using two standard deviation cut-points, while 30 (49.1%) would not be identified by either measure. Of the 21 children in the impaired range on the TLC (score <5), a total of 9 (42.9%) maintained relatively low rNRE (below 3.2504%). Conversely, of the 20 children who had elevated rNRE>3.2504%, 10 (50%) had TLC scores of 5 or higher (within 2 SD of mean). Of the 20 children who were within 1 SD of the mean on the TLC (>6), 15 (75%) were able to maintain an rNRE below 3.2504%, meaning 25% had elevated rNRE>3.2504%.

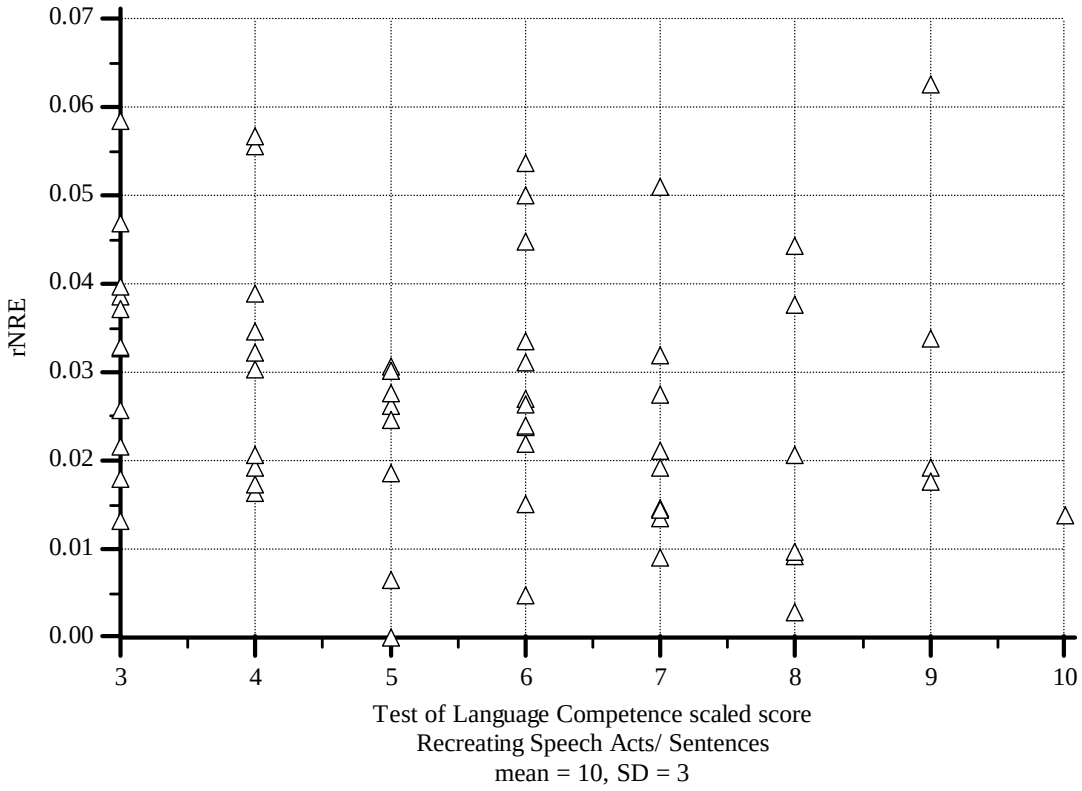


Figure 3.10: Distribution of rNRE by TLC "Recreating Sentences/Speech Acts" subtest score.

Hypothesis 2: 65 children had an available Language Impairment RANK of 1, 2 or 3. As predicted, based on a Chi-squared test and a Chi-squared test for trend, there was no significant difference and no linear trend between the three Language Impairment RANKS in the proportion of children that would be identified as having impairment based on an elevated rNRE>3.2504%. As can be seen in Figure 3.11 and Table 3.14, the largest proportion (42 % of RANK 1 compared to 30% of RANK 2 and 35% of RANK 3) was, in fact, found in the group with the lowest ranking of language impairment risk, RANK 1 = "normal" (i.e., no impairment).

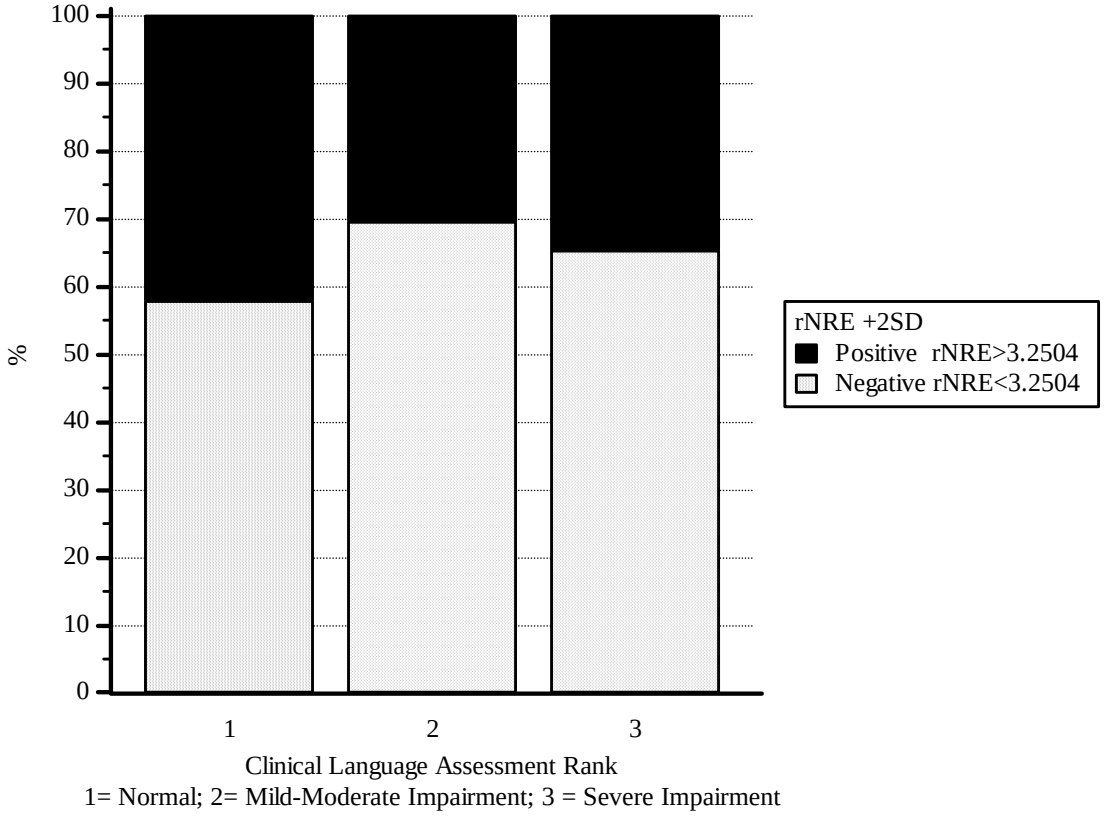


Figure 3.11: Distribution of children identified as having impairment based on rNRE>3.2504% across Language Impairment RANKS (1 = No Impairment; 2 = Mild-Moderate Impairment; 3 = Severe Impairment).

Table 3.14: Chi-square test for trend comparing rNRE to Language Impairment RANK.

Impairment (+2SD = rNRE>3.2504)	Language Impairment RANK			
	1	2	3	
Negative	11	16	15	42 (64.6%)
Positive	8	7	8	23 (35.4%)
	19 (29.2%)	23 (35.4%)	23 (35.4%)	65
Chi-Square test				
Chi-square	0.625			
DF	2			
Significance level	P = 0.7314			
Chi-square test for trend				
Chi-square (trend)	0.210			
DF	1			
Significance level	P = 0.6469			

SPECIFIC AIM FOUR:

While Specific Aims One, Two, and Three examined the validity of rNRE as a behavioral marker of underlying CNS impairment, it was also important to determine the clinical utility that a TREIN analysis has for use in identifying underlying impairment when a FASD is suspected. To explore the potential clinical utility of a TREIN analysis several hypotheses were tested. It was predicted that rNRE would be the most useful among the TREIN measures in elementary school aged children.

Specific Aim Four: To explore the relative clinical utility of each TREIN measure for use in identifying CNS impairments in elementary school-aged children suspected of having an FASD.

Study Five:

Hypotheses:

1. Among all TREIN outcome measures (including rNRE, rNREopp, rPRE, rPREopp, rALL, rALLopp) rNRE will be the more discriminative between the TD and FASD groups than those measures that incorporate pronominal reference errors as reflected in a higher Area Under the Receiver Operating Characteristic Curve (AUC) when predicting group status.
2. Among the various TREIN outcome measures (including rNRE, rNREopp, rPRE, rPREopp, rALL, rALLopp), a greater proportion of children will be classified as having impairment (i.e.,+2.0SD from the mean of the TD group) in the FASD group when using rNRE to define impairment than when using TREIN measures that incorporate pronominal reference errors to define impairment.

Design:

Group-level performance was compared between the two study groups: 1) children with impairment identified in an evaluation of suspected FASD (N = 75; 71 with a FASD and 4 with neurobehavioral disorder, alcohol unknown), and 2) typically developing peers (TD; n= 80).

Analysis:

Hypothesis 1: For each TREIN outcome measure, Area Under the Receiver Operating Characteristic Curve (AUC) was calculated. AUC represents a simple effect-size metric of the degree of separation between the performance distributions of the two groups. A stronger association between group membership and the measure will be reflected in a higher AUC. A pairwise comparison (2-tailed alpha set at 0.05) of AUC for each TREIN measure to rNRE was conducted following procedures from Delong et. al [125]. This procedure provides the probability that the observed difference between two AUC's is equal to zero.

Hypothesis 2: Proportion of cases classified as having an elevated rNRE (+ 2 SD from TD mean) was calculated for each measure and compared across measures. Cochrane's Q-test was used as an omnibus test of whether the various measures identified significantly different frequencies of children with elevated rNRE in the FASD group with pairwise comparisons conducted to identify which measures explained the omnibus result.

Results for Study Five:

Hypothesis 1: Area under the receiver operating curve (AUC) for predicting FASD group membership was calculated for each TREIN measure and a pairwise comparison to the rNRE was conducted [125] (see Figure 3.12 and Tables 3.15). While rNRE has the highest AUC among all measures, it is only significantly better than measures based on pronominal reference errors (rPRE and rPREopp). As a result of the lower accuracy of pronominal measures, the combination of errors in rALL and rALLopp is less accurate than use of nominal errors alone, but the

difference is non-significant in a pairwise comparison. Note that for all measures, performance is virtually identical for pairs that only differ by which denominator is used.

