Published in final edited form as:

Int J Lang Commun Disord. 2011 ; 46(6): 700–713. doi:10.1111/j.1460-6984.2011.00034.x.

Volubility as a Mediator in the Associations between Conversational Language Measures and Child Temperament

Laura Segebart DeThorne,

University of Illinois

Kirby Deater-Deckard,

Virginia Tech

Jamie Mahurin-Smith,

University of Illinois

Mary-Kelsey Coletto, and

University of Illinois

Stephen A. Petrill

Ohio State University

Abstract

Background—Despite support for the use of conversational language measures, concerns remain regarding the extent to which they may be confounded with aspects of child temperament, extraversion in particular.

Aims—This study of 161 twins from the Western Reserve Reading Project (WRRP) examined the associations between children's conversational language use and three key aspects of child temperament: Surgency (i.e., introversion/extraversion), Effortful Control (i.e., attention and task persistence), and Negative Affectivity (e.g., fear, anger, sadness). Child biological sex was considered as a possible moderating factor.

Methods & Procedures—Correlational analyses were conducted between aspects of temperament during early school-age years (i.e., 7 to 8 yrs), as measured by the Children's Behavior Questionnaire-Short Form (CBQ; Putnam & Rothbart, 2006), and six different measures of children's conversational language use: total number of complete and intelligible utterances (TCICU), number of total words (NTW), mean length of utterance (MLU), total number of conjunctions (TNC), number of different words (NDW), and measure D (i.e., a measure of lexical diversity). Values for NTW, TNC, and NDW were derived both on the entire sample and on the first 100 C-units. Correlations between language and temperament were compared between girls and boys using the Fisher *r*-to-*z* transformation to examine the significance of potential moderating effects.

Outcomes & Results—Children's reported variability in Effortful Control did not correlate significantly with any of the child language measures. In contrast, children's Negative Affectivity and Surgency tended to demonstrate positive, albeit modest, correlations with those conversational

language measures that were derived from the sample as a whole, rather than from a standardized number of utterances. MLU, as well as measures of NDW and NTW derived from standardized sample lengths of 100 C-units, did not correlate with any measure of child temperament. TNC demonstrated an unexpected negative correlation with child Surgency when it was derived from a standardized number of C-units but not when derived from the entire sample length. Child biological sex did not moderate the significant associations between language and temperament measures.

Conclusions & Implications—Overall, measures that control for volubility did not correlate significantly with child temperament; however, measures that reflected volubility tended to correlate weakly with some aspects of temperament, particularly Surgency. Results provide a degree of discriminant evidence for the validity of MLU and measures of type (i.e., NDW) and token use (i.e., NTW) when derived from a standardized number of utterances.

Introduction

Conversational language measures, first associated with the work of Brown in the 1970s, have become common tools in the study of child language. Despite being relatively time-consuming and cumbersome to derive, their inherent social validity is appealing. In addition, evidence of the psychometric properties for measures such as mean length of utterance (MLU) and number of different words (NDW) (Gavin & Giles, 1996; Heilmann, Miller, & Nockerts, 2010; Rice, Redmond, & Hoffman, 2006) rivals that of many standardized language tests (Plante & Vance, 1994; Mikucki & Larrivee, 2006). In spite of such support, concerns remain about the construct validity of conversational language measures, particularly in regard to the construct-irrelevant variance associated with child temperament. In particular, authors have suggested that conversational language measures may be confounded with child Surgency (i.e., degree of introversion/extraversion). Gregarious children, as opposed to shy ones, might receive higher values on measures of vocabulary and sentence complexity simply because they produce more words (cf. Hutchins, Brannick, Bryant, & Silliman, 2005 for a review of this position). As a result, numerous methods of controlling for differences in volubility (i.e., the amount of talk) have evolved, particularly in relation to vocabulary measurement.

Means to control for volubility in language sample measurement

To highlight the potential influence of volubility on child language sample measures we refer readers to Table 1, in which two brief fictional language samples are provided. In theory, both samples would have been collected by the same sampling procedure, and yet sample A includes 8 total utterances and sample B includes 5 total utterances. If one simply counts the total number of different words in the two samples, A exceeds B by 3 words. Although a difference of 3 words is minimal, note that differences could be substantially larger with longer samples. Consequently, one might question whether the higher value reflects stronger vocabulary skills or a general tendency to talk more. As a result, two general means of controlling for variance in volubility have developed: (a) utilizing standard sample sizes and (b) calculating ratios. Using the vocabulary diversity in samples A and B as an example, we will provide an overview of both methods.

The first method focuses on calculating measures such as the number of different words (NDW) from a standard sample size to 'level the playing field' in terms of amount of talk. This method can also facilitate comparison of achieved values to reference databases such as those provided by Systematic Analysis of Language Transcripts (Miller et al., 2005). Standard sample sizes can be established based on number of utterances or tokens (i.e., words). Using Samples A and B as examples, one might use the number of utterances in the smaller sample (i.e., five utterances) as the common denominator and consequently derive number of different words from the first five utterances in each sample, resulting in counts of 12 different words for Sample A and 19 for Sample B. With the standardization of sample length, the measure of vocabulary diversity now favors sample B by 7 words. A concern with standardizing based on number of utterances, however, is that utterances vary in terms of length (i.e., number of tokens), and thereby this strategy does not completely eliminate volubility as a confound (see Hutchins et al., 2005). Consequently, standard sample sizes can also be calculated based on the number of tokens. For example, if we used a standard sample cut of 21 tokens (i.e., the maximum available in Sample B), then the number of different words for Samples A and B would be 16 and 19 respectively. Whether samples are standardized by utterance or token number, the inherent variability across children's samples leads to the loss of data because sample sizes have to be truncated to the lowest common denominator. In our example, approximately one fourth to one third of Sample A was not included in the analysis when sample length was standardized.

A second method of controlling for volubility has focused on calculating vocabulary diversity as a function of sample size, leading to a ratio value. Compared to standardizing sample lengths, this method confers the advantage of utilizing all available data. Note that MLU utilizes a ratio method by dividing total morphemes by total number of utterances. Specific to vocabulary assessment, an early and intuitive version of the ratio method is exemplified by the type-token ratio (TTR), a ratio of word types over word tokens. Despite the pioneering nature of this effort, TTR did not prove to be a discriminating measure in terms of child language competence (Watkins, Kelly, Harbers, & Hollis, 1995). The relationship between types and tokens is not linear; the more words or utterances a child produces, the more certain words tend to be repeated. Note that even in the brief samples provided in the Appendix, high-frequency function words such as the pronouns 'he,' 'it' and 'I' begin to repeat themselves. Without such repetition, longer samples would become increasingly incoherent forms of 'word salad.' Consequently, children who produce more words in a sample are likely to produce fewer types relative to tokens, leading to reduced TTRs that may mask meaningful individual or group differences (see relevant discussion in Watkins et al., 1995, p.38).

