AJSLP

Research Article

The Index of Productive Syntax: Psychometric Properties and Suggested Modifications

Ji Seung Yang,[a] Brian MacWhinney,[b] and Nan Bernstein Ratner[c]

Purpose: The Index of Productive Syntax (IPSyn) is a well-known language sample analysis tool. However, its psychometric properties have not been assessed across a wide sample of typically developing preschool-age children and children with language disorders. We sought to determine the profile of IPSyn scores by age over early childhood. We additionally explored whether the IPSyn could be shortened to fewer items without loss of information and whether the required language sample could be shortened from a current required number of 100 utterances to 50.
Method: We used transcripts from the Child Language Data Exchange System, including 1,051 samples of adult–child conversational play with toys within the theoretical framework of item response theory. Samples included those from typically developing children as well as children with hearing loss, Down syndrome, and late language emergence.
Results: The Verb Phrase and Sentence Structure subscales showed more stable developmental trajectories over the preschool years and greater differentiation between typical and atypical cohorts than did the Noun Phrase and Question/Negation subscales. A number of current IPSyn scoring items can be dropped without loss of information, and 50-utterance samples demonstrate most of the same psychometric properties of longer samples.
Discussion: Our findings suggest ways in which the IPSyn can be automated and streamlined (proposed IPSyn-C) so as to provide useful clinical guidance with fewer items and a shorter required language sample. Reference values for the IPSyn-C are provided. Trajectories for one subscale (Question/Negation) appear inherently unstable and may require structured elicitation. Potential limitations, ramifications, and future directions are discussed.
Supplemental Material: https://doi.org/10.23641/asha.16915690

Language sample analysis (LSA) is viewed as a critical component of expressive child language assessment. Numerous analytical algorithms have been developed for LSA, starting with measures of vocabulary diversity, such as type–token ratio (Templin, 1957) and mean length of utterance (MLU; Brown, 1973), and progressing to more ambitious procedures for the analysis of syntactic structure, such as Developmental Sentence Scoring (DSS; Lee, 1974); the Language Assessment, Remediation and Screening Procedure (Crystal et al., 1981); and the Index of Productive Syntax (IPSyn; Scarborough, 1990). Because of its wide and active use in the literature, we will focus here on the use of the IPSyn for clinical assessment.

The IPSyn has undergone minor modifications recently, intended to simplify its scoring routines (Altenberg et al., 2018). In its current implementation, it consists of 59 target structures accorded points for up to two instances seen in the child's language sample. The IPSyn divides structures of interest into four categories: Noun Phrase (NP) development/elaboration, Verb Phrase (VP) development/elaboration, Question/Negation (Q/N) constructions, and Sentence (S) phrase structure. IPSyn analysis can be valuable for detecting group differences between typical and atypical/delayed expressive language in children (discussed in greater detail below), as well as for identifying structures absent from a client's repertoire. Identification of absent structures, in turn, can be used to create individualized and informed intervention goals for clients with expressive language weakness, as outlined in two recent clinical tutorials (Finestack et al., 2020; Pezold et al., 2020). In an analysis of samples from sixty-eight 3-year-old children who spoke

[a]Department of Human Development and Quantitative Methodology, College of Education, University of Maryland, College Park
[b]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA
[c]Department of Hearing and Speech Sciences, University of Maryland, College Park
Correspondence to Nan Bernstein Ratner: nratner@umd.edu

Editor-in-Chief: Erinn H. Finke
Editor: Lynne E. Hewitt
Received March 16, 2021
Revision received May 28, 2021
Accepted August 2, 2021
https://doi.org/10.1044/2021_AJSLP-21-00084

Disclosure: *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

African American English (AAE), Stockman et al. (2016) also found the IPSyn to be a relatively dialect-fair LSA tool, critical to any language sampling conducted with non–General American English (GAE) speakers.

A major obstacle to the widespread use of the IPSyn for diagnostic or therapy planning purposes is the commitment of time and linguistic knowledge necessary to conduct the analysis. First, the procedure, even in its recent revision, requires collection and transcription of 100 eligible utterances from the child. This can be difficult for the youngest or most delayed children or those who are less talkative during assessment interactions. Second, the scoring process takes a significant amount of time when conducted by hand (Long & Channell, 2001; Stockman et al., 2016). Third, if done by hand, the procedure requires an advanced degree of linguistic knowledge, particularly for accurate identification of later developing structures. One study estimated the accuracy of grammatical tagging by clinicians at around 70% (Justice & Ezell, 1999). Together, these three limitations have acted historically to discourage the widespread use of analytical tools such as the IPSyn (Finestack & Satterlund, 2018; Long, 2001; Long & Channell, 2001), despite its clear utility for setting customized treatment goals (Finestack et al., 2020; Long & Channell, 2001; Pezold et al., 2020).

A promising solution to the problem of clinician time and coding expertise involves the use of computerized routines for semiautomatic IPSyn computation (e.g., Hassanali et al., 2014; Sagae et al., 2005). Numerous versions of computer-assisted IPSyn have been developed, some of which have fairly good overall scoring agreement with informed manual scoring (Long & Channell, 2001) but less good reliability on an item-by-item basis (Altenberg & Roberts, 2016), important for goal-setting purposes. An early version of the IPSyn that was included in the Computerized Language Analysis (CLAN) program distributed via https://talkbank.org (MacWhinney, 2000) likewise appeared to have better agreement for total scores than for individual items (J. A. Roberts et al., 2020). However, working together, MacWhinney et al. (2020) were able to improve the item-level accuracy of CLAN's IPSyn analysis, performed on simple, uncoded transcripts of child utterances, to over 95%.

Despite these advances, four questions regarding IPSyn analysis remain, and the remainder of this article will address these in turn, using a large number of language samples from both typically developing (TD) and atypical young children from North American English–speaking communities. These four questions are (a) the validity of using IPSyn scores diagnostically to flag an individual child's language sample as falling outside age expectations, (b) whether the structures targeted by the IPSyn show a clear age-related profile of growth over the age range of 2–6 years, and (c) whether or not the IPSyn's individual subscales (NP, VP, Q/N, and S) appear to be equally informative in appraising child language skill. Finally, we will evaluate (d) whether or not a 100-utterance sample is required for IPSyn analysis in order to demonstrate the psychometric properties desired for an LSA measure. In comparison, the most typical alternative LSA measure for examining fine-grained grammatical development, namely, DSS, requires only 50 qualifying utterances. From the perspective of clinical time required to elicit and transcribe a language sample, even if machine scoring is used, a shorter sample requirement could make IPSyn analysis more attractive and accessible to a greater number of practicing clinicians. We will place these four questions about the IPSyn into context in the sections that follow.

Psychometric Properties of the IPSyn

The original IPSyn reference sample had only 15 children followed longitudinally from 24 to 48 months of age, and the recent revision only rescored 20 child language samples using each method to ascertain relative agreement between versions. However, in published research, the IPSyn generally shows statistically significant differences between TD children and children with language disorders. Specifically, the IPSyn yields distinct profiles for children not at risk for communication disorders in comparison with children with autism and children with Down syndrome (DS; Scarborough et al., 1991) through 48 months of age. These findings suggest that the structures targeted for scoring by the IPSyn do, in fact, bear on general patterns of age-appropriate and impaired or delayed expressive language.

Use of the IPSyn as a Diagnostic Measure

Our first concern is whether or not the IPSyn's psychometric properties are sufficiently robust to provide the practicing speech-language pathologist (SLP) with diagnostic information regarding the typicality of a child's expressive language profile. There is a very real distinction between comparing IPSyn scores from cohorts of children with various diagnoses or background experiences to examine subtle indices of linguistic delay or to set therapeutic goals and use of the IPSyn as a normed device capable of discriminating typical from atypical performance. To date, reference values for expected IPSyn score performance by age are based on relatively few children provided in relatively broad age bands. More data have been collected in research publications to describe differences between children with and without risk factors for language delay than to validate use of the IPSyn as a diagnostic tool.

To our knowledge, only one study has examined whether IPSyn scores can robustly distinguish typical from impaired performance. Oetting et al. (2010) found that although the IPSyn was appropriate to use with children who speak "nonstandard" varieties of American English, it was "insensitive to child language impairment" in their sample of 4- to 6-year-old children, of whom one third were previously diagnosed with specific language impairment (SLI).

IPSyn Growth Over Early Childhood

A second question is whether the IPSyn is sensitive to grammatical development over the wider age range of 2–6 years. Hadley et al. (2016) used IPSyn analysis of their twenty 24- to 30-month-old children's verb lexicon development

and noted that numerous prior researchers had chosen this measure for analysis of language samples collected from 3-year-old children (e.g., Horton-Ikard et al., 2005; Sanchez et al., 2020; Scarborough et al., 1991). Tomblin et al. (1999) found that for children with roughly 3 years of cochlear implant (CI) or hearing aid use, clear differences in IPSyn scores were seen when language samples were contrasted with those of children with typical hearing.

Given its wide range of targeted structures and potential for clinical guidance, the IPSyn might be particularly valuable as an assessment tool with older children. Indeed, some researchers have employed IPSyn analysis with older children and found statistically significant differences between 5-year-olds with and without expressive language delay (Rescorla & Turner, 2015). Two studies have suggested that the IPSyn may also be an age-sensitive syntactic structure tool for measuring language in TD children up to 6 years of age (Hewitt et al., 2005; Oetting et al., 1999). Pushing the age range further, Geers and Sedey (2011) found differences between students with more than 10 years of CI experience and peers with normal hearing, suggesting that the IPSyn may be informative at later ages of linguistic development.