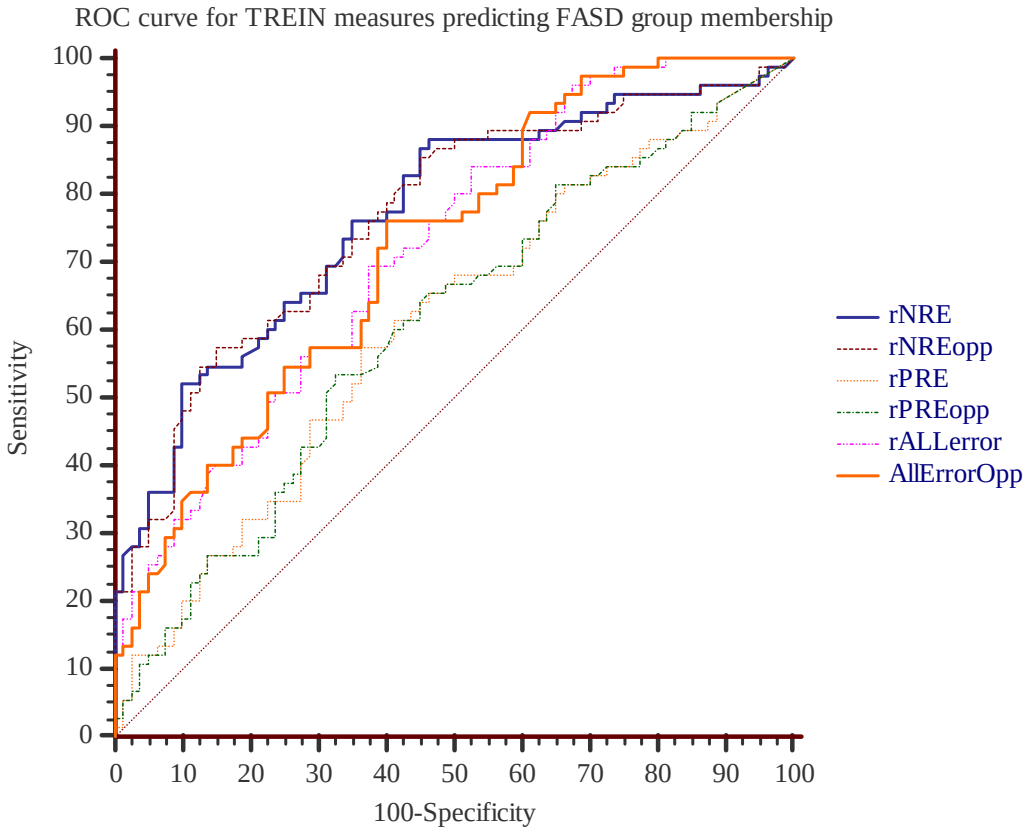


Figure 3.12: ROC curves for TREIN measures predicting *FASD group membership (N = 155). *FASD: n = 75: 71 with a FASD and 4 with neurobehavioral disorder, alcohol unknown; TD: n = 80.

Table 3.15: Pairwise comparison: AUC for TREIN measures predicting *FASD group membership.

Sample size	N = 155	(Positive = *FASD group, n = 75; Negative = TD, n = 80)	
	AUC	SE [a]	95% CI [b]
rNRE	0.769	0.0381	0.694 to 0.833
rNREopp	0.766	0.0383	0.692 to 0.830
rPRE	0.604	0.0456	0.523 to 0.682
rPREopp	0.603	0.0456	0.522 to 0.681
rALL	0.718	0.0403	0.640 to 0.787
rAllopp	0.720	0.0403	0.642 to 0.789

[a] DeLong et al., 1988; [b] Binomial exact

Pairwise comparison of ROC curves to rNRE

rNRE ~ rPRE	
Difference between areas	0.164
Standard Error [c]	0.0503
95% Confidence Interval	0.0658 to 0.263
z statistic	3.268
Significance level	**p = 0.0011**
rNRE ~ rPREopp	
Difference between areas	0.165
Standard Error [c]	0.0503
95% Confidence Interval	0.0668 to 0.264
z statistic	3.287
Significance level	**p = 0.0010**
rNRE ~ rALL	
Difference between areas	0.0508
Standard Error [c]	0.0265
95% Confidence Interval	-0.00110 to 0.103
z statistic	1.918
Significance level	p = 0.0550
rNRE ~ rAllopp	
Difference between areas	0.0493
Standard Error [c]	0.0264
95% Confidence Interval	-0.00258 to 0.101
z statistic	1.862
Significance level	p = 0.0625
rNRE ~ rNREopp	
Difference between areas	0.00258
Standard Error [c]	0.00593
95% Confidence Interval	-0.00904 to 0.0142
z statistic	0.436
Significance level	p = 0.6632

[c] DeLong et al., 1988

*n= 75: 71 with a FASD and 4 with neurobehavioral disorder, alcohol unknown.

Hypothesis 2: As can be seen in Table 3.16 below, while the proportion of FASD cases that would be defined as having impairment of Integrated Language ability (i.e., +2SD from mean of TD group on a measure) was greatest (33.33%) when impairment was defined using rNRE, this proportion was only significantly greater when rNRE was compared to pronominal reference errors (rPRE and rPREopp).

Table 3.16: For all TREIN measures, a comparison of the frequencies of *FASD cases above and below a +2SD cut-point (Cochran's Q Test).

	COUNT (whole FASD group)		Proportion Above
TREIN Measure	Below Cut	Above Cut	%
rNRE	50	25	33.33
rNREopp	53	22	29.33
rPRE	71	4	5.33
rPREopp	66	9	12.00
rALL	59	16	21.33
rALLopp	59	16	21.33
		FASD cases in group (n)	75
		Cochran's Q Test (DF = 5)	45.1961
		Significance	<0.001
		rNRE is significantly larger than (p<0.05)	rPRE, PREopp
		Minimum required difference for significance base on (n)	14.4326%

*N= 75: 71 with a FASD and 4 with neurobehavioral disorder, alcohol unknown.

Post-Hoc analyses related to impact of age on Clinical Utility:

Hypothesis 1: Non-significant differences between unadjusted and covariate adjusted ROC curves were found for all measures when age was used as a covariate [126]. That being said, when ROC curves are created using only cases aged 9 and up, AUC for all measures improves by between 0.05 and 0.07 points. For example AUC for rNRE improves to 0.84 (95% Confidence Interval 0.749 to 0.908) and rNREopp improves to 0.837 (95% Confidence Interval 0.746 to 0.906; see Figure 3.13).

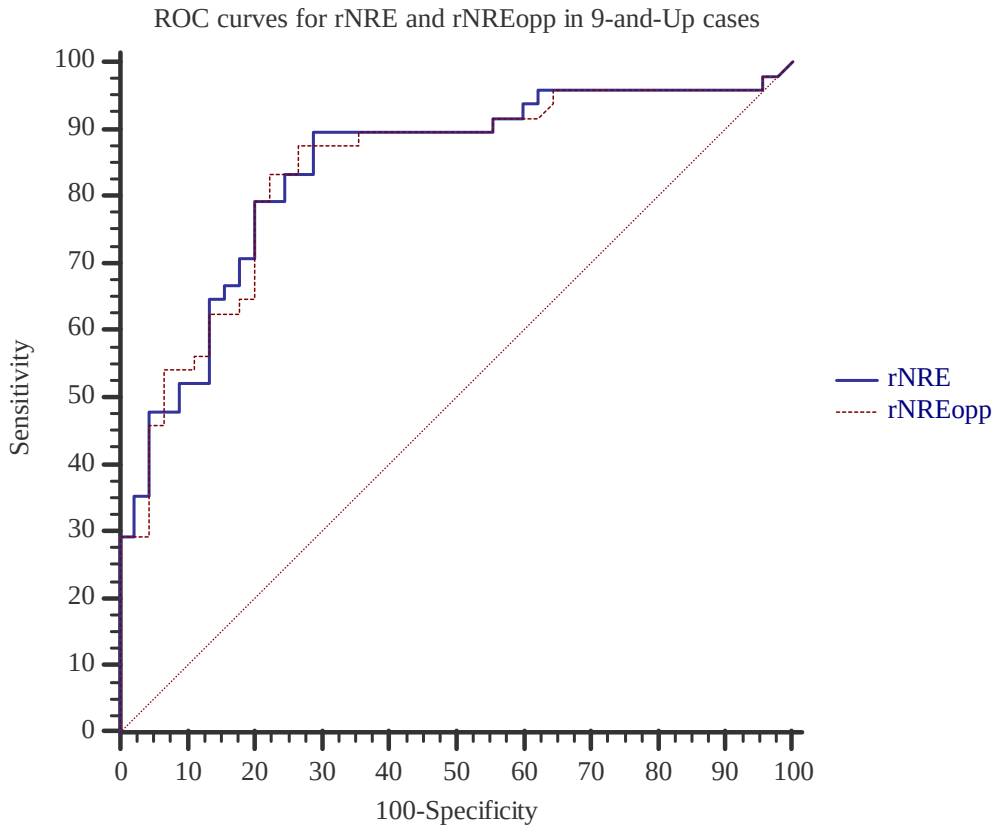


Figure 3.13: ROC curves for rNRE and rNREopp in age 9-and-up group (AUC = 0.84).

Hypothesis 2: In the whole-group analysis Cochrane's Q-test was used as an omnibus test of whether the various measures identified significantly different frequencies of children with elevated rNRE in the FASD group. To better understand the impact of age on this result, groups were stratified into the 3 age-strata: ages 6-8y, ages 9-11y, and ages 12-14y. Within-stratum mean and standard deviation for TD cases defined elevated errors for each measure. Cochrane's Q-test was used in an exploratory manner for each age-stratum when sufficient sample size was available. When age-stratified results were examined (see Table 3.17), it was found that the advantage of rNRE was maintained across age-strata. Due to the mean and SD of each subsequent age-stratum being smaller in the TD group, the advantage increased as one moved from the 6-8y

group, to the 9-11y group, to the 12-14y group with the difference being significant for all measures in the 9-11y stratum. Because of the small number of FASD cases in the 12-14y group, no Q-test was run for this stratum. In the youngest age-stratum the advantage of rNRE was non-significant.

Table 3.17: For all TREIN measures, a comparison of the frequencies of *FASD age-stratified cases above and below a within-stratum +2SD cut-point used to define Impairment (Cochran's Q Test).

	COUNT (6-8y)		Proportion Above	Count (9-11y)		Proportion Above	Count (12-14y)		Proportion Above
TREIN Measure	Below Cut	Above Cut	%	Below Cut	Above Cut	%	Below Cut	Above Cut	%
rNRE	18	9	33.33	27	17	38.64	2	2	50.00
rNREopp	18	9	33.33	29	15	34.09	2	2	50.00
rPRE	21	6	22.22	42	2	4.55	3	1	25.00
rPREopp	21	6	22.22	41	3	6.82	3	1	25.00
rALL	19	8	29.63	36	8	18.18	3	1	25.00
rALLopp	19	8	29.63	38	6	13.64	3	1	25.00
FASD cases in group (n)			27	44			4		
Cochran's Q (DF = 5)			3.8889	43.6466			insufficient n		
Significance			0.566	<0.001			n/a		
rNRE is significantly larger than (p<0.05)			none	rPRE, rPREopp, rALL, rALLopp			n/a		
Minimum required difference for significance base on (n)			33%	19.86%			n/a		
+2 SD cut-point for rNRE in this age-stratum			>3.662%	>3.047%			>1.6516%		

*N= 75: 71 with a FASD and 4 with neurobehavioral disorder, alcohol unknown.

CHAPTER 4

DISCUSSION

Tallying Reference Errors In Narrative (TREIN) is a tool designed to measure an important aspect of Integrative Language functioning—cohesive endophoric reference in both nominal and pronominal phrases during narrative production. To determine the degree to which outcome measures from the TREIN can serve as behavioral markers of underlying CNS impairments, a retrospective comparison of performance on the TREIN was conducted between two groups: a clinical population consisting of children who had a previously diagnosed CNS impairment found during a diagnostic evaluation for suspected Fetal Alcohol Spectrum Disorders (referred to as the "FASD group"), and a group of typically developing peers (the "TD group").