Based on the relatively predictable relationship between TTR and number of tokens, McKee, Malvern, and Richards (2000) developed software to assess vocabulary diversity independent of sample size. In short, the numbers of word types across numerous subsamples are compared to a theoretical distribution of curves that is based on a model of the relation between TTR and number of tokens. The theoretical curve that best fits the curve derived from the real transcript provides a specific measure referred to succinctly as 'D'. In addition to being difficult to interpret, D has yielded smaller effect sizes than other more traditional measures in studies of children with language difficulties (e.g., Owen &

Leonard, 2002; DeThorne et al., 2008). Note that measure D has not been calculated from our fictitious examples in Table 1 because the program requires samples with at least 50 tokens (McKee, Malvern, & Richards, 2000)

Prior literature on child language and temperament

The rationale for controlling for volubility in the calculation of vocabulary diversity and other language sample measures is based largely on implicit assumptions regarding the relationship between child temperament and language use rather than direct empirical support. Evidence supporting the intuitive notion that children who are more extraverted actually talk more is relatively sparse. Most studies of child language and temperament have focused on word learning in infancy and early childhood. Work by Dixon and colleagues has found associations between child vocabulary acquisition and aspects of temperament, attention and task persistence in particular (Dixon & Salley, 2007; Dixon, Salley, & Clements, 2006; Dixon & Smith, 2000). As an example, a correlational design by Dixon and Smith (2000) examined longitudinal parent-reports of toddler temperament and early language development in 40 mother-child dyads. The authors found that aspects of children's temperament at 13 months significantly predicted vocabulary size and morphological inflections at 20 months of age. The most consistent relationships emerged in regard to child persistence, with coefficients ranging from .25 to .44 across language variables, and in regard to child adaptability, with coefficients ranging from .21 to .48. The temperament variables most similar to the construct of Surgency, activity and approachability, did not correlate significantly with any category of vocabulary acquisition.

A similar study by Morales et al. (2000), found that aspects of temperament at 6 months of age, specifically activity level, duration of orienting, and smiling and laughter, correlated positively with receptive vocabulary at one year of age, with coefficients ranging from .31 to .40. The interpretation from both studies (Dixon & Smith, 2000 and Morales et al, 2000) was limited by the possibility of 'common source bias' in which shared variance is due to the informant, in this case the parent, rather than similarities between developmental domains.

Stronger experimental evidence attributing differences in vocabulary learning to differences in temperament, attentional skills in particular, comes from studies of novel word learning in toddlers. Specifically, Dixon, Salley, and Clements (2006) evaluated the word learning capabilities of 39 toddlers under various conditions of environmental distraction. Child temperament, in particular attentional focus, was evaluated by parent questionnaire, and the sample was divided into two groups: low-focused and high-focused toddlers. In sum, the authors found that children's attentional focus moderated the influence of environmental distractions on word learning. Specifically, children who were high in attentional focus were less adversely affected by social distraction during novel word-learning tasks (see also Dixon & Salley, 2007).

Whereas most studies of language and child temperament have focused on word learning in younger children, a study by Slomkowski, Nelson, Dunn, and Plomin (1992) revealed associations between aspects of child temperament and language ability using direct evaluations of language competence through early school age. Aspects of temperament were

assessed at ages two and three years via parent report, while both receptive and expressive language skills were directly evaluated through standardized tests at 2, 3, and 7 years of age. The authors found that children's extent of affect-extraversion and task orientation both correlated positively with standardized assessments of child language at ages 2, 3, and 7 years. Effect sizes were modest, and when the contribution of prior language scores was factored out, the only associations that remained significant at age 7 were between affect-extraversion and receptive language measures. Although Slomkowski et al. eliminated the potential confound of common source bias affiliated with previous studies and encompassed children in our target range of early school age, the language measures were from standardized test scores rather than conversational language samples. Results from one form of assessment, such as standardized tests, do not always generalize to other forms of assessment, such as more naturalistic measures (cf. DeThorne & Watkins, 2006).

Although prior literature has demonstrated some associations between language and key aspects of temperament, no study to our knowledge has directly examined the relationship between temperament and conversational language measures. Consequently the purpose of the present study was to directly examine the potential influence of child temperament on measures of conversational language use. Specific questions were as follows:

1. To what extent can variance in conversational language use be attributed to differences in child temperament, specifically Effortful Control, Negative Affectivity, and Surgency?
2. Do observed relationships vary as a function of children's biological sex?

Methods

General Procedures

The present study utilized data from the Western Reserve Reading Project (WRRP), a longitudinal population-based twin study of reading, math, and associated cognitive abilities (Petrill et al., 2006). Twins were recruited primarily from Ohio, with the majority of families living in the metropolitan areas of Cleveland, Columbus, and Cincinnati. Given the longitudinal nature of the study, families were invited to participate in four home visits within the first three years of the project that began after the children had entered kindergarten but before they completed first grade. Each home visit included approximately two hours of individualized assessment, with each twin within a pair being assessed by a separate examiner. The bulk of the assessment focused on reading-related abilities (see DeThorne et al., 2006 and Petrill et al, 2006 for details). This particular paper focused on conversational language samples that were embedded within the larger assessment protocols of the second and third annual home visits (hereafter referred to as HV2 and HV3), as well as a caregiver questionnaire of child temperament administered between the second and third home visits.

Participants

The present study did not utilize twin methodology, but focused instead on the first-born twin within each pair in order to achieve an independent sample. Participants for the present

study were selected based on the presence of child temperament data and language sample data from the second and third annual home visits. The selection resulted in a total of 161 twins: 73 monozygotic (MZ, 41% male), 85 dizygotic (DZ, 41% male), and 3 with undetermined zygosity (33% male). The twins' mean age at their second home visit was 7.17 years ($SD = .68$), with a mean of 8.33 years ($SD = .75$) at their third home visit. The vast majority of caregivers were self-identified as White (91%), with the next largest group being Black at 3%. The remaining 6% classified themselves as Asian, Hispanic, or Other. The range of parental education achieved was distributed as follows: 1% no high school diploma, 9% completed a high school diploma or equivalent, 59% attended college, 27% received some graduate education, and 4% Other. Based on the 112 caregivers who completed a survey on their children's speech-language development, 22% of the children had received speech-language services at some point in their development, although only 8% were receiving services at the initiation of the study (see DeThorne et al., 2006 for additional details regarding the Speech-Language Survey).

Language samples—Examiners elicited a language sample from each child while the two of them played with modeling clay for a fifteen-minute period. Examiners were trained to follow general guidelines from Leadholm and Miller (1992) in the collection of conversational samples such as (a) limit closed-questions, (b) offer comments, (c) allow the child plenty of time to take a conversational turn, and (d) avoid overt correction (see DeThorne & Hart, 2009 for a full copy of elicitation guidelines). Topics of conversation included, but were not limited to, clay creations, school activities, holidays, sports, movies, and pets. The entire conversational exchange was recorded onto audiocassette or memory card and shared with the first author's laboratory for transcription by trained research assistants. During transcription, the samples were segmented into Communication units (C-units; see Loban, 1976; Nippold, 1998), which separate all independent clauses joined by coordinating conjunctions (i.e., *and*, *but*, *or*). This procedure helps systematize transcription and prevents inflation of MLU and other measures due to frequent use of coordinating conjunctions during this developmental stage. For example, if a child said "In the movie WALL-E, the Earth was polluted and WALL-E lived there alone," this pair of independent clauses would be segmented after the word 'polluted' to form two separate utterances, or communication units. In all other regards, the samples were coded according to Systematic Analysis of Language Transcripts conventions (SALT; Miller et al., 2005). Transcription reliability on 43 transcripts from the second home visit yielded a mean agreement of 90% ($SD = .05$) for C-unit boundaries and 91% ($SD = .04$) for individual morphemes. The 45 selected transcripts from the third home visit led to a mean agreement of 93% ($SD = .04$) for boundaries and 92% ($SD = .04$) for individual morphemes.