However, Rice et al. (2006) found a relatively poor correlation between age and IPSyn total scores for children between 3 and 5 years of age, although IPSyn total scores were significantly lower for children with diagnoses of SLI than for age- or language-matched peers. Bernstein Ratner and MacWhinney (2016) found that IPSyn scores computed for over 600 children with presumed typical language development tended to plateau after roughly 36–42 months of age. Although Eigsti and Cicchetti (2004) found statistically significant lower IPSyn values for 60-month-old children with a history of maltreatment, compared to those without maltreatment, both groups functioned well below age expectations, with scores more appropriate to 41- and 46-month-old children in Scarborough's (1990) sample, respectively. Estigarribia et al. (2012) found that the IPSyn differentiated the expressive language of 9- to 10-year-old children with fragile X syndrome and DS from language- and age-matched 5-year-old peers. Observing 6-year-old children with and without SLI, Hewitt et al. (2005) found that while IPSyn total scores were lower for the children with SLI, subscale totals did not reliably distinguish between groups.

At least some studies have examined the potential for IPSyn scores to plateau due to a maximum of two targeted structures per phrase structure rule. In addition to the large-scale analysis by Bernstein Ratner and MacWhinney (2016), Tomblin et al. (1999) found that by 2 years postimplantation, many of their CI users achieved maximum scores and attributed this to "ceiling effects in the language measures."

Consistency of Subscale Performance on the IPSyn

A third concern regarding psychometric properties of the IPSyn is that there also have been some reports that not all scales of the IPSyn appear to reflect language maturation equally well. In a relatively large sample ($N = 62$) of Jamaican Creole–English bilingual preschoolers, the NP, VP, and S subscales were significantly correlated both across languages and with other expressive language measures. However, the Q/N subscale was not (Washington et al., 2019). Similarly, J. E. Roberts et al. (2007) found differences between boys with fragile X syndrome and mental age–matched typical peers on the NP, VP, and S subscales, but not Q/N. Hewitt et al. (2005) found that 6-year-old children with SLI did not uniformly achieve statistically lower scores on all subscales of the IPSyn than typical peers. Hadley (1998) found the VP subscale to be superior to the NP subscale in distinguishing between groups of typical children and children with SLI at ages 19–38 months.

Finally, both the original and the revised IPSyn are premised on a language sample that provides at least 100 eligible utterances for scoring. Although not targeted as a concern in prior studies, Bernstein Ratner and MacWhinney (2016) found, not surprisingly, that the youngest children in their corpus of more than 600 children ages 2–6 years often failed to provide a sufficiently large sample for IPSyn analysis. Children who are less voluble or show language delay (e.g., Rescorla & Bernstein Ratner, 1996) might also require extensive sampling to provide a sufficiently large sample. Thus, because shorter samples might be more appropriate in the assessment of younger, less verbal children and because shorter samples may ease SLPs' time in data collection and analysis, a fourth question is the necessity of a 100-utterance language sample in producing informative IPSyn analysis results for clinical use.

In summary, the IPSyn appears to be a sensitive tool for discriminating groups of children differing broadly in language skill, although less information is available to suggest that IPSyn scores can reliably identify individual children as TD or not. The IPSyn also holds great promise as an informative tool for identifying phrase structures that individual children appear to be able to use productively and those that they cannot, for therapeutic goal-setting purposes. However, as noted, no large-scale reference data are available to benchmark individual children's performance across the early years of English language development robustly for the purpose of identifying language delays. There is some evidence that not all items or subscales of the IPSyn have equivalent discriminative properties. Finally, even with open-access tools to enable machine scoring of the IPSyn using simple, uncoded language transcripts, saving time and the need for specialized expertise, the traditional IPSyn analysis requires at least 100 qualifying utterances, a fairly sizable commitment of language sample collection time and transcription. Thus, we undertook to explore these issues, using a large, open-access repository of adult–child conversational language samples from children 2–6 years of age.

Method

We examined the psychometric properties of IPSyn items with respect to its four subscales and overall scale and appraised the developmental patterns of scores seen in four different groups: children with no known or suspected language disorder (TD group) and three clinical samples,

namely, children under 42 months of age diagnosed as late talkers (LTs; LT group), children with hearing loss (HL; HL group), and children with DS (DS group). The three clinical cohorts were analyzed, with the primary goal of identifying possible differences in IPSyn profiles that might strengthen the validity of the IPSyn as a diagnostic tool. It was not our intent to analyze specific profiles for children with these diagnoses; rather, we were interested in the ability of IPSyn score items to grossly distinguish between children who are presumed to show typical language development and children identified as showing language impairment or delay or its risk factors. We analyzed children's item response data from a large collection of language samples within the theoretical framework of item response theory. "Item response theory (IRT) is a collection of mathematical models and statistical methods used for two primary purposes: item analysis and test scoring" (Thissen & Steinberg, 2009). While IRT is widely and routinely used to develop tests or surveys in education and social sciences (e.g., the Scholastic Aptitude Test), it has not yet been used to analyze the items scored by the IPSyn. One reason for this could be the fact that the sample size (number of independent observations) required for latent variable modeling, such as IRT analysis, is typically larger than 200 at the minimum to estimate model parameters (e.g., Hoe, 2008; Singh et al., 2016). Most studies using LSA collect far fewer samples than this required minimum and, therefore, would not qualify for analysis through IRT. However, by using the large number of uniformly transcribed samples collected in the Child Language Data Exchange System (CHILDES) data repository system (https://childes.talkbank.org) as well as the automated scoring system provided by the CLAN program that permits uniform assignment of IPSyn scores to these samples (https://dali.talkbank.org), we were able to identify sufficient language samples to permit IRT analysis.

Sample

A total of 1,127, 504, 169, and 72 (total n = 1,872) adult–child dyadic play files were collected respectively from 14, five, three, and two CHILDES corpus studies whose target population groups were TD, LT, HL, and DS, respectively. The vast majority of these samples were recorded between the children and their mothers, engaged in joint play, using a study-specific set of materials that were consistent across recording sessions within individual projects. Our primary interest was the applicability of IPSyn analysis across childhood in children with typical communication development. However, we added samples from children with late language emergence, HL, and DS for exploratory comparison purposes, to ascertain whether or not children with known diagnoses or risk factors for expressive language impairment might display profiles distinct from those obtained from children with presumptive typical development. As such, we mapped the performance of these additional groups of children as a preliminary assessment of the validity of the IPSyn in distinguishing between typical and clinical populations. Because up to 80% of children with diagnoses of late talking are well known to "catch up" to typical peers by the age of 4 years, we only analyzed LT data from the "Rescorla" and "EllisWeismer" corpora for children up to 42 months of age.

These data also reflect a mix of original study designs. Thus, in Table 1, we specify how these corpora are coded within the CHILDES database for North American English: "Toy" marks a design involving adult–child play with toys; "Cross" indicates a corpus collected cross-sectionally over child age groups; and "Long" codes a corpus in which individual children were tracked longitudinally.

The original number of 1,872 files was reduced by requirements of minimum sample length: Samples that met the age criterion (24–74 months) and the 100-eligible-utterance criterion for the IPSyn were 338 (TD), 127 (LT), 20 (HL), and 16 (DS), whereas the samples that satisfied both the age criterion and the lower 50-utterance criterion were 639 (TD), 354 (LT), 31 (HL), and 27 (DS). The number of samples per study and the descriptive statistics for gender and age for these four groups of children are presented in Table 1.

Analysis

Calibration of IPSyn Items

Because each item in the IPSyn is scored as 0, 1, or 2 (the maximum types scored for a given structure), we fit a graded response model (GRM; Samejima, 1969) that can handle ordered polytomous item responses. The full information maximum likelihood estimation with the expectation–maximization algorithm (Bock & Aitkin, 1981) implemented in the IRT software package flexMIRT (Version 3.5; Cai, 2017) was used to obtain the model parameter estimates and cross-product standard errors. The GRM equation and the interpretation of model parameters are available in Supplemental Material S1 for readers who are interested in further details. In the IRT analyses, item characteristic curves (ICCs, or item trace lines) are often presented based on the estimated parameters; test characteristic curves (summation of ICCs) or test information function can be also presented to demonstrate how the set of items is associated with latent ability being measured. As an illustration, Figure 1 shows an example of the estimated ICC for Noun Item 8 (two-word NP before a verb) and the corresponding item parameter estimates. The upper panel shows the probability of receiving an item score greater than a particular score (0, 1, 2) given latent trait levels, the middle panel shows the probability of receiving a particular score (0, 1, 2) given latent trait levels, and the bottom panel presents the expected item score given the latent trait level.

We first fit the unidimensional GRM for each subscale of the IPSyn and the overall IPSyn score using the language sample data from the TD group only (N = 338 child samples meeting 100-utterance sample size constraints), given that the sample size for the other three groups was inappropriately small for item calibration. While we are also aware that a multidimensional GRM can be an option, we decided not to do so as this sample size is not sufficient for fitting complex models. Goodness-of-model-fit assessment was evaluated

Table 1. Number of language samples per study and child sample characteristics.