Four specific aims were examined in this research that sought to 1) determine the degree of association between elevated rates of Nominal Reference Errors (rNRE) and previously diagnosed CNS impairments; 2) to explore the degree to which an elevated rNRE was more common in children with FASD, other than FAS, than in the general population; 3) to explore to degree to which Integrative Language functioning as reflected in the rNRE was independent of general Expressive Language abilities in our sample of children; and 4) to explore the relative clinical utility of the various TREIN outcome measures for discriminating between children with impairments and their typically developing peers.

The accumulated evidence presented in this dissertation shows a clear association between elevated rNRE and previously diagnosed impairment. The incidence of elevated rNRE appears to be more common in children in our FASD group than in their typically developing peers. In addition, results indicate that this measure of Integrative Language functioning holds promise for providing important clinical information about communicative functioning that would be missed by evaluations of language ability that focused solely on Expressive and Receptive Language abilities. It also provides evidence that this information about Integrative

Language Functioning has the potential to help clinicians distinguish between children with underlying CNS impairments and their typically developing peers.

The discussion that follows is built around the specific aims that motivated this research. Each aim will be briefly summarized, including a quick summary of the associated hypotheses and pertinent results (including information from post-hoc analyses). For each aim in turn, an analysis and discussion of these results will follow this summary of findings.

Specific Aim One: To assess the degree of to which the rate of Nominal Reference Errors (rNRE) can meet validity criteria (i) described above by showing that rNRE is "associated with disorder (i.e., with clinically diagnosed impairment)."

Hypotheses Tested and Results:

1. The mean rate of errors for all TREIN measures in the *Frog Where Are You* narratives will be greater for FASD group than for the TD group.
    a. Hypothesis accepted for all TREIN measures incorporating nominal reference errors including rNRE, rNREopp, rALL, and rALLopp.
    b. Hypothesis rejected for TREIN measures based on pronominal reference errors including rPRE and rPREopp.
2. The proportion of children in the FASD group who generate a narrative with an rNRE in the impaired range will be 50% or greater with impairment defined by the following performance criteria: greater than 2.0 standard deviations from mean of the TD group.
    a. Hypothesis rejected: however, 33% of combined group met criteria.
    b. Post-hoc analysis indicated that the hypothesis would be accepted if the analysis were restricted to children aged 9 and up (see Table 3.10).
3. As severity of *4-Digit Code* CNS RANK increases in severity from RANK 1 (TD group) to RANK 2 (*possible* CNS damage) to RANK 3 (*probable* CNS damage) to RANK 4

(*definite* CNS damage) the proportion of children with an elevated rNRE (+2 SD of the TD mean) will increase.

   a. Hypothesis accepted with a linear trend indicated by Chi-squared test for trend.
   b. Linear trend found across the 4 ranks appears as a 3-step trend with the middle 2 ranks having similar proportions of children with elevated rNRE.

4. For those children in the FASD group for which both measures are available, rNRE will be inversely correlated with a clinical estimate of Full-scale IQ (i.e., higher rNRE will be associated with lower IQ).
   a. Hypothesis accepted, but many with a high IQ estimate also have elevated rNRE.
5. For those children in the FASD group for which both measures are available, increasing rNRE will be negatively correlated with head size as reflected in clinically measured Occipital-Frontal Circumference percentile (OFC).
   a. Hypothesis rejected, no relationship found.
6. In the FASD group, the proportion of children performing in the impaired range based on rNRE will be greater among those who have a *4-Digit code* FACE RANK of 4 than among those without this FACE RANK.
   a. Hypothesis rejected, no relationship found.

Analysis and Discussion:

Specific Aim One sought to demonstrate an association between elevated error rates from a TREIN analysis and impairment status reflected in an existing clinically diagnosed CNS impairment. Results provide evidence that this association exists to a degree that the association can be considered clinically important.

As was predicted in hypothesis 1, mean group-level performance of the FASD group (who had CNS impairment identified during an assessment for suspected FASD), was significantly worse than that of the TD group on all TREIN measures incorporating nominal

reference errors (rNRE, rNREopp, rALL, and rALLopp) with the mean of the FASD group approximately twice that of the TD group.

Hypothesis 2—which predicted that a majority of children in the FASD group would have rNRE more than 2 SD above the mean of the TD group—was designed as a test of the strength of the association between elevated rNRE and impairment status. It was designed using the logic that any measure for which a majority of the FASD group fell in the impaired range would have a clear and strong association with group membership. The prediction of 50% was based, in part, on an estimate coming from our previous research which found impairment using rNRE in nearly 60% of the FASD group examined. However, the current larger study population included a significant number of children in the early elementary school years (6-8y), a group that was not well represented in our preliminary research. Because age was found to be important in predicting the rNRE in the TD group, inclusion of larger numbers of younger children resulted in a reduction in the proportion of TD children who were maintaining low rNRE relative to the group mean and SD, with many younger children were making frequent errors. As the whole-group mean and resulting 2 SD cut-point was elevated by the presence of these younger children, many older children in the FASD group fell below the cut-point when it was based on the whole-group mean and standard deviation. When the bottom age-strata was removed in post-hoc analysis, the proportion of children performing in the impaired range in the FASD group was statistically consistent with the prediction of 50%, largely due to a decrease in the variance in this older group accompanying the reduced mean rNRE.

Hypothesis 3 is supported by our results and indicates a strong relationship in our sample between the CNS RANK and the proportion of children with elevated rNRE, further supporting the idea that elevated rNRE is associated with disorder. It is important to note that, while there was a clear linear trend across these CNS RANKS based an omnibus Chi-squared test for trend, this appeared as a three step trend, with children in the middle two ranks (RANK 2 and RANK 3) having similar likelihood of having elevated rNRE. Because of the way that CNS RANKS are

determined in the *4-Digit Code*, an increase in CNS RANK indicates an increase in the number of domains of significant impairment identified during a clinical assessment. Because of this, the probability is greater that Integrative Language impairment identified with an elevated rNRE would be found in children with more domains of impairment (CNS RANK 3) than it is for those with fewer domains of impairment indentified (CNS RANK 2). While results strictly support this in the whole group analysis with 27% of CNS RANK 2 and 31% of CNS RANK 3 having elevated rNRE, this difference is smaller than expected and goes away when analysis is restricted to the upper two age-strata in our sample (ages 9 and up). Although it would only be speculative at this point, this pattern of results may suggest that Integrative Language functioning may be particularly sensitive to the types of underlying CNS impairment found in our clinical sample, with elevated rNRE showing up at higher than expected levels in the CNS RANK 2 group.

Hypothesis 4 predicted a correlation between performance as measured using rNRE and the degree of cognitive impairment reflected in a clinical estimate of full-scale IQ. Results indicated a small to moderate relationship between these two measures of cognitive functioning. Looking at Figure 3.5, only one child with an IQ estimate more than 1.5 SD below the mean for their age was able to maintain a low rNRE compared to same-aged peers (recall that in the 9-and-up age-range, an rNRE of 2.72% is 2 SD above the mean). This supports the idea that underlying impairment increases the likelihood a child will have difficulty with narrative cohesion. However, elevated rNRE was found in children even with IQ's estimates in the above average to superior range, suggesting that while there *might* be a minimal IQ level required to maintain a low rNRE, the skills that result in a high IQ estimate are not sufficient to maintain a low rNRE. In other words, rNRE is not redundant to clinical measures used to estimate IQ, but is measuring a substantially separate skill.

Hypothesis 4 and hypothesis 5 were not supported by our findings. While, on balance, our results support the idea that rNRE has potential as a behavioral marker of underlying impairment, the lack of association found between rNRE and the two available structural markers

of CNS impairment provides boundaries on the inferences that can be drawn from the associations found. While these structural measures are certainly valid proxy indicators of underlying CNS impairment, they are non-specific markers of structural impairment in the CNS. For example, while head size is related to the size of the brain, and while large deviations in the size of the brain indicate abnormal brain growth, overall differences in brain size that are equivalent in terms of their impact on head size may result from very different underlying processes in the brain. The lack of association between rNRE and deviation in head size suggests (coherent with our findings based on behavioral measures), that if rNRE is associated with underlying CNS impairment, that association is not with a generalized impairment in CNS integrity, but, rather, with impairment in specific brain systems. As a gross measure of brain growth, head-size does provide specific enough information about the nature of the underlying CNS to make any definitive statements about what type of structural impairment might be associated with an elevated rNRE. Future research using more refined structural measures of CNS impairment across the brain would be needed to find such an association if it indeed exists.

Likewise, while the face of FAS can be considered a proxy of frontal lobe abnormality (because midline facial features and frontal lobe structures share embryological origins) a FACE RANK of 4 provides no details about the extent or nature of that CNS impairment in an individual. Similar to the discussion above related to head size, the lack of specific information about the extent and/or specific nature of the frontal lobe abnormality conveyed by the FACE RANK of 4 only allows us to say that rNRE is not associated with generalized risk of frontal lobe impairment, but provides no ability to make inferences about more specific types of damage. Again, more refined structural measures would be needed to identify specific frontal lobe damage associated with rNRE, if such an association exists.

What is unclear is why these non-specific structural markers of underlying CNS impairment were not associated with elevated rNRE when similarly non-specific, behavior-based clinical assessments of underlying CNS impairment (e.g., CNS RANK, Diagnostic Category)

were associated with elevated rNRE. Speculatively, this may result from the fact that although these structural markers are indeed non-specific, they are more specific than the diagnostic and behavior-based clinical markers used in this research. Specifically, both diagnostic category and CNS RANK largely reflect the *extent* of impairment, with more severe CNS RANK and more severe diagnostic category reflecting an increase in the degree to which impairment is found across neurocognitive and physical feature domains. In this situation, it is perhaps not surprising that as severity using these risk measures increases, so does the likelihood that one of the domains effected is related to the skills needed to maintain a low rNRE. On the other hand, increased deviations in head size reflect changes on a single parameter indicative of underlying CNS abnormality (e.g., hypoplasia) that may or may not have functional consequences related to the episodic control hypothesized to support performance measured using rNRE. Similarly, since the face of FAS is, essentially, an all-or-nothing measure used as a proxy of frontal lobe damage, its presence may reflect a class of damage that is unrelated to episodic control as measured using rNRE.

Additional evidence of association. Additional group-level evidence of the association between impairment as reflected in FASD group membership and performance on TREIN measures comes from unexpected difference between the TD and FASD groups for the within-group correlation of age and narrative performance. While there were children from both groups with very-low or error-free performance in all 3 age-strata on all measures, the proportion of these children in the TD group increased in the older age-strata until the group mean was very low for all measures. In addition the variance in performance in the TD decreased for all measures for each subsequent age-stratum such that no TD children were making significant numbers of errors by the upper age-stratum on any measure. This pattern is what would be expected for any skill in typically developing children. What is striking is the lack of this pattern in the FASD group when performance is measured using the TREIN, particularly those TREIN outcome measures based on nominal reference errors. As a result, there is a significant separation

between the groups in the distribution of TREIN scores based on Nominal Reference Errors (rNRE and rNREopp; see Table 3.15). This significant distributional separation is reflected in the AUC of both rNRE (0.769; 95% Confidence Interval between 0.694 and 0.833) and rNREopp (0.766; 95% Confidence Interval between 0.692 and 0.830). These AUC values indicate the probability that a randomly selected child in the FASD group would have a higher error rate than a randomly selected child from the TD group. With probabilities in the whole group of 70% or higher, and probabilities that increase to 75% or higher in the older age-strata (ages 9 and up; see Figure 3.13), these results provide evidence of an association that is not only statistically significant, but one that is a clinically important association.