Conversational vocabulary measures—Six specific measures were derived from the conversational language samples in order to reflect (a) common assessment practices, (b) expressive skills in both vocabulary and morphosyntax, and (c) various means of controlling for volubility. All measures were derived either via SALT (Miller, 2004) or the Computerized Language Analysis suite of programs (CLAN; MacWhinney & Spekter, 2009).

- *Total complete and intelligible c-units (TCICU)*. TCICU represented the number of complete and intelligible C-units a child produced within the 15-minute conversational sample. Considered a direct measure of volubility or fluency, this measure has revealed group differences between school-age children with language disabilities and their same-age peers (e.g., Scott & Windsor, 2000).
- *Number of total words (NTW)*. NTW represented a frequency count of all word tokens produced within a sample. We derived NTW both from (a) all the complete and intelligible C-units within a sample, referred to hereafter as NTW-uncut, and (b) from the first 100 complete and intelligible C-units, referred to simply as NTW. Deriving NTW in both ways allowed us to more systematically review the impact of controlling for volubility on the potential associations between child language and temperament. Though often considered a measure of volubility or fluency (Miller et al., 2005; p.36), NTW tends to correlate significantly with measures of linguistic complexity, such as MLU and NDW (e.g., DeThorne et al., 2008).
- *Mean length of utterance (MLU)*. Mean length of utterance in morphemes, one of the earliest developmental measures of children's conversational language use, was calculated across all complete and intelligible C-units within a sample. The numerator was a sum of all morphemes divided by the total number of utterances, or in this case, C-units. Consistent with convention (Miller et al., 2005), words with inflectional morphemes (e.g., kicked) were considered as two morphemes, whereas words with derivational morphemes (e.g., quickly) were considered as one. Due to concerns with reliability of measurement, MLU values were not included for samples of fewer than 50 complete and intelligible C-units. Though commonly referred to as a measure of grammatical complexity, MLU correlates strongly with a variety of other conversational measures, including vocabulary diversity (DeThorne, Johnson, & Loeb, 2005).
- *Total number of conjunctions (TNC)*. Total Number of Conjunctions was derived as a frequency count both (a) within the entire sample (i.e., TNC-uncut) and (b) across the first 100 complete and intelligible C-units within each sample (i.e., TNC). Twelve types of coordinating and subordinating conjunctions were included: *after, and, as, because, but, if, or, since, so, then, until, and while*. TNC was intended as an explicit correlate of grammatical complexity, although our previous work has found that it loads on the same factor as NTW, MLU, and NDW (DeThorne et al., 2008).
- *Number of different words (NDW)*. NDW has been widely used as a measure of semantic diversity or productive vocabulary size based on the total number of different root words (e.g. Miller et al., 2005; Ukrainetz & Blomquist, 2002; Watkins et al., 1995). In the present study, we calculated NDW both from (a) the entire sample (i.e., NDW-uncut), and (b) the first 100 complete and intelligible utterances (i.e., NDW). Validity evidence for NDW comes from documented developmental change during the school-age years (e.g., Miller et al, 2005), correlation with standardized vocabulary measures (e.g., Ukrainetz & Blomquist, 2002), and differentiation of child language ability (e.g., Watkins et al., 1995).

- *Measure D.* We selected measure D as an additional assessment of lexical diversity due to its unique means of controlling for volubility. It was derived using a suite of programs available through the Child Language Data Exchange System (<http://childes.psy.cmu.edu>). Our process included four steps. First, the transcripts were converted from SALT to a format compatible with the Computerized Language Analysis suite of programs (CLAN) using the SALTIN utility. Second, the CHECK command was used to find formatting errors within the converted transcripts. In the third step, abandoned and interrupted utterances were marked with the CHSTRING utility, so that they would be excluded from measure D calculations. Finally, VOCED software was used to derive measure D on the entire sample. The +r6 and -s*^%% options were employed to exclude mazed utterances and morpheme variations, respectively. Emergent validity evidence for measure D includes developmental change with age, correlation with other measures of expressive vocabulary, and significant group differences (Durán et al., 2004; Klee, Stokes, Wong, Fletcher, & Gavin, 2004; Owen & Leonard, 2002).

Temperament questionnaire—Information regarding child temperament was obtained through the Children’s Behavior Questionnaire-Short Form or CBQ-SF (Putnam & Rothbart, 2006). Caregivers indicated for each twin separately the accuracy of a series of statements using a 7-point scale ranging from “extremely true” to “extremely *untrue*.” The CBQ-SF measures Surgency (facets: impulsivity, high-intensity pleasure, activity level, shyness), Negative Affectivity (facets: fear, anger, sadness, discomfort, soothability), and Effortful Control (facets: attentional focusing, inhibitory control, low-intensity pleasure, perceptual sensitivity). The individual facets of shyness and soothability were reverse scored for inclusion in the factor composites, so that higher values represented higher Surgency and Negative Affectivity respectively. In previous analyses, we confirmed the measurement structure of the CBQ-SF and found the reliability (i.e., alpha coefficients) to be satisfactory (Mullineaux, Deater-Deckard, Petrill, Thompson, & DeThorne, 2009). Despite concerns regarding bias in the use of parent reporting, such measures are able to reflect children’s behaviors across time and context, thereby offering a fuller view of development than a single observational or standardized assessment.

Analyses

Study questions were addressed primarily through bivariate Pearson correlations across language sample and temperament measures. The question regarding child biological sex as a potential moderator was examined by comparing correlation coefficients between girls and boys using the Fisher *r*-to-*z* transformation. Throughout all analyses, alpha was set at .01 to reduce Type I error.

Findings

Descriptives

Descriptive data for all individual language sample measures at both home visits are provided in Table 2. Note the differences in number of cases across variables, which was due to differences in the sample length required for calculation. For example, TCICU, D,

and measures from ‘uncut’ transcripts were derived from all available samples. In contrast, MLU was not included for samples with fewer than 50 utterances, and the remaining measures (i.e., NTW, TNC, & NDW) required a minimum of 100 complete and intelligible utterances. Note also from Table 2 that all language measures, except measure D, showed a relatively slight but consistent increase between the second and third home visits, a shift that is generally in line with developmental expectations. In addition, measures from uncut transcripts (i.e., NTW-uncut, TNC-uncut, and NDW-uncut) were higher than the same measures derived from the first 100 C-units within each sample, which also makes sense given that the average sample length was over 100 utterances at both home visits.

The temperament scores were widely and normally distributed, with the means for all factors near “4” (the center of the 7-point Likert scale on the CBQ-SF) and the full range of possible scores represented for most of the facets. Means ranged from 5.00 to 5.70 (*SDs* from .56 to .88) for Effortful Control and its individual facets, from 3.52 to 4.67 (*SDs* from .70 to 1.20) for Surgency/Extraversion and its facets, and from 3.58 to 4.25 (*SDs* from .66 to 1.10) for Negative Affectivity and its facets. As reported elsewhere (Mullineaux et al., 2009), the correlations between the three temperament factors on the CBQ-SF were generally modest in magnitude.