Group	CHILDES	Studies	*N* of all corpora	100-utt criterion satisfied			50-utt criterion satisfied		
				Age range	*N*	Summary	Age range	*N*	Summary
Toy/Cross/ Typical	Eng-NA	Bates	101	NA	0	$N = 338$ $M_{age} = 40.4$	28–28	7	$N = 639$ $M_{age} = 38.0$
		Bliss	7	27–73	7	Range: 24–74	27–73	7	Range: 24–74
		Morisset	196	NA	0	52% male	30– 39	53	47% male
		NewmanRatner/24	124	24–24	4		24–24	59	
		Tardif	25	NA	0		NA	0	
		Valian	43	25–32	36		25–32	37	
		VanHouten (free-play)	45	39–41	3		38–43	22	
		VanKleeck	40	42–48	31		37–48	36	
		Warren	20	24–74	13		24–74	17	
	Clinical-MOR	EllisWeismer/TD	296	30–66	126		30–66	271	
		Feldman/ParentChild/ TD	57	24–42	22		24–55	28	
		Hooshyar/TD/play	29	NA	0		NA	0	
		Nicholas/TD	103	24–54	68		24–55	71	
		Rondal/TD	41	24–54	28		24–55	31	
Toy/Long/ Atypical	Clinical-MOR	EllisWeismer/LT	280	30–66	68	$N = 126$ $M_{age} = 48.6$	30–66	221	$N = 354$ $M_{age} = 45.8$
		Rescorla	70	36–48	38	Range: 30–82	36–49	55	Range: 30–82
		EisenbergGuo/LT	17	37–47	16	68% male	36–47	49	70% male
		Hargrove	82	NA	0		44–55	4	
		UCSD/SLI	55	60–82	4		47–82	25	
Hearing loss	Clinical-MOR	Ambrose/HL	106	26–36	13	$N = 20$ $M_{age} = 36.2$	26–53	21	$N = 31$ $M_{age} = 33.3$
		Bliss	7	36–76	4	Range: 26–76	36–76	5	Range: 26–76
		Nicholas/HL	56	36–47	3	20% male	36–47	5	26% male
Down syndrome	Clinical-MOR	Hooshyar/DS/play	31	65–65	1	$N = 16$ $M_{age} = 53.6$	38–65	4	$N = 27$ $M_{age} = 49.2$
		Rondal/DS	41	38–80	15	Range: 38–80 56.3% male	38–80	23	Range: 38–80 55.6% male

Note. Age unit is months. CHILDES = Child Language Data Exchange System; utt = utterance; Eng-NA = North American English; Clinical-MOR = Clinical sample with morphological tagging; TD = samples from typically developing children; LT = samples from children with diagnoses of late talking; UCSD = University of California, San Diego; SLI = samples from children with diagnoses of specific language impairment; HL = samples from children with hearing loss; DS = samples from children with Down syndrome; NA = not applicable.

using limited information fit statistics[1] root-mean-square error of approximation (RMSEA) based on M_2 statistics (Maydeu-Olivares & Joe, 2005). In interpreting our results, we note that an RMSEA value smaller than 0.05 is considered to be an excellent fit, whereas 0.05–0.07 is considered an acceptable fit (Maydeu-Olivares & Joe, 2014).

IPSyn Scale Scores Over Early Childhood

After examining item properties and refining the measure using the TD sample, we calculated IRT scaled scores (expected a posteriori [EAP]; Thissen & Wainer, 2001) for all samples to examine the pattern of score trajectories across chronological age. EAP is analogous to standardized factor scores rather than summed or averaged observed scores. The calculated EAP scores were regressed on the child's chronological age using different models (linear, segmented, curve–linear, and nonparametric) within each sample to find the best representation of the data and theoretical expectations. Scaled scores were calculated in flexMIRT (Version 3.5; Cai, 2017), and the regression analysis was conducted in R (R Core Team, 2020). Full details of our statistical analysis are provided in Supplemental Material S1.

[1]We used fit statistics to compare the observed and expected probabilities of the second-order margin of item response patterns because the comparison of the full frequency tables of response patterns (called Pearson's residual chi-square) could not be used to the large number of possible response patterns (e.g., the number of response patterns for 10 items with 3 score categories is $3^{10} = 59{,}049$, whereas the sample size is thousands at the most).

Results

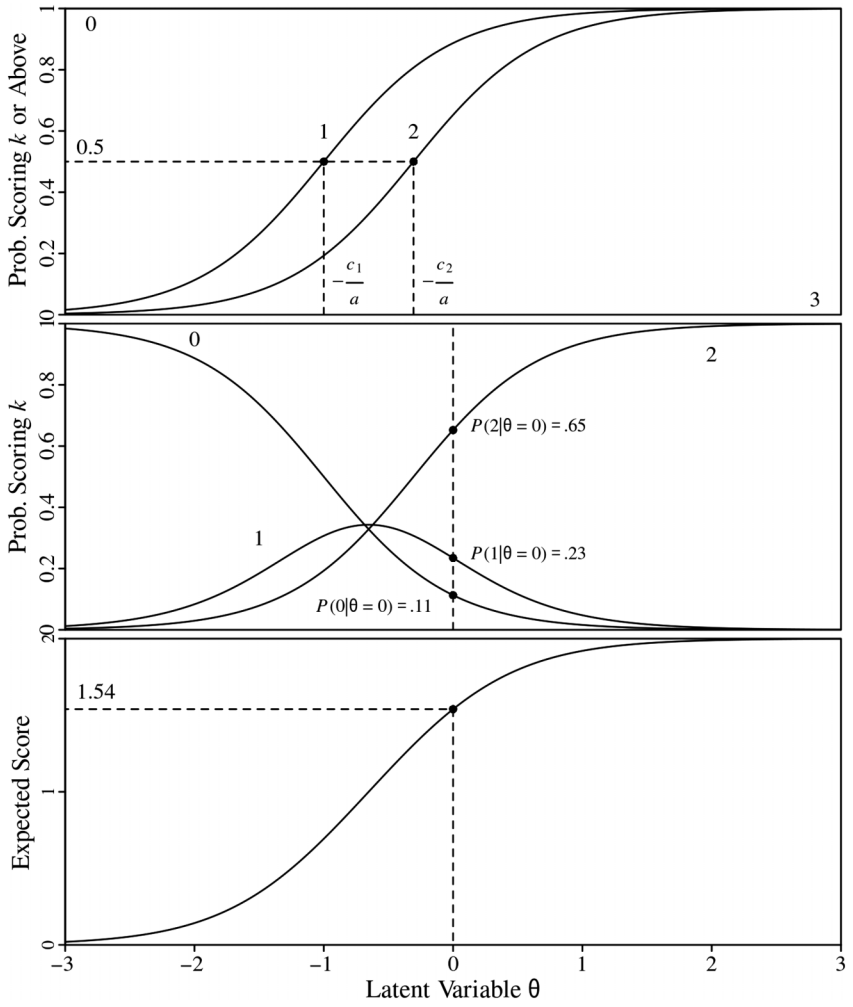
Calibration of IPSyn Items

In the sections that follow, we evaluate the information provided by specific IPSyn target structures. To preview our discussion, we found that some items do not add substantially to an informative profile of the child's expressive language. When these items are dropped from scoring, model fit is improved. In Tables A1 and A2 in the Appendix, we provide a suggested revision of target items for scoring, along with clinical guidance for score interpretation. To disambiguate the proposed revision from prior versions of the IPSyn (Altenberg et al., 2018), we call this adaptation of IPSyn administration the IPSyn-C.

NP Assembly Ability Items

The NP subscale has a total of 11 items (N1–N11). However, for the first four items (N1–N4) from 338 TD

Figure 1. The item characteristic curve for Index of Productive Syntax Item N8.



samples, all achieved the maximum score of 2, which means that these items are too easy for the target age group and basically a constant value across all observed cases. These items consist of NPs composed simply of a noun, a pronoun, a modifier, or two words. Even the youngest children showed at least two of each of these structures in their samples. Therefore, these items were omitted from the statistical modeling. Next, when a unidimensional GRM was fitted, the slope of Item 5 (article + noun) was extremely high (10), indicating that N5 functions nearly as a Guttman (1944) item (see the Appendix for details about Guttman-like items), which means it is measuring a construct tapped by other NP items. Accordingly, we decided to drop N5 (article + noun), as it duplicates information captured by the rest of the retained items. The most parsimonious explanation for N5 duplicating other items is its overlap with N3 (modifier + noun) and N4 (a two-word NP). If a child provided 100 eligible utterances, the majority of these early NP constructions were invariably present. After dropping Item N5 and retaining the remainder of NP structures, the model fit was excellent, with $M_2 = 47.83$ ($df = 54$, $p = .70$). However, the marginal reliability (Green et al., 1984) that corresponds to the internal consistency coefficient (e.g., alpha coefficient in classical test theory) for the NP assembly ability was 0.61, which can be considered low. The item parameter estimates and standard errors are reported in Table A1 in the Appendix, and the ICCs and test information function are reported in Supplemental Materials S2–S4. It is noteworthy that N7 (noun plural) also exhibited a statistically insignificant slope of change over time, which means that this item also represents a candidate for exclusion or revision. Examination of ICCs for N7 implies that this structure is mastered very early in children's conversational speech, as also might be suggested by Brown's (1973) work and subsequent examination of morphological development; therefore, this item is not very informative. N11 (any bound morpheme on a noun or an adjective not credited previously) also has a relatively flatter slope of development than other remaining NP items. In summary, many NP structures credited by the IPSyn do not discriminate well over early development

if a child is capable of providing 100 eligible utterances for analysis.

VP Assembly Ability Items

The VP subscale has a total of 17 items (V1–V17). However, most of the sample, if not all, achieved a score of 2 on the first two items (V1 and V2; use of any verb or verb particle or preposition). Again, this might be expected given historical work showing prepositions to be among the first morphemes mastered by young English-speaking children. Hence, the model fitting was conducted only for Items V3–V17. The initial model showed that Item V5 (catenative) and Item V8 (present progressive) had too high a slope estimate, which is a signal of redundant information or local dependency with other items (this is also noted in the IPSyn's guidance that crediting some verb items, such as V8, concurrently credits an earlier item, such as V1). In addition, V15 exhibited the lowest slope among this set of items. This is the uncontractible[2] copula or auxiliary used for ellipsis (e.g., "Yes, he is," but not *"Yes, he's") typically expected in response to questions such as "Is he happy/going?"

As shown in the next section, questioning/answering activity during spontaneous language sampling may be highly unpredictable, in addition to any inherent difficulty of the item itself. We note that Brown (1973) found uncontractible auxiliaries and copulas to be acquired *before* contractible forms, presumably because the uncontractible form is more salient (usually due to utterance-initial or utterance-final position), and ellipsis is one of the environments for uncontractibility. Thus, we consider the shallow slope of this item to potentially reflect sampling constraints when relying on unstructured conversation. Therefore, we also dropped V15, and the final fitting model only included V3 (prepositional phrase), V4 (copula linking two nouns), V6 (auxiliary BE, DO, HAVE), V7 (progressive *–ing*), V9–V14 (modal, third-person singular present, past tense modal, regular past tense, past tense aux, and medial adverb), V16 (past tense copula), and V17 (other bound morpheme). This model appeared to be an acceptable fit: $M_2 = 385.15$ ($df = 252$, $p < .01$, RMSEA = 0.04). The marginal reliability for the verbal phrase was 0.81, which would also be considered acceptable. The item parameter estimates are reported in Table A1 in the Appendix, and the ICCs and test information function are reported in Supplemental Materials S2–S4. In general, the final set of items exhibit fair and significant slopes (larger than 1).