Indeed, substantial numbers of children in the FASD group appear to have failed to master nominal reference skills even into the upper age-stratum. This is shown by the fact that 33% of the entire FASD group and approximately 50% of the children in the two upper age-stratum of the FASD group were making errors of nominal reference (rNRE and rNREopp) more than +2SD above the mean of their peers in the TD group (see Table 3.7 and 3.10). It is this difference in the proportion of children in the upper age-strata of the FASD group making frequent Nominal Reference Errors that results in substantial lack of overlap between the distribution of scores for the TD and FASD groups as reflected in the high AUC for both rNRE and rNREopp. This is also reflected in increasing proportions of children in the FASD group above a within-group cut-off for impairment based on rNRE> +2 SD moving from the 6-8 to the 9-11 to the 12-14 age-strata (see Table 3.17).

Specific Aim Two: To estimate what proportion of elementary school-aged children present with significant integrative language impairment, when impairment is defined as rNRE greater than 2.0 standard deviations from the mean of the TD group, across four diagnostic groups with increasing severity of diagnostic outcome.

Hypothesis tested and Results:

As severity of diagnosis increases from TD to Neurobehavioral Disorder to Static Encephalopathy to FAS, the proportion of children performing in the impaired range based on rNRE will increase.

a. Hypothesis accepted with a linear trend indicated by Chi-squared test for trend.
b. Linear trend across the 4 diagnostic categories appears to be a 3-step trend with the middle two categories having similar proportions of children with elevated rNRE.

Analysis and Discussion:

Because this research was based on a retrospective data set, and prenatal alcohol exposure (PAE) was not directly assessed in our control group, it would have been impossible to directly establish the degree to which elevated rNRE is associated with prenatal alcohol exposure. For this reason, a relaxed version of this question was examined, asking whether "difficulty maintaining narrative cohesion in narrative as reflected in an elevated rNRE is found in individuals with Fetal Alcohol Spectrum Disorders associated with PAE that do not have full FAS at a higher rate than in the general population."

In our sample, the answer to this weaker version of the question is yes. Results indicate that, compared to the TD group, there is an significantly increased likelihood of having an elevated rNRE among those children with previously identified CNS impairments including neurobehavioral disorder and static encephalopathy, with the greatest likelihood found in those children with FAS. While there was a clear linear trend across the four groups based an omnibus Chi-squared test for trend, this appeared as a three step trend, with the proportion of children with elevated rNRE in the two middle diagnostic categories (neurobehavioral disorders and static encephalopathy) being similar. As with the finding for CNS RANK, this 3-step pattern appears more robustly when analysis is restricted to the two upper age-strata (i.e., children ages 9 and up; see Figure 3.8 and Figure 3.9). If this result can be confirmed in prospective research that examines performance across the full range of outcomes associated with prenatal alcohol

exposure, rNRE would appear to be a promising candidate for use in the endophenotype approach to understanding the effect of PAE on the Integrative language functioning of children. Results reported here provide sufficient evidence to suggest that this type of research may be worth pursuing in the future development of the TREIN.

Specific Aim Three: To explore the association between impairments of Integrative Language impairment indicated by an elevated rNRE and more general language impairment identified during clinical assessment of language ability.

Hypotheses tested and Results:

1. Among the in the FASD group for whom both measures are available, the rNRE will be negatively but weakly correlated (Tau<0.30) with the scaled score from the TLC subtest "Recreating Sentences/Speech Acts".
    a. Hypothesis provisionally accepted with Tau = -0.192 (p = 0.028).
    b. 95% Confidence Interval for Tau included predicted value of -0.30.
    c. Of the 41 children identified as having impairment with one or the other measure, there was an overlap of only 11 children (26.8%).
2. Among the in the FASD group for whom both measures are available, the proportion of children performing in the impaired range based on rNRE will *not* increase reliably as severity of Language Impairment RANK increases in severity from RANK 1 (no impairment) to RANK 2 (Mild to Moderate Impairment) to RANK 3 (Severe Impairment).
    a. Hypothesis accepted (Chi-squared test for trend = 0.210; p = 0.6469).
    b. 42% of children in RANK 1 (i.e., no impairment) had rNRE in the impaired range (rNRE>3.2504%), the highest of the three RANKS.

Analysis and Discussion:

Both hypothesis 1 and hypothesis 2 were supported by our findings. There was no apparent relationship between the ability to maintain cohesive endophoric reference in nominal phrases as measured by rNRE and more general language ability as reflected in a clinical ranking of degree of language impairment (Chi-square test for Trend = 0.210; p = 0.6469) and only a weak association between rNRE and performance on a common standardized measure of expressive language from the TLC (Tau = -0.192; p = 0.028). Because the 95% confidence interval for Tau included the predicted upper limit of Tau <0.30, however, this association *may* be somewhat stronger than predicted. Similar to the findings related to IQ estimates, the degree to which this relationship exists seems to stem primarily from a decrease in the chances that a child who tests in the impaired range on the TLC subtest will be able to maintain a low rNRE. This slight trend can be seen in Figure 3.10 where 9 of 20 children (45%) with TLC scores in the impaired range maintain rNRE below 3.2504% while 15 of 20 (75%) of the children in the average range on the TLC (>6) were able to maintain this level of rNRE. The state-independence of impairment identified using rNRE from impairment identified using TLC, however, is clearly suggested by the distribution of rNRE across TLC scores; with one-third of the children being in the impaired range on only the TLC, one-third being impaired solely based on an elevated rNRE, and one-third of children meeting impairment criteria for both measures. In addition, the highest rNRE in this group was produced by a child with a TLC score of 9, and the only error free narrative (based on rNRE) was produced by a child with a TLC subtest score of 5 (-1.67 SD below the normative mean), a score that might be considered in the moderately impaired range. Indeed, 25% of the children who would have screened as “unimpaired” using this TLC subtest (scores of 7 or above), would be identified as having impairment if the rNRE from a TREIN assessment were also included in the screening—potentially a clinically meaningful increase in the number of children identified.

If Integrative Language functioning is a separate domain of functioning from Expressive and Receptive Language functioning, the validity of Integrative Language measures depend, in

part, upon an ability to show independence from measures of Expressive and Receptive Language. While our results are certainly preliminary and will need to be confirmed with more systematic research, they suggest that the approach to tallying Nominal Reference Errors utilized in the TREIN has promise in this regard. However, there is a need to clarify the degree to which Expressive Language abilities may be necessary, if not sufficient for successfully maintaining cohesive reference during narratives. While our results indicate that some children with clinically significant Expressive Language impairments may have little or no trouble maintaining cohesive reference in their narratives, the existence of that Expressive Language impairment seems to mildly increase the chances that they will have an elevated rNRE. A need for more systematic exploration of the relationships between Expressive impairments and Integrative Language functioning reflected in the rNRE is indicated by these findings.

Specific Aim Four: To explore the relative clinical utility of each TREIN measure for use in identifying CNS impairments in elementary school-aged children suspected of having an FASD.

Hypotheses tested and Results:

1. Among the various TREIN outcome measures (including rNRE, rNREopp, rPRE, rPREopp, rALL, rALLopp), rNRE will be the more discriminative between the TD and FASD groups than those measures that incorporate pronominal reference errors as reflected in a higher Area Under the Receiver Operating Characteristic Curve (AUC) when predicting group status.
    a. Hypothesis accepted. While rNRE has the highest AUC among all measures, it is only significantly better than measures based on solely on pronominal reference errors (rPRE and rPREopp).
    b. The combination of pronominal and nominal reference errors in rALL and rALLopp is less accurate than use of nominal errors alone, but the difference is non-significant in a pairwise comparison.

c. For all measures, performance is virtually identical for pairs that only differ by which denominator (NTW.a or opp) is used (see Figure 3.12).

2. Among the various TREIN outcome measures (including rNRE, rNREopp, rPRE, rPREopp, rALL, rALLopp), a greater proportion of children will be classified as having impairment (i.e.,+2.0SD from the mean of the TD group) in the FASD group when using rNRE to define impairment than when using TREIN measures that incorporate pronominal reference errors to define impairment.

    a. Hypothesis is accepted as rNRE outperformed rPRE and rPREopp. However, even though inclusion of pronominal errors in rALL and rALLopp reduced the performance of these measures, there was not a statistically significant difference between them and rNRE.

    b. The advantage of rNRE increased moving from the 6-8 to the 9-11 to the 12-14 age-strata with a large enough advantage for rNRE in the middle age-strata to accept this hypothesis for the 9-11 age-range (see Table 3.17).

Analysis and Discussion:

For our discussion related to Specific Aim Four, we will focus on several implications that flow from the results summarized above. These include the relatively greater clinical utility of TREIN measures based on Nominal Reference Errors when compared to those that incorporate Pronominal Reference Errors—and the importance of age in the clinical utility for both types of measures. Next we will look at the clinical impact of different methods available for controlling for the length of the story in a TREIN analysis. Finally, we will discuss the clinical implications regarding the relative performance of TREIN outcomes when compared to the performance of the available standardized language measures used in our research.

Nouns, pronouns, and age. Relative clinical utility of the various TREIN measures was examined in a head-to-head comparison of their ability to discriminate between our TD and FASD groups, and to identify children with clinically important impairments of functioning. As

expected for this age range, there was a significant performance gap between those measures based on Nominal Reference Errors (i.e., rNRE and rNREopp) and those that were based on the ability maintain referential cohesion in pronominal phrases (i.e., rPRE, rPREopp). This advantage was reflected both in a statistically significant advantage seen in pairwise comparisons of AUC when predicting group membership, and in a statistically significant difference in the number of children identified by when a +2 SD cut-point was used to identify Integrative Language impairment. The advantage of measures based on Nominal Reference Errors resulted not because children in the FASD group did not make significant numbers of errors on pronominal phrases (they did, as expected), but was, rather, the result of the greater number of errors on pronominal phrases made by the TD group in the two lower age-strata. This can be seen clearly below in Table 4.1 below, which shows performance of the TD group on the two pronominal errors rates, rPRE and rPREopp, broken down into three age-strata (6-8y, 9-11y, 12-14y). Notice that the mean and standard deviation for each error measure is similar for the two lower age-strata. In the upper age, stratum, however, mean error rates drop to about one-quarter of those in the younger strata for both measures. The range of error rates also decreases with age. For example, while the maximum rPREopp for children in the younger two TD age-strata was between 16% and 24% of opportunities, the worst performing children in the upper TD age-stratum maintained an error rate below 6%.

Table 4.1: Performance of TD group on measures of cohesive pronominal reference errors.

	rPRE			rPREopp		
TD	N= 80 — Age Strata—			N= 80 — Age Strata—		
	6-8y	**9-11y**	**12-14y**	**6-8y**	**9-11y**	**12-14y**
N	35	27	18	35	27	18
Mean	1.43%	1.53%	0.355%	5.13%	5.51%	1.35%
Variance	0.0114	0.0297	0.0014	0.146	0.373	0.0193
SD	1.065	1.725	0.3751	3.817	6.106	1.390
SEM	0.1801	0.3319	0.08842	0.6451	1.175	0.3276
Maximum error rate observed	4.15%	7.14%	1.54%	**16.3%**	**23.9%**	**5.68%**

This difference in performance on pronominal reference in the younger age-strata was not unexpected and motivated the focus on rNRE in the studies above. What was not predicted, however, was the clinical importance of age when using measures of nominal reference. Practically, this finding means that when Nominal Reference Errors are elevated, inferences related to the presence of underlying CNS impairment need to be made taking the age of the child into account. Clinical cut-points will need to be developed for children across the elementary school years, so that appropriate comparisons can be made. Even in the youngest age-stratum, 33% of children in the FASD group had rNRE outside the range seen in the same age range in the TD group. This points to a clinically important utility for rNRE even in the younger age-range as long as age is taken into consideration during clinical assessment. The efficiency of rNRE as a clinical measure, however, clearly increases during the later elementary school years. To put this 33% rate of impairment into a clinical context, Astley [127] reported on clinical outcomes for 1270 children undergoing interdisciplinary assessment at the University of Washington FASDPN. Of all the children considered “at risk for CNS impairment” (i.e., having a *4-Digit Code* CNS RANK of 2-4), only 8% would be identified as “at risk” if a Full Scale Intelligence Quotient

below 70 was considered in isolation, and only 35% would be identified if a clinical diagnosis of ADHD were considered in isolation. While rNRE did not meet our strict test of 50%, the 33% impairment rate found in the whole group appears to have as much utility as other measures currently being used to identify impairment when children are suspected of FASD.