Correlations across Language Measures

Given the overlapping constructs and procedures for the language sample measures, collinearity was expected and examined via a correlation matrix. Significance was based on an alpha of .01 in all cases. Correlations across NTW, MLU, TNC, and NDW (including cut and uncut versions) at the second home visit were all statistically significant and of medium to large effect sizes, ranging from .39 to .97. Similarly for data from the third home visit, correlations for the same measures ranged from .54 to .95.

Measure D demonstrated less robust associations with the other conversational language measures. Specifically at HV2, D correlated significantly with NDW ($r=.29$), NDW-uncut ($r=.48$), NTW-uncut ($r=.29$) and TCICU ($r=.31$). At HV3, D continued to correlate significantly with the same variables at small to medium effect sizes, and also gained a significant correlation with MLU ($r = .22$). In addition to its correlation with D, TCICU correlated with NTW-uncut, NDW-uncut, and TNC-uncut at both home visits with medium to large effect sizes ($r = .55$ to $.88$), thereby indicating a strong tendency for frequency variables from uncut samples to ‘hang together.’ In addition, the correlation between TCICU and MLU was significant at HV3 with a coefficient of .31. It makes sense that TCICU would not correlate significantly with NTW, NDW, and TNC since the number of C-units was controlled (i.e., 100 C-units) in the calculation of those three measures.

Stability of Language Measures across Home Visits

The stability in language sample measures between HV2 and HV3 was evaluated through bivariate correlations. Associations were statistically significant for all measures ($p < .01$) with coefficients ranging from .47 for D to .59 for TNC. Based on the stability of the values, each language sample measure was averaged across the home visits to address the primary research questions. Averaging the language measures across home visits to form the

dependent variables also made sense due to the timing of the temperament questionnaires which were completed between home visits two and three. In sum, each of the following nine language measures were averaged across HV2 and HV3 in order to examine their individual associations with child temperament: TCICU, NTW, NTW-uncut, MLU, TNC, TNC-uncut, NDW, NDW-uncut, & D. Descriptives for the averaged measures are provided in Table 2. Note that the number of cases for the averaged values exceeded the number of cases for either of the individual home visits for certain measures (i.e., NTW, MLU, TNC, and NDW). This is due to the fact that MLU, NDW, NTW, and TNC required sample sizes of at least 100 utterances. Some children produced a 100-utterance sample during one home visit but not the other. In such cases, the ‘averaged’ variable consisted of the value from the one completed home visit to maximize power.

Associations with age and sex

To understand the potential confounding effects of child age on the variables of interest, a correlation matrix involving child age at HV2, the averaged language measures, and temperament measures was generated. Five of the nine averaged language measures demonstrated significant positive associations with child age, generally of small effect size: NTW ($r=.24$), MLU ($r=.28$), TNC ($r=.30$), TNC-uncut ($r=.27$), and NDW ($r=.27$). Consistent with developmental expectations, such associations indicated increases in linguistic complexity with increasing age.

In terms of the temperament measures, only one measure revealed a significant association with age: attention focusing, a facet of Effortful Control, with a coefficient of $-.25$. Because a scatter plot revealed two bivariate outliers, specifically two older children with very low scores in attention focusing (below 2.5 on the 7-point scale), the correlation between child age and attention focusing was derived without these two cases. Without the two bivariate outliers, the correlation between child age and attention focusing was no longer statistically significant ($r = -.19$).

Mean differences in the language and temperament variables based on children’s biological sex were evaluated via independent t -tests with alpha set at .01. In sum, no group differences emerged for the nine language measures. Two temperament facets were significantly lower for boys in comparison to girls. Specifically, boys scored lower on perceptual sensitivity ($t=3.03$), a facet of Effortful Control, and lower on sadness ($t=2.64$), a facet of Negative Affectivity. Given the relatively limited associations of child age and biological sex with the variables of interest, we decided to control for such influences on a case-by-case basis if needed rather than factoring out the effects of age and sex across all dependent variables.

Correlations across language and temperament

To focus explicitly on the primary research question regarding the extent to which variance in conversational language measures can be attributed to differences in child temperament, a correlation matrix was derived for each temperament factor and its various facets with the nine averaged conversational language measures: TCICU, NTW, NTW-uncut, MLU, TNC, TNC-uncut, NDW, NDW-uncut, and D. As in the previous correlation analyses, alpha was

set at .01 to reduce the likelihood of Type I error given the large number of associations examined.

Effortful Control and its facets—Effortful Control and its individual facets (i.e., attention focusing, inhibitory control, intensity pleasure, and perceptual sensitivity) did not correlate significantly with any language measure. The individual coefficients ranged from $-.15$ to $.17$.

Negative Affectivity and its facets—In regard to Negative Affectivity, three language measures demonstrated significant correlations with the factor as a whole and/or the individual facet of sadness. Specifically, both TCICU and NTW-uncut demonstrated positive associations with the Negative Affectivity factor as a whole ($r=.26$ and $.21$ respectively) and with the individual facet of sadness ($r=.27$ and $.23$ respectively). Similarly, NDW-uncut correlated significantly with sadness ($r=.23$). Although small in effect size, the positive correlations indicated that children who were reported to display more negative affective, sadness in particular, tended to produce a higher frequency of word tokens, word types, and total C-units in their samples. When volubility was explicitly controlled, whether through ratio method, as in the case of MLU and D, or standardized transcript cuts, in the case of NTW and NDW, associations with Negative Affectivity and its facets were not significant.

Surgency/Extraversion and its facets—Of the three temperament factors, Surgency correlated with the largest number of conversational language measures; however, effect sizes were mostly small (see Table 3). TCICU correlated with Surgency ($r=.34$) as well as with two of its individual facets: high intensity pleasure and shyness. The negative correlation between TCICU and shyness is actually consistent with the positive correlation between TCICU and the Surgency factor given that shyness was reverse scored when entered into the Surgency composite. In essence, children higher in Surgency, especially children lower in shyness and more prone toward high intensity pleasure, tended to produce more complete and intelligible utterances within their conversational samples. Similarly, measure D correlated positively with the Surgency factor and the individual facet of high intensity pleasure. Of particular interest, NTW-uncut and NDW-uncut correlated positively with Surgency and the individual facet of shyness but not when the same measures were derived on only the first 100 C-units of each transcript. Unexpectedly, TNC demonstrated a significant association with Surgency, but in a negative direction, both with the overall factor ($r=-.22$) and with the facet of high intensity pleasure ($r=-.27$). This negative association indicated that children with higher Surgency, particularly those more prone to high intensity pleasure, tended to have fewer conjunctions within their samples of 100 C-units. When TNC-uncut was derived from the entire sample, no significant association emerged with child temperament.