Q/N Formation Ability

The Q/N subscale has a total of 11 items (Q1–Q11). Our analysis showed that this subscale exhibited the least desirable item properties in general. First, both Q1 (intonationally marked questions, signaled in our corpora simply by transcription of the question mark) and Q2 (*wh*-word alone or "routine question with or without a verb") were problematic because item slopes were very extreme (above 130) and made the estimation unstable. The examination of the frequency tables for these two items found that the distributions of the scores were almost identical between these two items; there were only two children out of 338 who scored 0, and four children achieved a score of 1 on these items. In other words, most of the children obtained a maximum score of 2 on these two items. Once Q1 and Q2 were dropped, Item Q4 (*wh*-question) and Item Q5 (negative morpheme between subject and verb) exhibited a similar problem of information redundancy and high slopes. Furthermore, the distribution of item scores for Q10 (child use of a tag question) was also extreme, in that only one child did not use any, but none produced two examples of Q10. Accordingly, a total of five items were dropped (Q1, Q2, Q4, Q5, and Q10), and only six items were left (Q3, Q6, Q7, Q8, Q9, and Q1: simple negation; *wh*-question with inversion of copula or modal; negation of copula, modal, or aux; yes/no question with inversion; *why, when, which, whose* questions; and question with negation and inversion, respectively). However, the reliability for this subscale was low at 0.71. The model fit assessment result was also not good: $M_2 = 110.24$ ($df = 54$, $p < .01$, RMSEA = 0.06). The item parameter estimates are reported in Table A1 in the Appendix, and the ICCs and test information function are reported in Supplemental Materials S2–S4. Although Items Q3, Q6, Q7, and Q11 have relatively high slopes, the other items, namely, Q8 (yes/no question with subject–auxiliary inversion) and Q9 (questions that specifically use *why, when, which, whose*), have much lower slopes.

We believe that there is a pragmatic reason for the instability of this subscale that we return to in the Discussion section. While structures within this subscale can be ordered in terms of their expected difficulty of grammatical development, it is not clear that all adult–child language samples provide equivalent opportunities for the structures in this subscale to be used in conversational interaction. Thus, a major difficulty with this subscale is whether or not a given child's sample can be reliably scored for expressive ability to construct these items without uniform elicitation procedures that assure that children will ask questions or create negative constructions.

S Assembly Ability

The S subscale has a total of 20 items (S1–S20). However, no child achieved a score of 2 on the last item, namely, S20 (full or truncated passive construction or tag comment/intrusion containing a clause), indicating that the item is the most difficult. Accordingly, the GRM with only two categories was fitted for this item. For the initial model, S1 (any two-word utterance) was excluded because all of the samples achieved a score of 2 on that item. Then, S2 (a subject–verb sequence) was subsequently dropped because the item exhibited extreme slope. After dropping these two items (S1 and S2), the marginal reliability of the response pattern score was 0.9, which is considered desirable.

[2]Not to be confused with contracted. Even an uncontracted copula or auxiliary, as in "He is nice/running," is contractible.

The model fit was also acceptable: $M_2 = 880.44$ ($df = 560$, $p < .01$, RMSEA = 0.05). For this subscale, most items exhibited significant slopes, suggestive of development over the preschool years, except for S9 (*let/make/help/watch* verb introducer) and S18 (gerund used as an NP). The ICC for S9 indicates that the item is too easy and that most of the children receive a score of 2 on this item, regardless of their ability or age level, whereas the ICC for S18 indicates that the item is too difficult and that most of the children received a score of 0 on this item, regardless of age. Either dropping or revising of these items may be advisable. See Table A1 in the Appendix for item parameter estimates and Supplemental Materials S2–S4 for ICCs and test information function.

Overall Productive Syntax Score

While the original total number of items on the IPSyn's subscales is 59 (NP = 11, VP = 17, Q/N = 11, and S = 20), items that were excluded following the subscale analyses were not included in the Overall Productive Syntax item analysis because those items did not contribute to subscale abilities, as discussed earlier. Therefore, a total of 42 items (six NP, 12 VP, six Q/N structures, and 18 S patterns) were included in this analysis; these appear as our suggested revision of the IPSyn in Table A1 in the Appendix. When a unidimensional GRM was fitted, generally, S and VP subscale items exhibited high discrimination, with the exceptions of S9, S18, and V15. Additionally, N11 as well as Q8, Q9, and Q11 also did not show significant slopes, indicating that these items are weakly associated with the Overall Productive Syntax. Therefore, these items could be further considered for elimination if a shorter IPSyn list of target structures is desired, without losing too much in score reliability. The marginal reliability for the Overall Productive Syntax with 42 items was 0.93. The model fit was acceptable: $M_2 = 1377$ ($df = 778$, $p < .01$, RMSEA = 0.05). The item parameter estimates are reported in Table A1 in the Appendix, and the ICCs and test information function are reported in Supplemental Materials S2–S4.

IPSyn Scale Scores Over Early Childhood

NP Assembly Ability

A linear–linear combination of segmented regression models fits well to represent the trajectory of NP assembly ability across ages ranging from 24 to 80[3] months. This implies that two separate trajectories can be identified before and after a "cut-point" in development. Figure 2 shows the fitted lines for the TD group and the LT group as well as for the HL and DS group samples as individual data points (circles and dots). Note that we did not fit regression lines for children who were deaf or hard of hearing or children with DS, as their sample sizes are small and there were not many data points across the full age range. It should be noted that none of the scores go beyond 1, which indicates a ceiling effect for the NP subscale. First, the estimated break point (age when the slopes change) of the two regression lines for the TD group was 27.4 months ($SE = 0.618$), which indicates that there is rapid development ($b = 0.31$, $SE = 0.08$, $p < .01$) in NP assembly ability for TD children up to 26–29 months of age (95% CI [26.21, 28.64]); after that age, the slope becomes notably flatter ($SE = 0.003$, $p < .01$). A similar pattern is observed for the LT group; however, their early-development slope is much flatter than that for the TD group ($b = 0.11$, $SE = 0.02$, $p < .01$), and the slope for this group changes at the age of 42 months ($SE = 1.93$) when it becomes very similar to the slope of the TD group ($b = 0.02$, $SE = 0.005$, $p < .01$). No children with DS reached the fitted lines, and most of the children with HL we analyzed did not achieve the TD or LT trajectory, with a few exceptions.

VP Assembly Ability

Similarly, a linear–linear combination of segmented regression models was fit for VP assembly ability. Figure 2 shows the fitted lines for the TD and LT groups, with the HD and DS group samples displayed as individual data points (triangles and squares). It should be noted that fitting a segmented regression model was only reasonable for the TD group; the LT group exhibited a constant slope across ages. The estimated break point of the two regression lines for the TD group was 30 months ($SE = 0.674$), which indicates that there is rapid development ($b = 0.21$, $SE = 0.03$, $p < .01$) in VP assembly ability for TD English-speaking children up to roughly 29–31 months of age (95% CI [31.40, 28.76]). After this age, the slope becomes noticeably flatter ($b = 0.04$, $SE = 0.004$, $p < .01$). In contrast, the LT group showed steady development, without the same rapid development stage seen in the TD cohort ($b = 0.05$, $SE = 0.005$, $p < .01$). Similar to NP assembly ability, the children's VP assembly ability in the DS and HL groups plots below the expected trajectories for either the TD or the LT group.

Q/N Formation Ability

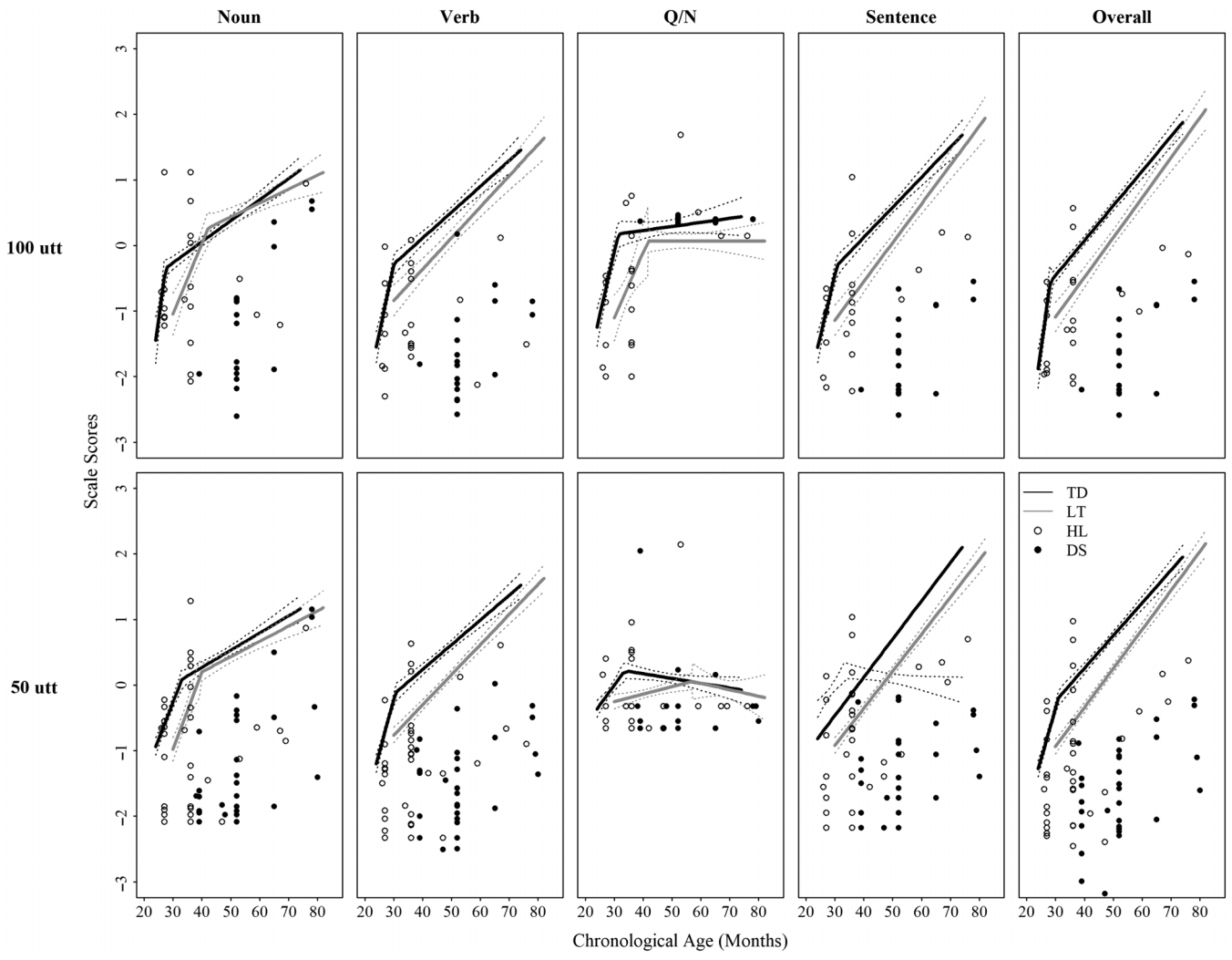
The fitted lines for the TD and LT groups and the data points for children with HL and DS are presented in Figure 2 for Q/N formation ability. The most noticeable phenomenon is that both trajectories become completely flat after a rapid development stage. For the TD group, the break point was 31.40 months ($SE = 0.85$), whereas the corresponding age was 10 months later for the LT group ($b = 41.686$, $SE = 2.848$). It should be also noted that these trajectories could not effectively discriminate the HL and DS groups.