Controlling for Length. Results reported above clearly showed that there was not a clear empirical advantage for measures from the TREIN that controlled for narrative length using the length of the story (NTW.a from a SALT analysis) when compared to measures that controlled for length based on the number of referential opportunities. In terms of clinical utility, this means that practical matters can be used to decide between measures without fear of reducing the validity of an assessment. The clear advantage of using NTW as a control for length comes from its ease of calculation. Because a standard SALT analysis will include NTW, calculations of TREIN measures based on it are computationally simple. The clinical disadvantage of using NTW is the relatively more complex interpretation of the score that results (e.g., rNRE represents the percentage of words in the narrative that are core nouns in ambiguous nominal phrases). Conversely, while using the number of referential opportunities is computationally more labor intensive (although still relatively simple); the interpretation of scores based on opportunities is much more direct. In addition, controlling for length based on number of referential opportunities raises the possibility that TREIN style reference errors can be tallied on-line, without the need for transcription—something that would be impossible if the total number of words were needed to create an error rate.

Comparison to a standardized language measure. The clinical utility of the rNRE as a way to measure integrative language functioning is also strengthened by results from our comparison of rNRE to a standardized language measure, the “Recreating Sentences/Speech Acts” subtest of the TLC. As mentioned above, similar percentages of children in the FASD group would be identified as having impairment whether a clinician chose to use a two standard deviation cut-point on the TLC or the rNRE as a screen for the presence of communication

impairment. Specifically, 21 of 61 children (34.4%) had TLC subtest scores <5 (-2 SD from the normative mean) and 20 of 61 children (32.8%) had rNRE>3.2504, (+2 SD above the mean of the TD group). In the type of screening conducted during a diagnostic evaluation for FASD like those conduced at the FASDPN, of course, a strict cut-point like this is would be used to set a lower limit on the number of false positives resulting from that screening. This is motivated in part because, in the final diagnosis, multiple measures will be combined using "OR" rules, a practice, as discussed in Chapter One, that increases sensitivity for impairment (broadly defined). When this is done, there is a need to minimize the unavoidable increase in the number of false-positives this combining of measures may produce. In this context, the clinical utility of a tool is determined in part by the sensitivity to impairment it has at the predetermined false-positive rate. As rNRE only identified one less child in the group of 61 with previously identified impairments, it appears to have similar validity for use in this context as use of the TLC. If both tools combined using an "OR" rule were implemented using these strict cut-points, 40 out of 61 children would be identified with impairments in our FASD group, with an overlap of 11 children. The fact that elevated rNRE was found in 42% of children who received a clinical Language Impairment RANK of 1 (i.e., no impairment) emphasizes this point, indicating the inclusion of rNRE in a clinical assessment has the potential to outperform current clinical practices by identifying children with Integrative Language impairments that would be missed otherwise. Given the exploratory nature of the examination done here, future research would, of course, be needed to verify this potential.

Conclusion:

The valid measurement of Integrative Language functioning is a clinical problem that presents a number of interesting challenges. Chief among these is the need, when evaluating underlying CNS capacities, to measure behaviors of interest in a standardized context; a significant challenge when the capacity being measured is the ability to respond to a dynamically

changing context. Narrative discourse analysis is one promising avenue for addressing this challenge in elementary school aged children. If that analysis is going to produce a valid clinical measure of Integrative Language functioning, however, there is a need to strategically choose which aspects of the narrative performance are measured and to verify that the choices made provide clinically important information. For instance, it might be expected that the tool provide information that helps clinicians to distinguish between children with and without underlying impairments. The development of the TREIN was premised on the idea that this is most effectively done by focusing on tallies of errors in obligatory forms in the narrative. For Integrative Language functioning those obligatory forms need to be ones that meet discourse obligations that operate across sentence boundaries (in Tier-2 of the narrative structure). This is done in a TREIN analysis by quantifying errors of endophoric reference during a story generation task for a naïve listener.

Validation of a new clinical measure also presents a number of interesting challenges, particularly when that measure focuses on aspects of behavior for which there are not existing well-validated tools. Since the TREIN is designed to measure an aspect of Integrative Language functioning for which there is not an existing well validated tool, validation could not be conducted by comparing its performance to another well validated measure. In this context, construct validity needs to be demonstrated in other ways. This was done in this research project by showing an association between elevated rNRE and previously diagnosed CNS impairments, by showing that elevated rNRE is more common in children with FAS and other FASD than their typically developing peers, and that it is correlated with other indicators of CNS impairment (e.g., IQ estimates, and a clinical ranking of severity of CNS impairment risk). Preliminary evidence also suggests that elevated rNRE is independent of general Expressive Language functioning captured in a standardized assessment of Expressive Language. These findings provide an important first step towards validating the TREIN for clinical use, but leave open several questions worth exploring.

For instance, why would TREIN measures of Nominal Reference Errors be associated with CNS impairments like those found in our FASD group and what is the role of prenatal alcohol exposure in this assocation? At this point any answer to these questions would be purely speculative. Certainly, as was discussed in the introduction, since these measures were designed to be measures of episodic control abilities as they manifest during expressive discourse, it would be tempting to say that those systems that are involved in other episodic control tasks would be involved in episodic control during narrative production. These fronto-striatal brain systems are known to be vulnerable to prenatal alcohol exposure, and difficulty with episodic control appears to be common in children with FAS and other FASD. If a more direct link between elevated rNRE and damage in these fronto-striatal systems can be found, this would certainly move us forward in our understanding of Integrative Language functioning in children. Demonstrating this association, however, is a non-trivial matter. Research into the question would need to be conducted using carefully designed studies of children identified with narrow and specific impairment profiles. I would argue for researchers to utilize an approach to case ascertainment based upon "AND" rules to maximize specificity rather than the "OR" rules that are used clinically to increase sensitivity to impairment (see discussion on page 5 in the introduction).

In other words, cases would be equated, for instance, when they have **exposure X** AND **CNS impairment Y** AND **behavioral outcome Z**. Minimally different groups, then, can be compared based on a contrast along a single dimension of structure, function, or exposure (Similarly, correlational research can be conducted on those factors not held constant by the "AND" rule that is being applied.). Of course, this approach will be most efficient when **CNS impairment Y** and **behavioral outcome Z** are sufficiently prevalent among those with **exposure X**. So, for instance, a researcher might contrast groups of children with similar PAE (e.g., heavy exposure in the 1st trimester) and a specific structural impairment commonly found in this exposure group (e.g., frontal lobe hypoplasia) based on whether or not they perform above some clinically relevant cut-off on a specific cognitive task associated with impairment in the

population with PAE (e.g., having an elevated rNRE). This grouping would then be used to examine, for instance, structural differences between the groups in the CNS to help understand the difference in performance on the integrative language task; or, in contrast, to examine differences in social communication outcome between groups to understand the role that the referential cohesion plays in that success. Of course an iterative cross validation between etiology/exposure, neurocognitive impairment, and the behavioral footprints left by impairment will need to take place to determine if a particular behavioral outcome such as elevated rNRE can serve as a useful diagnostic marker. As these interrelationships are better understood, they can be used to define and refine clinical diagnostic categories to maximize utility for various clinical and research purposes. This iterative process has been described as the "logical spiral of diagnosis" [128, c.f., 129] that ideally moves in a step-wise fashion towards ever refined diagnostic categories of impairment and improved understanding of the etiology, course, prognosis and treatment of those impairments.

This approach is similar to the "symptom-based approach" to endophenotyping recommended by Viding & Blackmore [72] who point out the need to guard against the potential logical circularity of defining a study group using a narrow cognitive impairment while concurrently using that impairment to explain membership in the larger diagnostic group. The multidimensional approach I am recommended avoids this circularity as long as the inclusionary criteria do not include the impairment of interest. In our example above, for instance, note that grouping based on exposure, frontal lobe hypoplasia, and elevated rNRE facilitates examination of other structural (e.g., caudate volume) or functional (e.g., working memory) differences between the groups that do or do not perform above a rNRE cut-off.

The multidimensional "AND" rule approach I am advocating, of course, requires careful consideration of which parameters, when combined with an "AND" rule, allow researchers to appropriately balance homogeneity with generalizability in their case definitions. Based on results presented here, this will include careful consideration of the importance of factors such as age,

severity of impairment, and differences in the timing and severity of exposure. By contrasting children within a carefully defined exposure-based group in terms of both structural and unitary functional parameters, we can begin to tease apart relationships between the exposure and its consequences on the system. These studies will need to include prospective identification of exposure in non-clinical populations if they are to fully explore the effect of PAE on the CNS.

In addition to these studies elaborating the relationship between elevated rNRE and underlying impairment in the context of PAE, there are several other avenues of research indicated by our results. Perhaps primary among these would be studies aimed at better understanding the relationship between rates of Nominal Reference errors and age across the elementary school years, in both TD and clinical populations, including children with FASD. Our results show a significant developmental performance improvement as typically developing children move towards the later elementary school years. Normative studies aimed a determining the variability in typically developing kids across age-groups would be important. If tallies of Nominal Reference Errors are to be used clinically, appropriate age-based clinical cut-offs will need to be determined. These normative studies could also examine narrative performance using other eliciting stimuli besides *Frog Where Are You.* This would provide additional functionality for the TREIN by providing greater potential for its use in tracking progress in intervention (i.e., by allowing for multiple administrations of the measure without memory interfering with the presumption of a naïve listener). Studies aimed at identifying clinical populations of children who may have elevated rates of Nominal Reference Errors beyond these normative limits would, additionally, help to define the clinical populations for which a TREIN analysis may be most useful.

As a narrow measure, of course, the rate of nominal reference errors from the TREIN would, at best, provide only a narrow window into how CNS damage caused by prenatal alcohol exposure may lead to impairments of Integrative Language functioning. As there is a vast wealth of information contained in the narrative discourse production of children, research to identify

other aspects of narrative production that may be associated with underlying impairment would be important as well. These could include investigations of other Tier-2 features of narrative productions (such as conjunctives), as well as macro-structural features in Tier-3 of the narrative.

Measuring Integrative Language functioning in a way that allows comparisons between children is challenging, but the results of the research reported here point to careful narrative analysis as a useful clinical tool with the potential to meet that challenge. Our results also suggest that narrative analysis has the potential to provide important information to researchers who are interested in understanding the underlying differences in children that result in greater or lesser Integrative Language capacity.

Literature Cited:

1. Astley, S.J., *Diagnostic guide for fetal alcohol spectrum disorders: The four-digit diagnostic code*. 3rd ed. 2004, Seattle: FAS Diagnostic and Prevention Network, University of Washington, electronic version available from http://fasdpn.org.