Sex as potential moderator

In order to address question two regarding sex as a potential moderator in the relation between language and temperament, the same correlation matrices between temperament and language variables were regenerated separately for girls and boys and compared via

Fisher r-to-z transformation in order to determine statistically significant differences (see Tables 4 and 5). The only significant difference to emerge related to the relationship between attention focusing and measure D ($z=-2.85$, $p<.01$), with girls demonstrating a negative correlation of $-.21$ and boys showing a correlation of $.25$. Neither correlation emerged as significant at $.01$ when considered individually; the two-tailed p value in both cases was $.04$.

Discussion

The present study examined the potential relationship between conversational language measures and child temperament for the primary purpose of understanding the extent to which the latter may confound the former during assessment practices. To this end, the primary finding was that measures explicitly controlling for differences in volubility, either through averaging across C-units (i.e., MLU) or through a standardized number of C-units (e.g., NDW, NTW) did not correlate significantly with child temperament measures. In contrast, frequency measures, such as TCICU, NTW-uncut, and NDW-uncut, that were derived from entire samples tended to demonstrate modest but significant associations with aspects of child Surgency and Negative Affectivity. More specifically, higher Surgency and higher Negative Affectivity were associated with higher values on language measures, accounting for between 7 to 12% of total variance. Two noteworthy exceptions emerged. First, measure D, developed with the explicit intent to control for volubility via an elaboration of the ratio method (McKee, Malvern, & Richards, 2000), demonstrated a modest but significant positive association with child Surgency. Second, TNC derived from 100 C-units demonstrated an unexpected *negative* correlation with child Surgency, which was modest, but statistically significant. The only evidence that child biological sex moderated any relationship between temperament and conversational language was a statistically different correlation between attention focusing and D, but the individual correlation coefficients for both girls and boys were of small effect size and not statistically different from zero. The ensuing discussion will focus on integrating our findings with prior research, considering potential limitations in generalizability, and highlighting implications for child language assessment.

Integrating findings with prior research

Our findings extend prior work in three key ways. First, few studies have examined the relationship between temperament and language in school-age children. The longitudinal study by Slomkowski et al. (1992) found that aspects of child temperament at age two, specifically affect-extraversion and task orientation, correlated with standardized assessments of child language at ages 2, 3, and 7 years. Similar to our findings of associations between Surgency and child language measures, the effect sizes were relatively modest. When the contribution of prior language scores was factored out in Slomkowski et al. (1992), the only associations that remained significant at age 7 were between affect-extraversion and receptive language measures. Similarly in the current study, Surgency emerged as the temperament factor most consistently tied to conversational language measures during the school-age years, but the effect sizes were relatively modest.

Besides age, another relatively unique aspect of our study was the use of direct language assessment, conversational measures in particular. Many prior studies of language and temperament have relied exclusively on parent report measures. Consequently the possibility of a ‘common source bias’ clouded the interpretation of positive associations between child language and temperament in past studies (e.g., Dixon & Smith, 2000; Morales et al., 2000). The present study paired a parent-report measure of temperament with direct examination of children’s language use within a conversational context. Direct assessment of children’s language use is a critical distinction, not only in regard to consideration of a common source bias, but also in regard to what aspects of child temperament are most likely to be engaged. Whereas conversational exchanges may draw upon aspects of child Extraversion/Surgency, they may be less taxing in regard to the need for focused attention and executive control. This distinction may help explain why the current study failed to replicate previously-found associations between vocabulary learning and aspects of Effortful Control. Previous studies that have reported such associations have either been confounded by potential common source bias (e.g., Dixon & Smith, 2000; Morales et al., 2000) or utilized assessment contexts that are more likely than conversation to tax abilities to attend, such as novel word learning tasks (e.g., Dixon & Salley, 2007; Dixon et al., 2006) and standardized tests (e.g., Slomkowski et al., 1992). Of interest, our finding of statistically different correlations between attention focusing and measure D in girls versus boys suggest that future study of children’s language and attention skills may want to examine the possibility of biological sex as a moderating influence.

The important distinctions across assessment methods also highlight a third key aspect of the current study: the inclusion of multiple measures of child language. We included nine language measures in our analyses that strategically differed in the means and extent to which they controlled for volubility. Specifically, our standard version of NTW, TNC, and NDW were derived from transcript cuts of 100 C-units. This method of controlling for volubility has been critiqued as being insufficient due to the fact that single utterances, or C-units, can vary substantially in terms of length (see Hutchins et al., 2005). It was interesting to find that of the variables derived on standardized transcript cuts, only TNC correlated significantly with Surgency, and the relationship was in the unanticipated direction. Specifically, children with greater Surgency actually produced fewer total conjunctions in a 100 C-unit sample. This unanticipated negative association is difficult to interpret. One possibility, though speculative, is that children with higher versus lower Surgency tend to use conjunctions in different ways. Since the negative association emerged only for TNC on 100 C-units and not on TNC-uncut, it might reflect the tendency of children higher in Surgency to use a higher proportion of conjoining conjunctions (specifically *and*, *but*, and *or* to link independent clauses) than children lower in Surgency. Said another way, low-surgency kids might be using relatively more intra-clause conjunctions (e.g., “For Christmas I want this *and* this *but* not that.”) compared to high-surgency children, thereby leading to more counted conjunctions per utterance. Because C-units are segmented based on the use of conjunctions that coordinate independent clauses, then children who used more of such conjunctions would have depressed TNC values when calculated on samples of 100 C-units. Although intriguing, this interpretation is speculative and was not anticipated or verified in

any way, it is possible that the negative association represents a spurious finding that is specific to our sample.

Regardless of how we interpret the negative association between TNC and Surgency, the results consistently indicated that NTW and NDW did not correlate significantly with any measured aspect of child temperament when calculated on 100-utterance cuts. In contrast, measure D was designed explicitly to evaluate vocabulary diversity while controlling for differences in the number of words produced within the samples (McKee, Malvern, & Richards, 2000), yet D correlated significantly with child Surgency, accounting for approximately 6% of the variance. This finding suggests that D does not control for child temperament to any greater extent than NTW and NDW in particular. In fact, D demonstrated associations with Surgency that seemed relatively comparable to the associations observed for NDW-uncut (see Table 3).

Finally, frequency count measures such as TCICU, NTW-uncut, and NDW-uncut demonstrated modest but significant positive associations with child Surgency and Negative Affectivity. Such measures did not include any direct control for volubility other than the standardized sample time. Examiners were instructed to maintain the conversational interaction for fifteen minutes, although the children were not explicitly required to talk during this period. In fact, examiners were encouraged to tolerate sizeable lulls in the conversation (up to 30 seconds at a time) and to provide comments in addition to direct questions. Consequently, the measures of TCICU, NTW-uncut, and NDW-uncut theoretically provided the greatest opportunity for differences in child talkativeness to emerge. As expected, these variables correlated the most consistently with Surgency, although the strength of association left 78–96% of unexplained variance. Of interest, TNC-uncut did not correlate significantly with Surgency, perhaps for the same reasons previously discussed in relation to the negative association between TNC and Surgency.

Unlike Surgency, the positive associations of TCICU, NTW-uncut, and NDW-uncut with Negative Affectivity or its facets was not expected. At face value, it suggests that children reported to express more negative affect, sadness in particular, are more likely to produce higher frequencies of C-units, total words, and different words in their fifteen-minute samples. It is possible that children who experience more negative feelings tend to talk more, or that parents of children who talk more tend to be more aware of their children's negative feelings. Given that the association does not represent an a priori prediction, we recommend awaiting replication in a separate sample before attaching too much weight to this particular finding.