S Assembly Ability

The trajectories of the S assembly ability for the TD and LT groups appeared to be very similar to those for VP

[3]Note that IRT analysis allows the use of different samples for calibration and scoring. While the age of the calibration sample was from 24 to 74 months, we included up to 82-month-old children in scoring samples (LT, HL, and DS) to be able to model the trajectory with better precision.

Figure 2. Fitted regression lines for typically developing children (TD; black lines) and children classified as late talkers (LT; gray lines). Q/N = Question/Negation; utt = utterances; HL = children with hearing loss; DS = children with Down syndrome.



assembly ability, as one can see in Figure 2. The segmented regression line was only well fitted to the TD group; the LT group demonstrated a constantly developing profile. The estimated break point of the two regression lines for the TD group was 28.1 months (SE = 0.68), which indicates that there is rapid development (b = 0.18, SE = 0.026, p < .01) in S assembly ability for TD children up to that point, and then, the slope becomes flatter (b = 0.05, SE = 0.004, p < .01) from that age point forward. In contrast, the LT group showed steady development without evidence of a rapid development stage (b = 0.05, SE = 0.005, p < .01). Children in neither the HL nor the DS group reached the expected trajectories of either the TD or the LT group.

Overall Productive Syntax Score

As the S and VP structure assembly ability items had high slopes in our item analysis, the score trajectory for the Overall Productive Syntax is very similar to the development of these two subscales. The estimated break point of the two regression lines for the TD group was 28.2 months (SE = 0.66), which indicates that there is rapid development (b = 0.32, SE = 0.067, p < .01) in Overall Productive Syntax for children who are presumed to show typical development, and then, the slope becomes flatter (b = 0.05, SE = 0.003, p < .01) over later age points. In contrast, as noted earlier, the LT group developed steadily over the full age range (b = 0.06, SE = 0.003, p < .01). Again, most children with HL did not reach the expected trajectories of the TD and LT groups in the corpora we analyzed.

Fifty-Utterance Sample Results

First, as expected, many more children's samples were available for analysis when we lowered the criterion to 50 eligible utterances rather than 100 (1,051 vs. 501). This illustrates that lowering the required sample length makes it more

likely that a child will provide a sample amenable to clinical LSA. Next, the psychometric properties of items using 50 utterances from TD children (n = 639) appear to be generally consistent with the results from 100-utterance samples, except for some items meant to tap Q/N ability, which showed inherent instability as discussed in a prior section. For the NP subscale, all samples achieved the maximum score of 2 on Items N1 and N3, whereas N1–N4 met this score in the 100-utterance samples. One expects to see more variation in item responses as the number of samples gets larger. Typically, a larger sample size mediates the problem of quasi-complete separation (see the Appendix for an explanation of this phenomenon) among item responses, so we expected to see fewer items with extremely high slopes. As a result, the final set of NP subscale items in the 50-utterance sample included two items more (N2 [pronoun] and N5 [article + noun]) than the 100-utterance sample, and this also yielded an interconsistency reliability of 0.69, a value slightly higher than that of the 100-utterance sample (0.61). Problematic features of N7 (plural) and N11 (other bound morpheme) were also seen in the 50-utterance sample.

As with NP items, V2 (particle or preposition) and V8 (adverb) can be re-included in the final model when a shorter, 50-utterance sample is used. Because item responses to V8 in this sample did not include any scores of 1, a two-category model (0 or 2) was used. However, the prior problem with V15 (ellipsis) was also seen in shorter samples, and so, V15 was dropped from our final model, although the reliability coefficient was slightly higher at 0.83 in contrast to 0.78.

Very similar phenomena were seen in S assembly ability items. All children achieved a maximum score of 2 on S1 (two words). As in the 100-utterance sample analysis, there were only two children who scored 1 on S20 (passive or tag question), and no one received 2 for this structure; hence, S20 was included in the calibration model as a dichotomous item. The information redundancy issue with S2 (subject + verb) was consistent. Therefore, the final model for S for a 50-utterance sample reduced to the same items that were used in the 100-utterance sample; the reliability was 0.91.

Unlike other subscales, the profiles of Q/N formation ability items using 50-utterance samples did not necessarily follow trends seen in 100-utterance samples, and two additional items (Q6 and Q7) suffered from the extreme-slope problem described previously and needed to be dropped for the final model. Accordingly, only Q3 (simple negation), Q8 (yes/no question with inversion), Q9 (use of *why, when, which, whose*), and Q11 (question with negation and inversion) were included for the final model for shorter samples. The low slopes on Items Q8 and Q9, however, were consistent between the samples at both lengths. The basic instability of Q/N formation ability items was again confirmed via this analysis and suggests a need for reconceptualizing items or elicitation procedures. Possible reasons for this will be discussed in a later section, but this subscale does not demonstrate a reliable age-based trend over early development.

Overall scale items for shorter samples were thus composed of a total of 43 items, as shown in Table A1 in the Appendix (eight NP, 13 VP, four Q/N structures, and 18 S patterns). To make our suggested revision easier to comprehend, we list the current IPSyn structures and use a strike-out function to remove those items that our analysis suggests do not contribute meaningfully to a child's developmental profile of expressive language use. Despite the slight changes in the item compositions, the score reliability coefficient for shorter, 50-utterance samples was similar to that derived from 100-utterance samples, at 0.93. All of the item parameter estimates from the 50-utterance sample are presented in Table A1 in the Appendix and depicted in Supplemental Materials S2–S4, along with the results from 100-utterance samples.

Before moving on to the developmental pattern analysis, we calculated correlations between the two sets of scale scores (one from 100-utterance sample calibration and one from 50-utterance sample calibration) for 388 children. The Pearson correlations were .80, .87, .49, .92, and .93 for NP, VP, Q/N, S, and Overall, respectively. As one can see, the correlations are fairly high between short and long samples, with the exception of the Q/N formation subscale, in which we observed inherent, item-based instability across age ranges and sample lengths.

In summary, our analysis suggests that using a 50-utterance criterion allows us to obtain more qualifying samples and results in scores similar to those that would have been obtained via 100 utterances. This adjustment, therefore, can make the IPSyn more available to less talkative children, with less use of clinician transcription time.

The same regression models (linear–linear) were fitted using the EAP scores based on the 50-utterance sample calibration results (see the trajectories in Figure 2). For NP assembly ability, the trajectory pattern is similar between the 100- and 50-utterance samples, but the estimated break point was 33.11 months (SE = 1.186; 95% CI [30.78, 35.43]) for the TD group, compared to 27.43 months when using longer language samples. With a 50-utterance criterion, the early-stage slope was flatter, and the change in slope occurred later compared to that seen using 100 utterances, which may add to clinical utility with older preschoolers. However, the LT group appears to show basically the same trajectories between the two sample lengths; their early-development slope is similar to that of the TD group (b = 0.18, SE = 0.02, $p < .01$), and the slope for this group changes at the age of 40 months (SE = 1.85) and becomes very similar to the slope of the TD group (b = 0.02, SE = 0.004, $p < .01$). For VP assembly ability, the fitted lines are very much alike between the 100- and 50-utterance samples for both TD and LT groups, including the estimated break point of the two regression lines for TD children (at 30 months of age; SE = 0.688) and the constant slope for the LT group (b = 0.05, SE = 0.003, $p < .01$). For Q/N formation ability, the shallow slope for early-stage development and plateaus were consistently observed in 50-utterance samples. For the TD group, the break point was 33.65 months (SE = 1.5), which overlaps with the 95% confidence interval of the change point in 100-utterance samples; the corresponding age for the LT group was 57 months with a substantial standard

error of 7.7, which indicates that it would be more reasonable to fit a linear line for the TD group. Regardless of the regression models, however, it was again confirmed that Q/N formation ability is not sensitive to chronological age and that trajectories could not effectively discriminate the HL and DS groups from TD children using either 100- or 50-utterance samples. Finally, similar to VP assembly ability, which had high score reliability, the Overall Productive Syntax score showed developmental trajectories using 50 utterances similar to those obtained using 100-utterance samples (see Figure 2).

In summary, using 50 eligible utterances allowed more children to be included for calibration, and the results were generally consistent for most subscales, as well as the overall score. With the exception of the inherently unstable Q/N formation ability scale (and its impact on the Overall Productive Syntax score), 50-utterance sample scores were highly correlated with those from 100 utterances, and the trajectory patterns across chronological age were consistent between the two sample lengths. Simply put, an IPSyn based on a shorter, more easily elicited and transcribed sample of 50 utterances shows much the same benefits as that conducted using a sample twice as long.