2. Bertrand, J., L.L. Floyd, and M.K. Weber, *Guidelines for identifying and referring persons with fetal alcohol syndrome.* Morbitity and Mortatility Weekly Report: Recommendations and Reports, 2005. **54**(RR-11): p. 1-14.

3. Abel, E.L. and R.J. Sokol, *Incidence of fetal alcohol syndrome and economic impact of FAS-related anomalies.* Drug and Alcohol Dependence, 1987. **19**(1): p. 51-70.

4. Lupton, C., L. Burd, and R. Harwood, *Cost of fetal alcohol spectrum disorders.* American journal of medical genetics. Part C, Seminars in medical genetics, 2004. **127**(1): p. 42-50.

5. Riley, E.P. and C.L. McGee, *Fetal alcohol spectrum disorders: an overview with emphasis on changes in brain and behavior.* Experimental Biology and Medicine (Maywood, N.J.), 2005. **230**(6): p. 357-65.

6. Astley, S.J., et al., *Application of the fetal alcohol syndrome facial photographic screening tool in a foster care population.* Journal of Pediatrics, 2002. **141**(5): p. 712-7.

7. Astley, S.J. and S.K. Clarren, *Measuring the facial phenotype of individuals with prenatal alcohol exposure: Correlations with brain dysfunction.* Alcohol & Alcoholism, 2001. **36**(2): p. 147-159.

8. Sampson, P.D., et al., *On categorizations in analyses of alcohol teratogenesis.* Environmental Health Perspectives, 2000. **108 Suppl 3**: p. 421-8.

9. Streissguth, A.P., et al., *Risk factors for adverse life outcomes in fetal alcohol syndrome and fetal alcohol effects.* Journal of Developmental and Behavioral Pediatrics, 2004. **25**(4): p. 228-38.

10. Sampson, P.D., et al., *Incidence of fetal alcohol syndrome and prevalence of alcohol-related neurodevelopmental disorder.* Teratology, 1997. **56**(5): p. 317-26.

11. Lee, K.T., S.N. Mattson, and E.P. Riley, *Classifying children with heavy prenatal alcohol exposure using measures of attention.* Journal of the International Neuropsychological Society, 2004. **10**(2): p. 271-7.

12. Mattson, S.N., K.E. Calarco, and A.R. Lang, *Focused and shifting attention in children with heavy prenatal alcohol exposure.* Neuropsychology, 2006. **20**(3): p. 361-9.

13. Jacobson, S.W., et al., *Impaired Eyeblink Conditioning in Children With Fetal Alcohol Syndrome.* Alcoholism: Clinical and Experimental Research, 2008. **32**(2): p. 365-372.

14. Connor, P.D., et al., *Effects of prenatal alcohol exposure on fine motor coordination and balance: A study of two adult samples.* Neuropsychologia, 2006. **44**(5): p. 744-51.

15. Rasmussen, C., *Executive Functioning and Working Memory in Fetal Alcohol Spectrum Disorder.* Alcoholism: Clinical and Experimental Research, 2005. **29**(8): p. 1359-1367.

16. Green, C.R., et al., *Deficits in Eye Movement Control in Children With Fetal Alcohol Spectrum Disorders.* Alcoholism: Clinical and Experimental Research, 2007. **31**(3): p. 500-511.

17. Streissguth, A.P., et al., *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, 1994. **18**(2): p. 248-254.

18. Kodituwakku, P.W., et al., *Neurobehavioral Characteristics of Children with Fetal Alcohol Spectrum Disorders in Communities from Italy: Preliminary Results.* Alcoholism: Clinical and Experimental Research, 2006. **30**(9): p. 1551-1561.

19. Iosub, S., et al., *Fetal alcohol syndrome revisited.* Pediatrics, 1981. **68**(4): p. 475-9.

20. Becker, M., G.A. Warr-Leeper, and H.A. Leeper, Jr., *Fetal alcohol syndrome: A description of oral motor, articulatory, short-term memory, grammatical, and semantic abilities.* Journal of Communication Disorders, 1990. **23**(2): p. 97-124.

21. McGee, C.L., et al., *Impaired language performance in young children with heavy prenatal alcohol exposure.* Neurotoxicology and Teratology, 2009. **31**(2): p. 71-5.

22. Greenbaum, R., et al., *The Toronto experience in diagnosing alcohol-related neurodevelopmental disorder: a unique profile of deficits and assets.* The Canadian Journal of Clinical Pharmacology, 2002. **9**(4): p. 215-25.

23. Sowell, E.R., et al., *Mapping callosal morphology and cognitive correlates: Effects of heavy prenatal alcohol exposure.* Neurology, 2001. **57**(2): p. 235-244.

24. Kvigne, V.L., et al., *Characteristics of children who have full or incomplete fetal alcohol syndrome.* The Journal of Pediatrics, 2004. **145**(5): p. 635-40.

25. Cone-Wesson, B., *Prenatal alcohol and cocaine exposure: Influences on cognition, speech, language, and hearing.* Journal of Communication Disorders, 2005. **38**(4): p. 279-302.

26. Burd, L., et al., *Recognition and management of fetal alcohol syndrome.* Neurotoxicology and Teratology, 2003. **25**(6): p. 681-688.

27. Church, M.W. and J.A. Kaltenbach, *Hearing, speech, language, and vestibular disorders in the fetal alcohol syndrome: A literature review.* Alcoholism, Clinical and Experimental Research, 1997. **21**(3): p. 495-512.

28. Church, M.W., et al., *Hearing, language, speech, vestibular, and dentofacial disorders in fetal alcohol syndrome.* Alcoholism, Clinical and Experimental Research, 1997. **21**(2): p. 227-37.

29. Kodituwakku, P.W., *Defining the behavioral phenotype in children with fetal alcohol spectrum disorders: A review.* Neuroscience and Biobehavioral Reviews, 2006.

30. Coggins, T.E., et al., *On becoming socially competent communicators: The challenge for children with fetal alcohol exposure.* International Review of Research in Mental Retardation, 2003. **27**: p. 121-150.

31. World Health Organization. *International Classification of Functioning, Disability, and Health (ICF): Introduction*. 2001 [cited 2008 March 13]; Available from: http://www.who.int/classifications/icf/site/intros/ICF-Eng-Intro.pdf.

32. Talmy, L., *A cognitive framework for narrative structure*, in *Toward a cognitive semantics*. 2000, MIT Press: Cambridge, MA. p. 417-482.

33. Croft, W., *Radical construction grammar: Syntactic theory in typological perspective.* 2001, Oxford: Oxford University Press.

34. Van Hoek, K., *Anaphora and conceptual structure*. 1997, Chicago: The University of Chicago Press.

35. Hikosaka, O. and M. Isoda, *Switching from automatic to controlled behavior: cortico-basal ganglia mechanisms.* Trends in Cognitive Sciences, 2010. **14**(4): p. 154-161.

36. B. J. Casey, N.T.J.F., *Clinical, imaging, lesion, and genetic approaches toward a model of cognitive control.* Developmental Psychobiology, 2002. **40**(3): p. 237-254.

37. Koch, K., et al., *Structure-function relationships in the context of reinforcement-related learning: a combined DTI-fMRI study.* Neuroscience, 2010. **In Press, Accepted Manuscript**.

38. Hikosaka, O. and M. Isoda, *Switching from automatic to controlled behavior: cortico-basal ganglia mechanisms.* Trends in Cognitive Sciences, 2010. **In Press, Corrected Proof**.

39. Casey, B.J., *Frontostriatal and Frontocerebellar Circuitry underlying Cognitive Control.* . in press, Sackler Institute for Developmental Psychobiology, Weill Medical College of Cornell University, retreived 10/18/2007 from http://www.csbmb.princeton.edu/conte/pdfs/project4/Proj4Pub28.pdf.

40. Poldrack, R.A. and P. Rodriguez, *How do memory systems interact? Evidence from human classification learning.* Neurobiology of Learning and Memory, 2004. **82**(3): p. 324-332.

41. Longworth, C.E., et al., *The basal ganglia and rule-governed language use: evidence from vascular and degenerative conditions.* Brain, 2005. **128**(Pt 3): p. 584-96.

42. Tricomi, E., et al., *Performance feedback drives caudate activation in a phonological learning task.* Journal of Cognitive Neuroscience, 2006. **18**(6): p. 1029-43.

43. Robles, S.G., et al., *The role of dominant striatum in language: a study using intraoperative electrical stimulations.* Journal of Neurology Neurosurgery and Psychiatry, 2005. **76**(7): p. 940-946.

44. Teichmann, M., et al., *The role of the striatum in rule application: the model of Huntington's disease at early stage.* Brain, 2005. **128**(5): p. 1155-1167.

45. Teichmann, M., et al., *The Role of the Striatum in Processing Language Rules: Evidence from Word Perception in Huntington's Disease.* Journal of Cognitive Neuroscience, 2006. **18**(9): p. 1555-1569.

46. Monchi, O., et al., *Functional role of the basal ganglia in the planning and execution of actions.* Annals of Neurology, 2006. **59**(2): p. 257-264.

47. Seger, C.A. and C.M. Cincotta, *Dynamics of Frontal, Striatal, and Hippocampal Systems during Rule Learning.* Cerebral Cortex, 2006. **16**(11): p. 1546-1555.

48. Seger, C.A., *The Basal Ganglia in Human Learning.* Neuroscientist, 2006. **12**(4): p. 285-290.

49. Cools, R., R.B. Ivry, and M. D'Esposito, *The Human Striatum is Necessary for Responding to Changes in Stimulus Relevance.* Journal of Cognitive Neuroscience, 2006. **18**(12): p. 1973-1983.

50. Alexander, M.P., *Impairments of procedures for implementing complex language are due to disruption of frontal attention processes.* Journal of the International Neuropsychological Society, 2006. **12**(02): p. 236-247.

51. Bitan, T., et al., *Weaker top-down modulation from the left inferior frontal gyrus in children.* NeuroImage, 2006. **33**(3): p. 991-998.

52. Bava, S., et al., *Longitudinal Characterization of White Matter Maturation During Adolescence.* Brain Research, 2010. **In Press, Accepted Manuscript**.

53. Blakemore, S.-J., *The social brain in adolescence.* Nat Rev Neurosci, 2008. **9**(4): p. 267-277.

54. Santesso, D.L. and S.J. Segalowitz, *Developmental Differences in Error-Related ERPs in Middle- to Late-Adolescent Males.* Developmental Psychology, 2008. **44**(1): p. 205-217.

55. Castellanos, F.X., et al., *Developmental Trajectories of Brain Volume Abnormalities in Children and Adolescents With Attention-Deficit/Hyperactivity Disorder.* JAMA, 2002. **288**(14): p. 1740-1748.

56. Conklin, H.M., et al., *Working memory performance in typically developing children and adolescents: behavioral evidence of protracted frontal lobe development.* Dev Neuropsychol, 2007. **31**(1): p. 103-28.

57. Ofen, N., et al., *Development of the declarative memory system in the human brain.* Nat Neurosci, 2007. **10**(9): p. 1198-1205.

58. Thomas, K.M., et al., *Evidence of Developmental Differences in Implicit Sequence Learning: An fMRI Study of Children and Adults.* J. Cogn. Neurosci., 2004. **16**(8): p. 1339-1351.

59. Astley, S.J., et al., *Magnetic resonance spectroscopy outcomes from a comprehensive magnetic resonance study of children with fetal alcohol spectrum disorders.* Magn Reson Imaging, 2009. **27**(6): p. 760-78.

60. Connor, P.D., et al., *Direct and indirect effects of prenatal alcohol damage on executive function.* Dev Neuropsychol, 2000. **18**(3): p. 331-54.