In sum, similarities and differences across studies of child language and temperament could be attributed to such factors as child age, assessment method, and the specific measures utilized.

Considering potential limitations

Related to differences between our findings and prior studies, it is prudent to note that findings from the present study might differ substantially within other populations or with the use of different measures. In regard to select populations, it seems plausible to

hypothesize that language expression and aspects of temperament may be more difficult to distinguish during early stages of development (e.g., Bloom, 1993) or in children with frank deficits in either domain. For example, there is a substantial literature documenting the increased risk of challenging behaviors and impaired social interactions in children with speech-language disabilities, thereby suggesting there may be an inherent link between aspects of temperament and language in select populations. However, the nature of this link may be difficult to disentangle given the key role of communication skills in mediating social interaction (cf. Redmond & Rice, 1998).

In addition to subgroup differences, varied measures or assessment methods would likely lead to differing relations both within and across individual developmental domains, such as language and temperament (cf. Kagan, 2001). As mentioned previously, language tasks that require more focused attention are more likely to reveal associations with aspects of Effortful Control. Similarly, measures from a conversational context that provide little to no control for volubility, such as TCICU, are most likely to correlate with aspects of Surgency. It is possible that correlations in the present study underestimated the potential associations between talkativeness (as evaluated via TCICU, NTW-uncut, and NDW-uncut) and temperament due to the equated length of the children's conversational samples. In other words, examiners explicitly tried to maintain the conversational interaction for 15 minutes through continued prompts (i.e., comments and questions), thereby eliciting more utterances from children who might have chosen to end the exchange earlier if given a more naturalistic scenario. In fact, studies of narrative samples, in which children are free to tell as long or as short of a story as they like, elicit extremely short samples in some children (e.g., Miller, et al., 2006; Scott & Windsor, 2000). In other words, all our measures included some element of control for volubility. Consequently, talkativeness may well be associated with child temperament; however, that question was not the explicit focus of this study.

Another means in which conversational measures differ both from narrative samples and from standardized assessments is the relatively unconstrained role of the conversational partner. Although examiners were given general guidelines for the interaction, conversational interactions by nature are largely unscripted and require flexibility between both partners. As such, the examiner may play a significant role in shaping the interaction, including the frequency and complexity of a child's productions. Although it difficult to imagine any systematic bias this issue might have raised in the present study, such uncontrolled variability in the data could serve to mask meaningful associations or group differences. For a more in-depth discussion of the transactional nature of conversational interactions, readers are referred to DeThorne and Hart (2009).

Highlighting implications

The discussion of what factors influence language sample measures brings us to the original motivation for the study, which was to better understand the potential influence of child temperament on commonly-used conversational language measures. In short, findings suggest that measures of MLU, NTW, and NDW that control for volubility through standardizing number of utterances (or via ratio as in the case of MLU) do not correlate significantly with child temperament. Although correlational data are limited by their

inability to reveal causal factors, in this case, the lack of significant correlations provides a form of discriminant evidence that supports the validity of these language sample measures. Discriminant evidence of validity, also known as divergent evidence, demonstrates the specificity of measurement to the construct of interest, thereby helping address concerns regarding construct-irrelevant variance (cf. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; see also Plante & Vance, 1994). Together with prior evidence of construct validity (e.g., Heilmann et al., 2010; Klee et al., 1989; Owen & Leonard, 2002; Rice, Redmond, & Hoffman, 2006; Watkins et al., 2001), findings here support the use of NTW, MLU, and NDW as valid measures of children's conversational language use. It is also worth noting that the positive correlation between child age and all three of these conversational language measures provided additional evidence of construct validity given the expected development of language with age (see also Miller et al., 2005; Rice, Redmond, & Hoffman, 2006). Of interest, measure D, a vocabulary measure developed to minimize the influence of volubility, demonstrated small but significant correlations with some aspects of Surgency and did not correlate significantly with child age. Such findings do not conclusively rule out measure D as a valid measure of productive vocabulary. However when taken in conjunction with findings for the other measures, results suggest that measure D does not demonstrate inherent superiority to more traditional measures of vocabulary, such as NDW calculated on a standard number of utterances.

Despite the significant associations that emerged between child temperament and certain language measures, it is important to highlight that a single temperament measure never accounted for more than 12% of variance in a single conversational language measure, indicating that other factors accounted for the majority of the variance, i.e., the remaining 88% or more. Such factors could include the child's language proficiency (as one would hope), transient mood, or even qualities inherent to the conversational partner (e.g., disposition, familiarity, etc.). Certainly additional forms of validity evidence are needed to support the use of any of these measures for the intended assessment purposes (e.g., Heilmann et al. 2010).

Although the present study provides discriminant evidence for the validity of child language sample measures calculated on a standardized number of C-units, it does not directly address the question of whether or not examiners *should* control for differences in child temperament and volubility when deriving language sample measures. It is not unreasonable to think that volubility may represent a meaningful part of the construct of interest. For example, the more proficient a child's language skills, the more eager he/she may be to employ them. Perhaps by controlling for volubility we are actually removing a significant portion of meaningful variance in child language use. This is not to say that 'being a child of few words' is inconsistent with developing strong language skills. It is only to say that this pattern may be the exception rather than the general rule. If it is true that children with stronger language skills are likely to talk more than their peers with weaker language skills when given the opportunity, then controlling for volubility would actually remove variance associated with language competency. This interpretation is consistent with findings that group differences and other experimental effect sizes seem to diminish with the use of

language measures that control for volubility (e.g., DeThorne et al., 2008; Owen & Leonard, 2002). It would be beneficial for future work to directly examine the sensitivity and specificity of language sample measures, both before and after controlling for volubility (cf., Owen and Leonard, 2002), in order to help guide clinicians and investigators in designing their assessments.

In closing, currently-used language sample measures demonstrated relatively limited associations with child temperament, particularly MLU and measures of word use based on the first 100 C-units of a sample (i.e., NTW and NDW). However, the question regarding whether or not examiners should control for volubility when calculating language sample measures is open for debate and remains an important question to address in developing evidence-based practices for child language assessment.

Acknowledgements

The Western Reserve Reading Project is supported by NICHD (HD38075, HD46167, HD050307). In addition, transcription and analyses have been supported by the American Speech-Language-Hearing Foundation New Investigator Award, the UIUC Campus Research Board, and the Children Youth and Families Consortium at the Pennsylvania State University. Interdisciplinary collaborations have been enhanced by the American Speech-Language-Hearing Association Advancing Academic-Research Careers (AARC) Award. In addition, we sincerely appreciate the time and effort of all participating families and affiliated research staff. Special thanks to Amanda Austin for research assistance and to the following individuals for formative recommendations: Bonnie Johnson, Amanda Owen, Kerry Proctor-Williams, and Sean Redmond.