Finally, we examined the performance of each measure in separating TD and LT groups by plotting the expected distribution of scale scores and generating receiver operating characteristic curves with 50-utterance samples (see Figure 3). While NP assembly score was most effective in distinguishing 24- and 30-month-old children from the two groups, the false-positive rate (the probability of incorrectly diagnosing an LT when the child is not) is still 0.4, when the true-positive rate (the probability of correctly flagging an LT when the child is indeed an LT) target is 0.8. In contrast, VP assembly, S assembly, and Overall Productive Syntax scores were not informative in separating two groups at the age of 2 years. These scores were the most effective for children who are between 30 and 42 or 48 months of age. However, the false-positive rate is between 0.4 and 0.5, when a true-positive rate of 0.8 is targeted. Thus, it is not clear that we can obtain a cutoff score for either IPSyn subscales or total scores that reliably classify children as TD or delayed on an individual basis. With these caveats in mind, Tables A1 and A2 in the Appendix provide guidance in the interpretation of derived IPSyn scores, from a 50-utterance sample, in 1-year increments from 24 to 60 months, using our proposed revision—the IPSyn-C. These score interpretations are based on profiles observed in samples from 639 children who are presumed to be TD.

Discussion

We wish to start the discussion by noting that the IPSyn is a remarkably valuable tool that goes well beyond earlier measures of grammatical development, such as MLU, that were solely concerned with measuring the length of children's utterances—how much a child says rather than the variety of phrasal forms that the child can construct. IPSyn analysis can be incredibly useful in establishing treatment goals for children and for monitoring their progress in language learning. Its detailed analysis of specific structures in a child's expressive language provides both information about a child's pace of language development and numerous potential targets for intervention when an SLP identifies gaps in the child's syntactic repertoire. Our goal in conducting this exhaustive analysis is not to suggest that the IPSyn is not an appropriate measure of children's language development but, rather, to suggest ways in which the IPSyn might be modified to be easier for SLPs to both gather information and streamline its scoring (whether by hand or by computer) to identify the most sensitive items that can discriminate age-appropriate from delayed/disordered expressive language skills. Thus, we value the IPSyn and seek only to suggest ways in which its use can be made easier for practicing clinicians.

Streamlining the Number of IPSyn Items

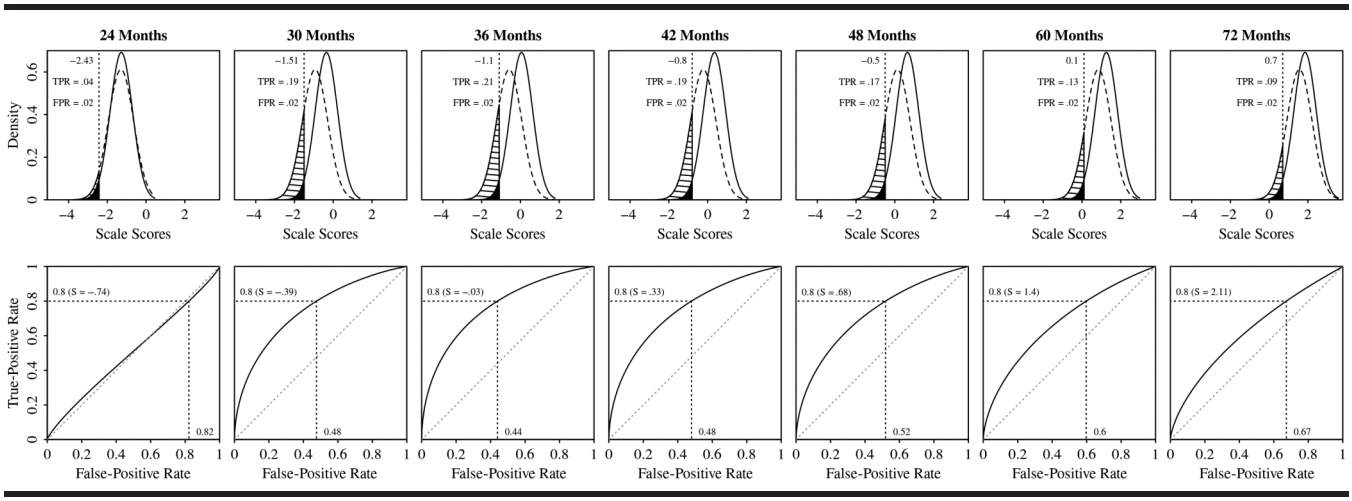
In this regard, our work suggests that not all items on the IPSyn are equally informative and that not all of its four subscales are equally predictive of language growth. As the IPSyn authors note in the construction of its scoring rules, many structures targeted by the IPSyn "cascade" into other structures; typically, if a child uses a construction considered to be more advanced, they receive credit for earlier phrase structures that are contained within the more elaborated constructions, something that its authors readily acknowledge (Altenberg et al., 2018). At a psychometric level of analysis, when the IPSyn is run on hundreds of presumably typical preschool children, this suggests the feasibility and appropriateness of culling the number of items scored to its most informative ones. This has the advantage of making either machine scoring or hand scoring of the IPSyn simpler and less prone to error. Our result is the proposed IPSyn-C.

Our analysis also suggests that the NP subscale is less informative than the VP and S subscales, primarily because many children "top out" on NP items on the IPSyn at relatively young ages. In this sense, if the SLP's goal is to identify children making less-than-optimal progress over the preschool age range, VP and S subscale items appear to present more challenge to the full range of young children we analyzed and, thus, would appear to be more discriminating in their ability to pinpoint language delay, as well as provide potential, well-delineated targets for intervention.

In this regard, our findings were not completely unexpected. For instance, Rispoli and Hadley (2001) found VP complexity to be uniquely predictive of fluency disruption in the expressive language of children between the ages of 2;6 and 4;0, suggesting that it does tap aspects of grammatical maturation beyond mere accuracy or variety of production. Classically, verb use has always been considered more difficult for young children to master in English than noun use (Gentner, 2006), and VP constructions embedded in the S subscale carry this concept further in tracking children's sentence assembly abilities over the preschool years.

Our analysis also raises potential cautions regarding the use of the Q/N subscale to benchmark children's language as age-appropriate, although the subscale can certainly be used

Figure 3. Density plots and receiver operating characteristic curves for the Overall Productive Syntax score. TPR = true-positive rate; FPR = false-positive rate.



to identify items for later prompting, assessment, and goal setting. When considering these items on a more pragmatic level, it seems that usage of many of the subscale items is not purely linguistic or obligatory in the conventional sense. Use of questions and negation is likely to be highly dependent upon the context of the adult–child interaction and even, perhaps, the child's personality when engaged in joint play, the setting for the language samples examined here (and typical of SLP sample elicitation procedures). We could not identify reliable age-related trajectories in their use in naturalistic adult–child play. It seems to us that the assessment of a child's ability to form negation and questions may require a more standardized set of probes, such as those used by the Rice/Wexler Test of Early Grammatical Impairment (TEGI). TEGI materials are now openly available for download at https://cldp.ku.edu/rice-wexler-tegi.

While there are reference scores available to benchmark children's performance on the IPSyn, our work suggests that when examined using a much larger cohort of samples than the original ones studied by Scarborough (1990), there is a time window for the use of the IPSyn in classifying individual children as functioning outside of typical age expectations. Furthermore, scores obtained from younger children or children functioning at lower levels of expressive language performance are more robust indicators of developmental status than scores obtained from older, more advanced children. The cut-point at which trajectories of language growth slow markedly is quite early for typical children, below 3 years of age. This is true for the NP and VP subscales as well as the S subscale. Given the level of detail in IPSyn analysis, this is somewhat surprising. However, for clinical purposes, an IPSyn analysis conducted on a younger child is more informative, for intervention purposes, than an alternative such as MLU. Even if MLU can also distinguish typical from delayed expressive language profiles, it leaves the clinician with little specific guidance on structuring language therapy, other than to guide the child to "say longer utterances." In contrast, the IPSyn provides a tangible list of structures that, if not observed in the IPSyn profiling process, make excellent targets for intervention.

Shortening the Language Sample

One possible way to make the IPSyn faster and easier to conduct is to premise it on a shorter language sample. This makes the time required to gather and transcribe the sample shorter, as well as the analysis time, if it is conducted by hand rather than automated. We have also found, in clinical practice, that younger and more delayed talkers are also less voluble during play interactions (Rescorla & Bernstein Ratner 1996); concretely, this has often led to difficulty in gathering a sample long enough to fulfill the requirements for IPSyn analysis. Our work with this large set of language samples suggests that it is feasible to adopt psychometric properties for the IPSyn on smaller language samples. For the samples we utilized, this actually "grew" the statistical data set considerably, and thus, we consider our analysis in this regard to be rather robust. Samples with 50 utterances have much the same psychometric properties as samples containing 100 utterances. We think this may be the most important of our findings, since it should encourage more clinicians to employ IPSyn analysis in their evaluation of children's expressive language profiles.

Using Computer-Assisted LSA

Before leaving this section, we wish to acknowledge that clinicians may not relish the process of coding for even the reduced number of IPSyn structures on a shorter sample. We undertook this analysis using a free, open-access computer program (CLAN; available for both Mac and PC) that does not require the clinician to tag or label any aspect of the child's sample—the program itself performs the grammatical analysis before conducting the IPSyn analysis. Thus, the assessment, therapy planning, and therapy monitoring potential of the IPSyn is available to any clinician

who creates a simple transcript of the child's language using conventionalized spelling, capitalization, and punctuation. Both written (Overton et al., 2020) and video (https://talkbank.org/screencasts/0SLP/) tutorials are available to guide clinicians in using computer-assisted IPSyn and other LSA measures.