61. Astley, S.J., et al., *Neuropyschological and behavioral outcomes from a comprehensive magnetic resonance study of children with fetal alcohol spectrum disorders.* Canadian Journal of Clinical Pharmacology, 2009. **16**(1): p. e178-201.

62. Kodituwakku, P.W., W. Kalberg, and P.A. May, *The effects of prenatal alcohol exposure on executive functioning.* Alcoholism: Clinical and Experimental Research, 2001. **25**(3): p. 192-198.

63. Astley, S.J., *Fetal alcohol syndrome prevention in Washington State: evidence of success.* Paediatrics & Perinatal Epidemiology, 2004. **18**(5): p. 344-51.

64. Kodituwakku, P.W., et al., *Executive control functioning and theory of mind in children prenatally exposed to alcohol*. 1997, July: Breckenridge, CO.

65. Mattson, S.N., et al., *Executive functioning in children with heavy prenatal alcohol exposure.* Alcoholism, clinical and experimental research, 1999. **23**(11): p. 1808-15.

66. Niccols, A., *Fetal alcohol syndrome and the developing socio-emotional brain.* Brain and Cognition, 2007. **65**(1): p. 135-142.

67. Schonfeld, A.M., et al., *Executive functioning predicts social skills following prenatal alcohol exposure.* Child Neuropsychology, 2006. **12**(6): p. 439-52.

68. Coggins, T.E., G. Timler, and L. Olswang, *Double jeopardy without the daily double: Impact of prenatal alcohol exposure and maltreatment on the social communicative abilities of school-age children with Fetal Alcohol Spectrum Disorders.* Language, Speech, and Hearing Services in Schools, 2007. **38**(2).

69. Morton, J. and U. Frith, *Causal modelling a structural approach to developmental psychopatholgy*, in *Developmental psychopatholgy*, D. Cicchetti and D.J. Cohen, Editors. 1995, Wiley: New York.

70. Morton, J. and U. Frith, *Why we need cognition: Cause and developmental disorders*, in *Language, Brain, and Cognitive Development*, E. Dupoux, Editor. 2001, MIT Press.: Cambridge, MA. p. 263-278.

71. Gottesman, II and T.D. Gould, *The endophenotype concept in psychiatry: etymology and strategic intentions.* American Journal of Psychiatry, 2003. **160**(4): p. 636-45.

72. Viding, E. and S.-J. Blakemore, *Endophenotype Approach to Developmental Psychopathology: Implications for Autism Research.* Behavior Genetics, 2007. **37**(1): p. 51-60.

73. Gazzaniga, M.S., *Neuroscience and the correct level of explanation for understanding mind: An extraterrestrial roams through some neuroscience laboratories and concludes earthlings are not grasping how best to understand the mind-brain interface.* Trends in Cognitive Sciences, 2010. **14**(7): p. 291-292.

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

75. Justice, L.M., et al., *The Index of Narrative Microstructure: A Clinical Tool for Analyzing School-Age Children's Narrative Performances.* American Journal of Speech-Language Pathology, 2006. **15**(2): p. 177-191.

76. Strong, C.J., *The Strong Narrative Assessment Procedure*. 1998, Eau Claire, WI: Thinking Publications.

77. Liles, B.Z., *Narrative discourse in children with language disorders and children with normal language: a critical review of the literature.* Journal of Speech and Hearing Research, 1993. **36**(5): p. 868-82.

78. Halliday, M.A.K. and R. Hasan, *Cohesion in English*. 1976, London: Longman. xv, 374 - -.

79. Ariel, M., *Discourse, grammar, discourse.* Discourse Studies, 2009. **11**(1): p. 5-36.

80. Hickmann, M., *Children's discourse: person, space, and time across languages*. 2003, Cambridge, UK: Cambridge University Press.

81. Egner, T., *Prefrontal cortex and cognitive control: motivating functional hierarchies.* Nat Neurosci, 2009. **12**(7): p. 821-822.

82. Koechlin, E. and T. Jubault, *Broca's Area and the Hierarchical Organization of Human Behavior.* Neuron, 2006. **50**(6): p. 963-974.

83. Koechlin, E. and C. Summerfield, *An information theoretical approach to prefrontal executive function.* Trends in Cognitive Sciences, 2007. **11**(6): p. 229-235.

84. Kouneiher, F., S. Charron, and E. Koechlin, *Motivation and cognitive control in the human prefrontal cortex.* Nat Neurosci, 2009. **12**(7): p. 939-945.

85. O'Leary, C.M., *Fetal alcohol syndrome: diagnosis, epidemiology, and developmental outcomes.* Journal of Paediatrics and Child Health, 2004. **40**(1-2): p. 2-7.

86. Olson, H.C., et al., *Association of prenatal alcohol exposure with behavioral and learning problems in early adolescence.* Journal of the American Academy of Child and Adolescent Psychiatry, 1997. **36**(9): p. 1187-94.

87. Kelly, S.J., N. Day, and A.P. Streissguth, *Effects of prenatal alcohol exposure on social behavior in humans and other species.* Neurotoxicology and Teratology, 2000. **22**(2): p. 143-9.

88. Whaley, S.E., M.J. O'Connor, and B. Gunderson, *Comparison of the Adaptive Functioning of Children Prenatally Exposed to Alcohol to a Nonexposed Clinical Sample.* Alcoholism: Clinical and Experimental Research, 2001. **25**(7): p. 1018-1024.

89. Kodituwakku, P.W., *Defining the behavioral phenotype in children with fetal alcohol spectrum disorders: A review.* Neuroscience & Biobehavioral Reviews, 2007. **31**(2): p. 192-201.

90. Coggins, T.E., G.R. Timler, and L.B. Olswang, *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, 2007. **38**(2): p. 117-27.

91. Thorne, J.C., et al., *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, 2007. **50**(2): p. 459-474.

92. Thorne, J.C. and T.E. Coggins, *A diagnostically promising technique for tallying nominal reference errors in the narratives of school-aged children with Foetal Alcohol Spectrum Disorders (FASD).* International Journal of Language & Communinication Disorders, 2008. **43**(5): p. 570-594.

93. Hoffman, L.M., *The utility of school-age narrative microstructure indices: INMIS and the proportion of restricted utterances.* Language Speech & Hearing Services in Schools, 2009. **40**(4): p. 365-75.

94. Kraemer, H.C., *Evaluating medical tests: Objective and quantitative guidelines*. 1992, London: Sage Publications.

95. Johnston, J.R., *Narratives: Twenty-Five Years Later.* Topics in Language Disorders, 2008. **28**(2): p. 93-98.

96. Wong, A.M.Y. and J.R. Johnston, *The development of discourse referencing in Cantonese-speaking children.* Journal of Child Language, 2004. **31**(03): p. 633-660.

97. Ariel, M., *Accessibility Marking: Discourse Functions, Discourse Profiles, and Processing Cues.* Discourse Processes, 2004. **37**(2): p. 91 - 116.

98. Lillo-Martin, D. and R.M.d. Quadros, *Acquisition of the syntax-discourse interface: The expression of point of view*. Lingua, 2010. **In Press, Corrected Proof**.

99. Brown-Schmidt, S., *Partner-specific interpretation of maintained referential precedents during interactive dialog.* Journal of Memory and Language, 2009. **61**(2): p. 171-190.

100. Gundel, J.K., et al., *Testing predictions of the Givenness Hierarchy framework: A crosslinguistic investigation.* Journal of Pragmatics, In Press. **In Press, Corrected Proof**.

101. Seuren, P.A.M., *Language in Cognition: Language From Within Volume I*. 2009, Oxford: Oxford University Press.

102. Grassmann, S., M. Stracke, and M. Tomasello, *Two-year-olds exclude novel objects as potential referents of novel words based on pragmatics.* Cognition, 2009. **112**(3): p. 488-493.

103. Grosse, G., H. Moll, and M. Tomasello, *21-Month-olds understand the cooperative logic of requests.* Journal of Pragmatics, 2010. **In Press, Corrected Proof**.

104. Scott-Phillips, T.C., S. Kirby, and G.R.S. Ritchie, *Signalling signalhood and the emergence of communication.* Cognition, 2009. **113**(2): p. 226-233.

105. Maratsos, M.P., *The use of definite and indefinite reference in young children: An experimental study of semantic acquisition*. 1976, New York: Cambridge University Press. xiv, 144 p.

106. Rozendaal, M. and A. Baker, *The acquisition of reference: Pragmatic aspects and the influence of language input.* Journal of Pragmatics, In Press. **In Press, Corrected Proof**.

107. Ariel, M., *Interpreting Anaphoric Expressions: A Cognitive versus a Pragmatic Approach.* Journal of Linguistics, 1994. **30**(1): p. 3-42.

108. van Hout, A., K. Harrigan, and J. de Villiers, *Asymmetries in the acquisition of definite and indefinite NPs.* Lingua, 2010. **120**(8): p. 1973-1990.

109. Liles, B.Z., *Cohesion in the narratives of normal and language-disordered children.* Journal of Speech and Hearing Research, 1985. **28**(1): p. 123-33.

110. Thorne, J.C., *Tallying reference errors in narrative*. 2006, University of Washington, available http://students.washington.edu/jct6/TRAINwebManual.pdf.

111. Miller, J.F., *Systematic Analysis of Language Transcripts*. 2004, University of Wisconsin-Madison: Madison, WI.

112. Thorne, J.C. and T.E. Coggins. *Signals of CNS damage in the discourse behavior of school-aged children with prenatal alcohol exposure during a narrative generation task*. in *Symposium on Research in Child Language Disorders*. 2008. Madison, WI. Electronic version available at http://students.washington.edu/jct6/Webversion.ppt.

113. Thorne, J.C., et al., *A narrative analysis tool for distinguishing between school-age children with and without impairments of the central nervous system.* In revision.

114. Carmichael-Olson, H. and S.J. Astley, *[Intervening with children/adolescents with FAS/ARND.] Unpublished raw data*. 2005, University of Washington.

115. Coggins, T.E., *[Narratives produced by typically developing school-aged children]. Unpublished Raw Data*. 1995, University of Washington.

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

117. Pepe, M.S., *The statistical evaluation of medical tests for classification and prediction*. 2003, Oxford: Oxford University Press.

118. Rice, M. and G. Harris, *Comparing Effect Sizes in Follow-Up Studies: ROC Area, Cohen's d, and r*. Law and Human Behavior, 2005. **29**(5): p. 615-620.

119. Kraemer, H.C., et al., *Measures of Clinical Significance.* Journal of the American Academy of Child & Adolescent Psychiatry, 2003. **42**(12): p. 1524-1529.

120. Thorne, J.C., *The semantic elaboration coding system*. 2004, University of Washington, available at http://students.washington.edu/jct6/Handbook.htm.

121. Plante, E. and R. Vance, *Selection of Preschool Language Tests: A Data-Based Approach.* Lang Speech Hear Serv Sch, 1994. **25**(1): p. 15-24.

122. Svensson, L., *Variability in classroom social communication: Performance of children with fetal alcohol spectrum disorders and typically developing peers*, in *Department of Speech and Hearing Sciences*. 2006, University of Washington: Seattle, WA. p. 224.

123. Kaufman, A.S. and N.L. Kaufman, *Kaufman brief intelligence test*. 1990, American Guidance Service, Inc.: Circle Pines, MN.

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

125. DeLong, E.R., D.M. DeLong, and D.L. Clarke-Pearson, *Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.* Biometrics, 1988. **44**(3): p. 837-45.