References

- American Educational Research Association, American Psychological Association, and National Council On Measurement In Education. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association; 1999.
- Bloom, L. The Transition from Infancy to Language: Acquiring the Power of Expression. New York: Cambridge University; 1993.
- Dethorne LS, Hart SA. Use of the Twin Design to Examine Evocative Gene-Environment Effects within a Conversational Context. 2009; 3:175–194.
- Dethorne LS, Hart SA, Petrill SA, Deater-Deckard K, Thompson LA, Schnatschneider C, Davison MD. Children's history of speech-language difficulties: Genetic influences and associations with reading-related measures. *Journal of Speech, Language, and Hearing Research*. 2006; 49:1280–1293.
- Dethorne LS, Johnson BW, Loeb JW. A closer look at MLU: What does it really measure? *Clinical Linguistics and Phonetics*. 2005; 19:635–648. [PubMed: 16147407]
- Dethorne LS, Petrill SA, Channell RW, Hart SA, Campbell RJ, Deater-Deckard K, Thompson LA, Vandenberg DJ. Genetic effects on children's conversational language use. *Journal of Speech-Language-Hearing Research*. 2008; 51:423–435.
- Dethorne LS, Watkins RV. Language abilities and nonverbal IQ in children with language impairment: Inconsistency across measures. *Clinical Linguistics and Phonetics*. 2006; 20:641–658. [PubMed: 17342874]
- Dixon WE, Salley BJ. "Shhh! We're trying to concentrate": Attention and environmental distracters in novel word learning. *The Journal of Genetic Psychology*. 2007; 167:393–414. [PubMed: 17645230]
- Dixon WE, Salley BJ, Clements AD. Temperament, distraction, and learning in toddlerhood. *Infant Behavior & Development*. 2006; 29:342–357. [PubMed: 17138290]
- Dixon WE, Smith PH. Links between early temperament and language acquisition. *Merrill-Palmer Quarterly*. 2000; 46:417–440.
- Durán P, Malvern D, Richards B, Chipere N. Developmental trends in lexical diversity. *Applied Linguistics*. 2004; 25:220–242.

- Gavin WJ, Giles L. Sample size effects on temporal reliability of language sample measures of preschool children. *Journal of Speech and Hearing Research*. 1996; 39:1258–1262. [PubMed: 8959610]
- Heilmann J, Miller J, Nockerts A. Using language sample databases. *Language, Speech, and Hearing Services in School*. 2010; 41:84–95.
- Hutchins TL, Brannick M, Bryant JB, Silliman ER. Methods for controlling amount of talk: *Difficulties, considerations and recommendations*. *First Language*. 2005; 25:347–363.
- Kagan, J. The structure of temperament. In: Emde, RN.; Hewitt, JK., editors. *Infancy to early childhood: Genetic and environmental influences on developmental change*. New York: Oxford University; 2001. p. 45-51.
- Klee T, Schaffer M, May S, Membrino I, Mougey K. A comparison of the age-MLU relation in normal and specifically language-impaired preschool children. *Journal of Speech and Hearing Disorders*. 1989; 54:226–233. [PubMed: 2709841]
- Klee T, Stokes SF, Wong AM-Y, Fletcher P, Gavin WJ. Utterance length and lexical diversity in Cantonese-speaking children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research*. 2004; 47:1396–1410.
- Leadholm, BJ.; Miller, JF. *Language Sample Analysis: The Wisconsin Guide*. Madison, Wisconsin: Wisconsin Department of Public Instruction; 1992. Analysis; p. 36-51.
- Loban, W. *Language development: Kindergarten through grade twelve*. Urbana, IL: National Council of Teachers of English; 1976. Research Report No. 18
- MacWhinney, B.; Spektor, L. *Computerized Language Analysis (Version 7.28.09)* [Computer software]. Pittsburgh, PA: Author; 2009.
- McKee G, Malvern D, Richards B. Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*. 2000; 15:323–337.
- Mikucki, BA.; Larrivee, L. Validity and reliability of twelve child language tests. Miami, FL: American Speech-Language-Hearing Association Annual Convention; Nov. 2006
- Miller, JF. *The Systematic Analysis of Language Transcripts Guide (Research Version 8.0)* [Computer software.]. Madison, WI: University of Wisconsin; 2004.
- Miller, JF.; Long, S.; McKinley, N.; Thormann, S.; Jones, MA.; Nockerts, A. *Language Sample Analysis II: The Wisconsin Guide*. Madison, Wisconsin: Wisconsin Department of Public Instruction; 2005.
- Morales M, Mundy P, Delgado CEF, Yale M, Neal R, Schwartz HK. Gaze following, temperament, and language development in 6-month-olds: A replication and extension. *Infant Behavior & Development*. 2000; 23:231–236.
- Mullineaux PY, Deater-Deckard K, Petrill SA, Thompson LA, Dethorne LS. Temperament in middle childhood: A behavioral genetic analysis of fathers' and mothers' reports. *Journal of Research in Personality*. 2009; 43:737–746. [PubMed: 20216930]
- Nippold, MA. *Later Language Development: The School-Age and Adolescent Years*. 2nd edn.. Austin, TX: Pro-ed; 1998.
- Owen AJ, Leonard LB. Lexical diversity in the spontaneous speech of children with specific language impairment: Application of D. *Journal of Speech, Language, and Hearing Research*. 2002; 45:927–937.
- Petrill SA, Deater-Deckard K, Thompson LA, Dethorne LS, Schatschneider C. Reading skills in early readers: Genetic and shared environmental influences. *Journal of Learning Disabilities*. 2006; 39:48–55. [PubMed: 16512082]
- Plante E, Vance R. Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools*. 1994; 25:15–24.
- Putnam SP, Rothbart MK. Development of Short and Very Short Forms of the Children's Behavior Questionnaire. *Journal of Personality Assessment*. 2006; 87:102–112. [PubMed: 16856791]
- Redmond S, Rice M. The socioemotional behaviors of children with SLI: Social adaptation or social deviance? *Journal of Speech, Language, and Hearing Research*. 1998; 41:688–700.
- Rice ML, Redmond SM, Hoffman L. Mean length of utterance in children with specific language impairment an in younger control children shows concurrent validity and stable and parallel growth trajectories. *Journal of Speech, Language, and Hearing Research*. 2006; 49:793–808.

- Scott CM, Windsor J. General language performance measures in spoken and written narrative and expository discourse of school-age children with language-learning disabilities. *Journal of Speech, Language, and Hearing Research*. 2000; 43:324–339.
- Slomkowski CL, Nelson K, Dunn J, Plomin R. Temperament and language: Relations from toddlerhood to middle childhood. *Developmental Psychology*. 1992; 28:1090–1095.
- Ukrainetz TA, Blomquist C. The criterion validity of four vocabulary tests compared to a language sample. *Child Language Teaching and Therapy*. 2002; 18:59–78.
- Watkins RV, Kelly DJ, Harbers HM, Hollis W. Measuring children's lexical diversity: Differentiating typical and impaired language learners. *Journal of Speech and Hearing Research*. 1995; 38:1349–1355. [PubMed: 8747826]

What this paper adds

What is already known

Conversational language measures, such as mean length of utterance and number of different words, are common tools in the study of child language. Despite evidence of construct validity, concerns remain regarding the extent to which such measures are confounded by aspects of child temperament, such as extraversion. Prior papers on child language and temperament have focused on infants and toddlers, rather than school-age children, and none have included direct observation of children's natural language use.