Limitations

Diversity of the Child Language Samples

First and foremost, the children represented in the CHILDES archive tend not to reflect the diversity of socioeconomic, linguistic, and cultural backgrounds that characterize most SLP caseloads. It is clear that middle-class, GAE-speaking children are overrepresented in our data set. Having said this, some of the individual corpora, such as HHSLD and Van Houten, did enroll children from low–socioeconomic status households. Work is clearly needed to replicate our findings with children across a variety of social and linguistic communities. In our laboratory, we are exploring IPSyn applicability to play samples gathered from children speaking AAE (Overton et al., 2020).

All of the samples used for analysis were, as specified in the Method section, adult–child play sessions, almost uniformly between the children and their mothers, although some corpora also included samples gathered while the child played with an investigator. As such, our findings cannot be generalized to analysis of other genres of verbal behavior, such as narratives or narrative retells, or interviews.

It is also possible that some cohorts of children within a given data set (such as an individual researcher's project contributed to CHILDES) could have common features (such as a data collection methodology that encourages use or avoidance of some structures because of toys/activities used to elicit the samples). If so, the resulting data would have a nested structure, and a multilevel IRT model would be needed for appropriate analysis. In the absence of such expectations, we decided to use a single-level IRT model given the sample size. A future approach to evaluating this possibility would be to analyze cohorts of longitudinal data obtained from individual children (also possible to do using CHILDES corpora), which would obviate the potential for a cross-sectional language study using an unvarying language sample elicitation approach to identify nested behaviors we were unable to detect in the current analysis.

Next, we also acknowledge that the four subdomains of the IPSyn are correlated with each other and that other multidimensional models (e.g., independent cluster models) can be used to investigate the psychometric properties of the IPSyn. However, the current sample size was not sufficient to conduct a four-dimensional model analysis. Future work that includes the substantial number of longitudinal corpora in CHILDES might be able to pursue such an approach.

Conclusions

An initial goal of our work was to employ IPSyn analysis on a much larger sample of children than previously used, in order to obtain robust reference scores that can be used to interpret IPSyn scores in clinical practice. We have obtained such scores and offer them to clinicians in the Appendix of this article as an adaptation of the IPSyn that we call the IPSyn-C. As we examined profiles obtained from hundreds of TD preschool English language speakers, as well as those obtained from children with language delay and children with HL or DS, we discovered response patterns that offered an opportunity to simplify IPSyn language sample elicitation, transcription, and scoring. Simply put, the proposed IPSyn-C can provide its excellent guidance to the practicing clinician with fewer eligible utterances and fewer scoring categories. Both findings should encourage its wider use, particularly if clinicians perform the required analyses using computer-assisted free software, such as TalkBank's CLAN program, or an equivalent utility. Interpreting score results should be more reliable.

One unexpected caution that arose from our analyses is the inclusion of Q/N structures on any "checklist" analysis of expressive language corpora. Children's use of such structures in free-play with adults appears to be unpredictable, and the absence of such structures does not then imply that they are absent from the child's skill set but, rather, absent from motivated use in conversation. We note that DSS (Lee, 1974) also includes a similar subscale that may be vulnerable to these concerns. We suspect that the field would be aided by further development of structured probes that elicit Q/N constructions from young children in sufficient quantity and scope to guide both clinical assessment and goal setting. Aside from stipulating a desired length of the language sample, clinical guidance is rather vague on uniform "best practices" for ensuring a diversity of potential language structures in a child's interaction with an adult during the data elicitation process.

Finally, it should not be surprising that children's mastery of some English sentence structure components does not proceed in a completely linear fashion. In this regard, we observed that growth in IPSyn-targeted constructions is very rapid early in development, making it easier to discriminate younger children who are "on target" developmentally from those who are not. As children age and their repertoires enlarge, IPSyn score slopes slow, making the difference between typical and delayed/impaired language sample results somewhat more subtle. Future work may wish to consider whether the current numerical "cap" on two exemplars for each targeted IPSyn structure could be adjusted to enable a more distinct developmental trajectory that does not asymptote over development. Having said this, the IPSyn's benefits to clinical practice in child language remediation go well beyond its role in diagnosis of impairment and include valuable guidance on potential targets for remediation, as well as monitoring of language growth over the course of intervention.

Acknowledgments

This work was supported by National Institute on Deafness and Other Communication Disorders Grant R01DC017152, awarded to Nan Bernstein Ratner.

References

Altenberg, E. P., & Roberts, J. A. (2016). Promises and pitfalls of machine scoring of the Index of Productive Syntax. *Clinical Linguistics & Phonetics, 30*(6), 433–448. https://doi.org/10.3109/02699206.2016.1139184

Altenberg, E. P., Roberts, J. A., & Scarborough, H. S. (2018). Young children's structure production: A revision of the Index of Productive Syntax. *Language, Speech, and Hearing Services in Schools, 49*(4), 995–1008. https://doi.org/10.1044/2018_LSHSS-17-0092

Bernstein Ratner, N., & MacWhinney, B. (2016). Your laptop to the rescue: Using the Child Language Data Exchange System archive and CLAN utilities to improve child language sample analysis. *Seminars in Speech and Language, 37*(2), 74–84. https://doi.org/10.1055/s-0036-1580742

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443–459. https://doi.org/10.1007/BF02293801

Brown, R. (1973). *A first language*. Harvard University Press. https://doi.org/10.4159/harvard.9780674732469

Cai, L. (2017). *flexMIRT®: Flexible multilevel multidimensional item analysis and test scoring* (Version 3.51) [Computer software]. Vector Psychometric Group.

Crystal, D., Fletcher, P., & Garman, M. (1981). *Grammatical analysis of language disability*. Cole and Whurr.

Eigsti, I. M., & Cicchetti, D. (2004). The impact of child maltreatment on expressive syntax at 60 months. *Developmental Science, 7*(1), 88–102. https://doi.org/10.1111/j.1467-7687.2004.00325.x

Estigarribia, B., Martin, G. E., & Roberts, J. E. (2012). Cognitive, environmental, and linguistic predictors of syntax in Fragile X syndrome and Down syndrome. *Journal of Speech, Language, and Hearing Research, 55*(6), 1600–1612. https://doi.org/10.1044/1092-4388(2012/10-0153)

Finestack, L. H., Rohwer, B., Hilliard, L., & Abbeduto, L. (2020). Using Computerized Language Analysis to evaluate grammatical skills. *Language, Speech, and Hearing Services in Schools, 51*(2), 184–204. https://doi.org/10.1044/2019_LSHSS-19-00032

Finestack, L. H., & Satterlund, K. E. (2018). Current practice of child grammar intervention: A survey of speech-language pathologists. *American Journal of Speech-Language Pathology, 27*(4), 1329–1351. https://doi.org/10.1044/2018_AJSLP-17-0168

Geers, A. E., & Sedey, A. L. (2011). Language and verbal reasoning skills in adolescents with 10 or more years of cochlear implant experience. *Ear and Hearing, 32*(1), 39S–48S. https://doi.org/10.1097/AUD.0b013e3181fa41dc

Gentner, D. (2006). Why verbs are hard to learn. Action meets word: How children learn verbs. In K. Hirsh-Pasek & R. M. Golinkoff (Eds.), *Action meets word: How children learn verbs* (pp. 544–564). Oxford University Press.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*(4), 347–360. https://doi.org/10.1111/j.1745-3984.1984.tb01039.x

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*(2), 139–150. https://doi.org/10.2307/2086306

Hadley, P. A. (1998). Early verb-related vulnerability among children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 41*(6), 1384–1397. https://doi.org/10.1044/jslhr.4106.1384

Hadley, P. A., Rispoli, M., & Hsu, N. (2016). Toddlers' verb lexicon diversity and grammatical outcomes. *Language, Speech, and Hearing Services in Schools, 47*(1), 44–58. https://doi.org/10.1044/2015_LSHSS-15-0018

Hassanali, K. N., Liu, Y., Iglesias, A., Solorio, T., & Dollaghan, C. (2014). Automatic generation of the index of productive syntax for child language transcripts. *Behavior Research Methods, 46*(1), 254–262. https://doi.org/10.3758/s13428-013-0354-x

Hewitt, L. E., Hammer, C. S., Yont, K. M., & Tomblin, J. B. (2005). Language sampling for kindergarten children with and without SLI: Mean length of utterance, IPSYN, and NDW. *Journal of Communication Disorders, 38*(3), 197–213. https://doi.org/10.1016/j.jcomdis.2004.10.002

Hoe, S. L. (2008). Issues and procedures in adopting structural equation modeling technique. *Journal of Quantitative Methods, 3*(1), 76–83.

Horton-Ikard, R., Weismer, S. E., & Edwards, C. (2005). Examining the use of standard language production measures in the language samples of African-American toddlers. *Journal of Multilingual Communication Disorders, 3*(3), 169–182. https://doi.org/10.1080/14769670500170768

Justice, L. M., & Ezell, H. (1999). Syntax and speech-language pathology graduate students: Performance and perceptions. *Contemporary Issues in Communication Science and Disorders, 26*(Fall), 119–127. https://doi.org/10.1044/cicsd_26_F_119

Lee, L. L. (1974). *Developmental sentence analysis: A grammatical assessment procedure for speech and language clinicians*. Northwestern University Press.

Long, S. H. (2001). About time: A comparison of computerized and manual procedures for grammatical and phonological analysis. *Clinical Linguistics & Phonetics, 15*(5), 399–426. https://doi.org/10.1080/02699200010027778

Long, S. H., & Channell, R. W. (2001). Accuracy of four language analysis procedures performed automatically. *American Journal of Speech-Language Pathology, 10*(2), 180–188. https://doi.org/10.1044/1058-0360(2001/017)

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, Volume II: The database*. Erlbaum.