126. Janes, H., G. Longton, and M. Pepe, *Accommodating Covariates in ROC Analysis.* Stata J, 2009. **9**(1): p. 17-39.

127. Astley, S.J., *Profile of the first 1,400 patients receiving diagnostic evaluations for fetal alcohol spectrum disorder at the Washington State Fetal Alcohol Syndrome Diagnostic & Prevention Network.* Canadian Journal of Clinical Pharmacology, 2010. **17**(1): p. e132-e164.

128. Kraemer, H.C., A. Noda, and R. O'Hara, *Categorical versus dimensional approaches to diagnosis: Methodological challenges.* Journal of Psychiatric Research, 2004. **38**(1): p. 17-25.

129. Gillberg, C., *The ESSENCE in child psychiatry: Early Symptomatic Syndromes Eliciting Neurodevelopmental Clinical Examinations.* Research in Developmental Disabilities, 2010. **In Press, Corrected Proof**.

130. Gresham, F.M. and S.N. Elliott, *Social Skills Rating System: Manual*. 1990, Circle Pines, MN: American Guidance Service.

131. Bishop, D., *Children's Communication Checklist-2 (British Edition)*. 2006, Upper Saddle River, NJ: Pearson.

Appendix:
On-line resources related to the TREIN

Tallying Reference Errors In Narratives (TREIN): training manual and details of the system as implemented in this research are available on-line at http://johncthorne.wordpress.com/tallying-reference-errors-in-narrative-trein/

Preliminary Studies.

Study 1: Thorne, Coggins, Carmichael Olson, & Astley [91] is available on-line at http://jslhr.asha.org/cgi/content/full/50/2/459

Study 2: Thorne & Coggins [92] is available on-line at http://depts.washington.edu/fasdpn/pdfs/thorne2008.pdf

Study 3: Details of Thorne & Coggins [112] are available on-line at http://johncthorne.files.wordpress.com/2010/05/thornecogginssrcld2008handout.pdf and the poster presented can be found at http://johncthorne.files.wordpress.com/2010/05/webversion.ppt

Study 4: Thorne, Coggins, Grittner, & Olswang [113]: Additional information related to study four are presented on the following pages.

Study 4: Additional Descriptive Information

Social behavior of the FASD group. As mentioned above, each child with FASD in the source study [122] had also been identified as having meaningful social problems. This was determined using the *Social Skills Rating System* [SSRS; 130], a norm-referenced test that uses teacher or parent report to rate social behaviors in children. Children were included in the original study based on a parent or teacher rating on the SSRS "Problem Behaviors" (SSRS-PB) subtest in the "clinical" or "borderline clinical" range (i.e. standard score >113). None of the children with FASD had a severe psychiatric diagnosis (e.g., schizophrenia). However, the caregivers of these children did report the following co-morbid conditions: 8 had attention deficits/hyperactivity, 9 had learning problems, 4 had speech/language problems, and 9 had trouble making friends.

Typically Developing Control Group

Typically developing controls were classmates of the FASD participants. In each case, the respective teachers were asked to choose a classroom peer who was "as close a cognitive match as possible" to the FASD participant [122]. Children in this group did not present with any academic concerns and caregivers did not *endorse* any of the following: attention deficits/hyperactivity, behavioral/emotional problems, learning problems, speech/language problems, or trouble making friends. Each TD participant was matched to an FASD participant based on gender and chronological age (mean difference $\pm$ 6.2 months, range $\pm$ 0-18 months). TD participants were excluded if they had a Composite IQ score more than -1.0 SD from the mean on the *Kaufman Brief Intelligence Test* [K-BIT; 123] and or receive scores on the SSRS-PB in the clinical range as rated by their teachers.

Additional Descriptive Statistics

All data were collected by an experienced and certified speech-language pathologist. Oral narratives were collected as a part of a longer testing protocol that was typically conducted across two days [122]. The oral narrative measure was the final measure collected and was recorded on audio cassette.

Standardized Testing. Testing conducted for the source study comprised a battery of standardized instruments of general cognitive and communication abilities—including: the *K-BIT, Children's Communication Checklist* [CCC-2; 131], and the *Test of Language Competence* [124]. The K-BIT and the TLC were administered to each participant in a quiet room at each child's school. Additional information on communicative performance was gathered using the CCC-2, a standardized parent screening report. For those scores where a group mean fell more than one-standard deviation from the normative mean, a T-test (p of 0.05) was performed to determine if group score means were significantly different between TD and FASD participant groups (see Table 4). Significant contrasts were found for the *Social Relations* scale of the CCC-2 (an expected result given that the inclusionary criteria for the FASD group included problem behaviors), for the TLC *Screening Composite* (mean of 83.90), and for the TLC *Recreating Speech Acts* subtest (TLC-RS, mean of 5.82). Although the TLC *Expressing Intents* group mean of 84 for the FASD group was more than one SD below the normative mean, it was not significantly different than the mean of 91.9 for the TD group.

FASD group test results. Table A.1 summarizes the SSRS-PB, K-BIT, TLC and CCC-2 scores for both groups. For the version of the CCC-2 used, a *General Communication Composite* score (GCC) below 55 suggests language impairment. Three of the children with FASD fell below this cut-off score. In addition, children with more than two scaled scores below the 10^{th} percentile score of 5 or any individual subscale below the 3rd percentile were considered at risk for language impairment. Using these CCC-2 criteria, 5 of the children in the FASD group would be considered at risk.

TD group test results. As seen in Table A.1, as per exclusionary criteria, all children in the TD group score within one standard deviation from the mean on the K-BIT composite IQ, and received SSRS-PB scores in the non-clinical range. For the TLC *Composite*, nine of the ten participants scored in the typical range (mean = 100; standard deviation = 15), while one child scored in the clinical range with a *Composite* score of 75 (1.67 SD below the normative mean). This child had a TLC *Expressing Intents Composite* of 73, an *Interpreting Intents Composite* of 82, an *Ambiguous Sentences* subtest score of 3 (2.3 SD below normative mean of 10), and a *Listening Comprehension* subtest score of 6. Three additional children from the TD group also received either domain composite scores or subtest scale scores more than one standard deviations below the normative mean: one for the *Screening Composite* (score 79); two for the *Expressing Intents Composite* (scores of 79, and 82); two for *Recreating Speech Acts* subtest (TLC-RS, scores of 4 and 5); and one for *Listening Comprehension* (score of 5). On the CCC-2, no children from the TD group received a GCC below 55, however, one child received a score at the cut-off of 55 with two scaled scores at or below 5 (*Speech* score of 4, *Appropriate Initiation* score of 5). This child received a TLC *Expressing Intents Composite* of 82 and a TLC-RS score of 4. An additional child received a score of 2 (1st percentile) on the CCC-2 *Speech* scale.

Table A.1: FASD and TD group data for *Social Skills Rating System* **(SSRS), K-BIT, TLC, and** *Children's Communication Checklist-2* **(CCC-2) scores (Preliminary Study 4).**

	Mean		Standard Deviation		Range (low|high)	
Group:	FASD	TD	FASD	TD	FASD	TD
SSRS						
Problem Behaviors	122.45	91.0	6.53	9.1	113|131	85|108
K-BIT						
Composite IQ (SS)	106.36	112.4	16.2	11.09	84|128	98|131
Vocabulary (SS)	101.91	106.30	14.9	8.6	78|118	98|126
Matrices (SS)	109.27	115.90	15.8	15.2	84|131	92|138
TLC						
Express Intents (SS)	[a]84.00	91.90	14.9	10.8	<65|112	73|106
Interpret Intents (SS)	97.36	103.00	15.5	11.8	<65|115	82|121
Screening Comp. (SS)	[b]83.90	96.10	11.67	10.3	<65|100	79|100
Language Comp. (SS)	89.6	96.8	15.6	11.7	<65|115	75|110
Recreating Speech (ss)	[b]5.82	[c]7.60	1.8	2.1	3|8	4|10
Ambiguous sent. (ss)	8.64	9.70	3.9	2.8	3|16	3 |13
Figurative Lang. (ss)	9.09	11.10	2.7	2.4	5| 13	8 |16
Listening Compr. (ss)	10.00	9.90	3.2	2.8	3 |13	5|13
CCC-2						
GCC	68.46	94.20	23.6	22.2	21 | 93	55|127
SIDC	-7.90	4.10	9.8	5.8	-24 |13	-2 |16
Coherence	9.64	11.80	4.2	2.3	2 |13	7 | 13
Inappropriate Initiation	7.73	12.40	3.0	3.8	4 | 14	5 | 16
Interests	8.46	13.30	3.3	2.7	4 | 13	8 | 16
Semantics	8.63	11.40	3.6	3.4	4 | 14	6 | 15
Social Relations	[d]4.46	10.70	3.8	2.4	0 |12	6 | 13
Speech	9.46	9.40	3.0	3.7	4 | 12	2 | 12
Stereotyped Language	9.82	11.90	3.0	1.9	5 | 13	8 | 13
Syntax	8.91	11.20	3.8	1.8	0 | 12	7 | 12
Use of Context	8.55	11.90	4.5	3.3	1 | 15	6 | 15

(SS) = Standard Scores: M of 100 (SD 15). (ss) = scaled scores M of 10 (SD 3). GCC = General Communication Composite. SIDC = Social Interaction Deviance Composite. [a]More than one SD below normative mean. [b]Significantly lower than TD mean (T-test, p =0.05). [c]Group mean more than one SD below normative mean. [d]Group mean below normative 10th percentile; significantly lower than TD group (T-test, p = 0.0003).

Nominal Reference Errors (rNRE): Table A.2 presents age and rNRE for the 21 newly added TD and FASD participants alongside performance from the 32 children in our initial research [91].

Table A.2: Age and rNRE for 32 original participants and 21 new participants (N=53; Preliminary Study 4).

Original 32 participants					21 newly added participants			
Age in months		rNRE			Age in months		rNRE	
TD	FASD	TD	FASD	Group	TD	FASD	TD	FASD
100	101	2.73%	2.12%		89	92	1.30%	3.82%
104	104	1.53%	2.61%		91	95	1.60%	5.00%
105	105	1.37%	3.896%		92	97	3.76%	1.15%
107	106	1.32%	2.79%		106	100	2.57%	3.09%
107	107	3.03%	3.902%		114	110	0.60%	0.30%
109	110	2.39%	4.38%		103	121	2.92%	0.88%
110	111	0.00%	3.86%		116	125	1.96%	4.45%
111	113	1.73%	2.64%		122	127	2.60%	1.50%
114	126	1.38%	1.10%		129	129	2.77%	4.62%
117	126	1.48%	3.72%		140	130	3.89%	12.36%
130	126	1.14%	3.31%		-	143	-	3.05%
131	128	0.88%	4.42%					
136	133	1.79%	5.73%					
137	134	0.57%	3.64%					
138	134	1.92%	2.11%					
139	137	1.87%	2.02%					
118	119	[a]1.57%	3.27%	Mean	111	115	[a]2.40%	3.66%
14	13	0.76%	1.16%	SD	17	17	1.05%	3.31%

rNRE = rate of nominal reference errors.
[a]Significantly different (T-test, p= 0.028).

Vita:

John C. Thorne was awarded a Bachelor of Arts in English and a Master of Arts in Curriculum & Instruction from New Mexico State University, as well as a Master of Science in Communication Disorders from the University of New Mexico. He has been an active speech-language pathologist certified by the American Speech-Language Hearing Association since 1999 and has been working with children with special needs for over 20 years. In 2010, he graduated with a Doctor of Philosophy in Speech and Hearing Sciences from the University of Washington. When he is not teaching or doing science John spends his time as a musician, composer, and recording artist.