What this paper adds

This paper addressed whether or not differences in child temperament, such as how Surgent (i.e., extraverted) a child is, influence language sample measures, such as number of different words (NDW) and mean length of utterance (MLU). The findings suggest that clinicians and investigators who are interested in limiting the confound of child temperament in the assessment of school-age language should consider using MLU, which averages across utterance number, or NTW and NDW, which can be derived from a standardized number of utterances. Differences across language sample measures are discussed, and the need to examine the sensitivity and specificity of language sample measures, particularly as a function of volubility, is highlighted.

Table 1

Two fictional language samples (A and B) provided to illustrate the differences across vocabulary measures depending on how they are calculated

Calculation Method	Sample A	Sample B
	Yeah. I like pets. This dog's brown. Like mine. Do you have a dog? I take mine for walks. Sometimes I don't want to. Mom makes me.	It's a Beagle. Mom calls him Captain_Exasperating*. He gnaws on everything. Like my Harry_Potter* library book. He shredded it to bits.
NDW Total sample	22	19
NDW on 5-utt. cut	12	19
NDW on 21-token cut	16	19
Type-Token Ratio	.81	.90

* *Note:* Consistent with transcription convention (Miller, 2004, p.43), proper names were transcribed as 'frozen forms' and counted as single words.

Table 2

Means and standard deviations (*SD*) for the individual conversational language measures at the second (HV2) and third (HV3) home visits, as well as the values averaged across home visits (AVE)

	HV2		HV3		AVE	
	n	Mean (<i>SD</i>)	n	Mean (<i>SD</i>)	n	Mean (<i>SD</i>)
Total Number of Complete & Intelligible C-Units	161	134.29 (44.67)	161	138.17 (47.92)	161	136.23 (39.90)
Number of Total Words ^a	125	531.84 (111.07)	124	552.48 (116.46)	145	538.66 (104.60)
Number of Total Words-uncut	159	710.75 (296.66)	160	759.04 (356.04)	160	734.95 (290.13)
Mean Length of C-Unit	155	5.81 (1.19)	154	5.98 (1.40)	159	5.86 (1.18)
Total Number of Conjunctions ^a	125	38.20 (19.58)	124	42.14 (20.53)	145	40.20 (18.78)
Total Number of Conjunctions-uncut	161	50.84 (31.66)	161	58.49 (39.75)	161	54.67 (31.06)
Number of Different Words ^a	125	194.29 (29.00)	124	198.20 (28.23)	145	195.81 (26.32)
Number of Different Words-uncut	159	228.31 (63.80)	160	237.83 (74.43)	160	233.20 (61.36)
Measure D	161	73.70 (12.91)	161	73.53 (14.26)	161	73.61 (11.65)

^a Values derived from the first 100 complete and intelligible utterances within a sample

Table 3

Correlations between language measures, Surgency, and individual facets of Surgency

	TCICU	NTW ^a	NTW-uncut	MLU	TNC ^a	TNC-uncut	NDW ^a	NDW-uncut	D
Surgency	.34*	-.10	.24*	.04	-.22*	.06	-.08	.27*	.24*
High Intensity Pleasure	.25*	-.16	.16	-.00	-.27*	-.02	-.10	.20	.22*
Activity	.15	-.10	.07	-.03	-.19	-.06	-.05	.09	.12
Impulsivity	.19	-.07	.14	.02	-.17	.02	-.03	.16	.16
Shyness	-.35*	-.01	-.30*	-.11	.06	-.19	.05	-.30*	-.19

* Denotes statistical significance at $p < .01$ ^a Values derived from the first 100 complete and intelligible utterances within a sample

Table 4Correlations between language and temperament measures reported for *girls* only

	TCICU	NTW	NTW-uncut	MLU	TNC	TNC-uncut	NDW	NDW-uncut	D
Effortful Control	-.15	.08	-.12	-.02	.10	-.05	.05	-.13	-.10
Attention Focusing	-.24	.06	-.20	-.05	.14	-.10	.05	-.19	-.21 [†]
Inhibitory Control	-.15	.15	-.08	.03	.20	.03	.12	-.10	-.12
Low Intensity Pleasure	.11	-.09	.04	-.09	-.23	-.10	-.04	.03	.09
Perceptual Sensitivity	-.11	.07	-.05	.05	.13	.04	-.01	-.07	.02
Negative Affectivity	.19	-.09	.11	-.05	-.11	.03	-.09	.12	.06
Fear	.08	.01	.09	.08	.06	.13	-.01	.10	-.04
Anger	.10	-.09	.03	-.06	-.14	-.05	-.06	.07	.15
Sadness	.24	-.09	.15	-.01	-.13	.04	-.04	.16	.10
Discomfort	.07	-.08	.02	-.16	-.05	-.01	-.15	-.03	-.14
Surgency	.40*	-.07	.30*	.09	-.20	.13	-.01	.36*	.31
High Intensity Pleasure	.32*	-.19	.18	-.03	-.27	.01	-.09	.24	.33*
Activity	.17	-.14	.08	-.04	-.19	-.04	-.04	.12	.17
Impulsivity	.27*	-.04	.21	.06	-.16	.08	.03	.27*	.27
Shyness	-.43*	-.10	-.38	-.21	.03	-.27	-.05	-.40*	-.20

* Denotes statistical significance at $p < .01$ [†] Denotes statistical difference in the correlations for girls versus boys using Fisher's r-to-z statistic ($p < .01$)

Table 5Correlations between language and temperament measures reported for *boys* only

	TICU	NTW	NTW-uncut	MLU	TNC	TNC-uncut	NDW	NDW-uncut	D
Effortful Control	-.02	.07	.01	.05	.06	.02	.08	.02	.11
Attention Focusing	-.07	.12	.04	.19	.11	.07	.19	.08	.25 [†]
Inhibitory Control	-.18	.04	-.13	-.00	.09	-.07	.05	-.13	-.08
Low Intensity Pleasure	.19	.03	.15	-.04	.00	.04	-.05	.10	.03
Perceptual Sensitivity	.02	-.00	-.01	-.04	-.03	.01	-.00	.00	.08
Negative Affectivity	.35*	.11	.34*	.13	.01	.26	.04	.29	-.05
Fear	.20	.01	.15	-.07	-.01	.05	.00	.11	-.03
Anger	.22	.02	.19	.05	-.07	.14	-.07	.14	-.12
Sadness	.29	.14	.33*	.21	.01	.28	.05	.29	-.03
Discomfort	.26	.09	.26	.11	.04	.20	.03	.21	-.02
Surgency	.24	-.14	.16	-.02	-.26	-.04	-.17	.15	.14
High Intensity Pleasure	.19	-.12	.14	.04	-.28	-.06	-.10	.15	.10
Activity	.14	-.03	.07	-.02	-.19	-.08	-.07	.06	.06
Impulsivity	.07	-.13	.02	-.06	-.23	-.09	-.14	-.00	.03
Shyness	-.24	.10	-.18	.03	.06	-.08	.16	-.16	-.17

* Denotes statistical significance at $p < .01$ [†] Denotes statistical difference in the correlations for girls versus boys using Fisher's r-to-z statistic ($p < .01$)