MacWhinney, B., Roberts, J. A., Altenberg, E. P., & Hunter, M. (2020). Improving automatic IPSyn coding. *Language, Speech, and Hearing Services in Schools, 51*(4), 1187–1189. https://doi.org/10.1044/2020_LSHSS-20-00090

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2^n contingency tables. *Journal of the American Statistical Association, 100*(471), 1009–1020. https://doi.org/10.1198/016214504000002069

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*(4), 305–328. https://doi.org/10.1080/00273171.2014.911075

Oetting, J. B., Cantrell, J. P., & Horohov, J. E. (1999). A study of specific language impairment (SLI) in the context of nonstandard dialect. *Clinical Linguistics & Phonetics, 13*(1), 25–44. https://doi.org/10.1080/026992099299220

Oetting, J. B., Newkirk, B. L., Hartfield, L. R., Wynn, C. G., Pruitt, S. L., & Garrity, A. W. (2010). Index of Productive Syntax for children who speak African American English. *Language, Speech, and Hearing Services in Schools, 41*(3), 328–339. https://doi.org/10.1044/0161-1461(2009/08-0077)

Overton, C., Pearson, B. Z., & Bernstein Ratner, N. (2020). Computer-assisted and dialect neutral child language sample analysis. *Language, Speech, and Hearing Services in Schools, 52*(1), 31–50. https://doi.org/10.1044/2020_LSHSS-19-00107

Pezold, M. J., Imgrund, C. M., & Storkel, H. L. (2020). Using computer programs for language sample analysis. *Language, Speech, and Hearing Services in Schools, 51*(1), 103–114. https://doi.org/10.1044/2019_LSHSS-18-0148

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rescorla, L., & Bernstein Ratner, N. (1996). Phonetic profiles of toddlers with specific expressive language impairment (SLI-E). *Journal of Speech and Hearing Research, 39*(1), 153–165. https://doi.org/10.1044/jshr.3901.153

Rescorla, L., & Turner, H. L. (2015). Morphology and syntax in late talkers at age 5. *Journal of Speech, Language, and Hearing Research, 58*(2), 434–444. https://doi.org/10.1044/2015_JSLHR-L-14-0042

Rice, M. L., Redmond, S. M., & Hoffman, L. (2006). *Mean length of utterance in children with specific language impairment and in younger control children shows concurrent validity and stable and parallel growth trajectories*. University of Nebraska–Lincoln, Faculty Publications, Department of Psychology. https://digitalcommons.unl.edu/psychfacpub/435

Rispoli, M., & Hadley, P. (2001). The leading-edge. *Journal of Speech, Language, and Hearing Research, 44*(5), 1131–1143. https://doi.org/10.1044/1092-4388(2001/089)

Roberts, J. A., Altenberg, E. P., & Hunter, M. (2020). Machine-scored syntax: Comparison of the CLAN automatic scoring program to manual scoring. *Language, Speech, and Hearing Services in Schools, 51*(2), 479–493. https://doi.org/10.1044/2019_LSHSS-19-00056

Roberts, J. E., Hennon, E. A., Price, J. R., Dear, E., Anderson, K., & Vandergrift, N. A. (2007). Expressive language during conversational speech in boys with fragile X syndrome. *American Journal on Mental Retardation, 112*(1), 1–17. https://doi.org/10.1352/0895-8017(2007)112[1:ELDCSI]2.0.CO;2

Sagae, K., Lavie, A., & MacWhinney, B. (2005, June). Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (pp. 197–204).

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded score* (Psychometric Monograph No. 17). Psychometric Society. http://www.psychometrika.org/journal/online/MN17.pdf

Sanchez, K., Spittle, A. J., Boyce, J. O., Leembruggen, L., Mantelos, A., Mills, S., Mitchell, N., Neil, E., John, M. S., Treloar, J., & Morgan, A. T. (2020). Conversational language in 3-year-old children born very preterm and at term. *Journal of Speech, Language, and Hearing Research, 63*(1), 206–215. https://doi.org/10.1044/2019_JSLHR-19-00153

Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics, 11*(1), 1–22. https://doi.org/10.1017/S0142716400008262

Scarborough, H. S., Rescorla, L., Tager-Flusberg, H., Fowler, A. E., & Sudhalter, V. (1991). The relation of utterance length to grammatical complexity in normal and language-disordered groups. *Applied Psycholinguistics, 12*(1), 23–46. https://doi.org/10.1017/S014271640000936X

Singh, K., Junnarkar, M., & Kaur, J. (2016). *Measures of positive psychology, development and validation*. Springer. https://doi.org/10.1007/978-81-322-3631-3

Stockman, I. J., Newkirk-Turner, B. L., Swartzlander, E., & Morris, L. R. (2016). Comparison of African American children's performances on a minimal competence core for morphosyntax and the Index of Productive Syntax. *American Journal of Speech-Language Pathology, 25*(1), 80–96. https://doi.org/10.1044/2015_AJSLP-14-0207

Templin, M. C. (1957). *Certain language skills in children: Their development and interrelationships*. University of Minnesota Press.

Thissen, D., & Steinberg, L. (2009). Item response theory. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 148–177). Sage. https://doi.org/10.4135/9780857020994.n7

Thissen, D. & Wainer, H. (Eds.). (2001). *Test scoring*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410604729

Tomblin, J. B., Spencer, L., Flock, S., Tyler, R., & Gantz, B. (1999). A comparison of language achievement in children with cochlear implants and children using hearing aids. *Journal of Speech, Language, and Hearing Research, 42*(2), 497–511. https://doi.org/10.1044/jslhr.4202.497

Washington, K., Fritz, K., Crowe, K., Kelly, B., & Karem, R. (2019). Bilingual preschoolers' spontaneous productions: Considering Jamaican Creole and English. *Language, Speech, and Hearing Services in Schools, 50*(2), 179–195. https://doi.org/10.1044/2018_LSHSS-18-0072

Appendix (p. 1 of 3)

Revised Items and Statistical Properties of the Proposed IPSyn-C

Table A1. Item parameter estimates and standard errors for overall scale items.

		Overall scale calibration											
		100-utt sample (*N* = 388)						50-utt sample (*N* = 639)					
Item	Description	a (*SE*)		c_1 (*SE*)		c_2 (*SE*)		a (*SE*)		c_1 (*SE*)		c_2 (*SE*)	
N1	**Noun**			*						*			
N2	Pronoun			*				3.8	(1.5)	11.8	(3.8)	8.9	(3)
N3	**Modifier**			*						*			
N4	**Two-word NP**			*						**			
N5	Det + N			**				1.6	(0.3)	4.4	(0.4)	3.2	(0.3)
N6	V + 2-wd NP	2.3	(1.8)	6.9	(3.8)	5.9	(2.5)	3.1	(0.5)	5.6	(0.6)	4.0	(0.5)
N7	N plural	1.0	(0.4)	4.0	(0.6)	2.9	(0.4)	0.7	(0.1)	2.4	(0.2)	1.1	(0.1)
N8	2-wd NP + V	1.5	(0.3)	1.7	(0.3)	0.5	(0.2)	1.4	(0.1)	0.3	(0.1)	–1.1	(0.1)
N9	3-wd NP	1.6	(0.3)	2.9	(0.4)	1.4	(0.3)	1.5	(0.1)	0.9	(0.1)	–0.7	(0.1)
N10	NP Adv	1.0	(0.2)	0.4	(0.2)	–1.3	(0.2)	0.8	(0.1)	–0.7	(0.1)	–2.6	(0.2)
N11	Bound morpheme	0.4	(0.2)	0.4	(0.2)	–0.1	(0.2)	0.4	(0.1)	–0.6	(0.1)	–1.1	(0.1)
V1	**Verb**			*						*			
V2	V part or prep			*				2.5	(0.5)	5.4	(1.5)	4.8	(1.7)
V3	Prep phr	2.1	(1.6)	7.1	(4.1)	5.2	(2.1)	2.3	(0.4)	4.5	(1.4)	2.7	(0.4)
V4	N + cop + N	1.9	(1.3)	4.7	(1.4)	4.2	(1.4)	1.5	(0.2)	2.6	(0.2)	2.1	(0.2)
V5	**Catenative**			**						**			
V6	Aux BE, DO, HAVE	2.6	(0.7)	4.7	(0.8)	4.1	(0.7)	1.9	(0.2)	1.9	(0.2)	1.3	(0.2)
V7	Prog *–ing*	1.0	(0.3)	3.3	(0.4)	1.8	(0.2)	0.7	(0.1)	1.4	(0.1)	0.0	(0.1)
V8	Adverb			**				2.4	(1.7)	–7.9	(3.2)	NA	
V9	Modal + V	1.7	(0.3)	2.4	(0.3)	1.0	(0.2)	1.4	(0.2)	0.6	(0.1)	–0.7	(0.1)
V10	3rd-pers sing pres	1.3	(0.2)	2.2	(0.3)	0.7	(0.2)	1.2	(0.1)	0.6	(0.1)	–1.1	(0.1)
V11	Past tns modal	1.1	(0.2)	–0.3	(0.2)	–1.5	(0.2)	1.1	(0.2)	–1.5	(0.2)	–2.8	(0.2)
V12	Reg past	1.1	(0.2)	0.4	(0.2)	–1.0	(0.2)	1.2	(0.2)	–1.0	(0.1)	–2.7	(0.2)
V13	Past aux	1.8	(0.3)	0.4	(0.2)	–1.8	(0.3)	1.2	(0.2)	–1.0	(0.1)	–3.1	(0.2)
V14	Medial Adv	1.9	(0.4)	3.4	(0.4)	1.8	(0.3)	1.6	(0.2)	1.2	(0.2)	–0.3	(0.1)
V15	**Ellipsis**			#						#			
V16	Past cop	1.9	(0.3)	–1.0	(0.3)	–2.1	(0.3)	1.9	(0.3)	–2.5	(0.3)	–3.7	(0.4)
V17	Bound morpheme	1.1	(0.3)	–2.2	(0.3)	–4.1	(0.6)	1.1	(0.2)	–3.0	(0.3)	–4.6	(0.5)
Q1	**Intonation**			**						**			
Q2	**Routine**			**						**			
Q3	Simple Neg	1.7	(0.5)	4.8	(0.7)	2.7	(0.4)	1.4	(0.2)	2.3	(0.2)	1.0	(0.1)
Q4	***Wh*-Q + V**			**						**			
Q5	**S + neg + V**			**						**			
Q6	***Wh*-Q with S-Aux inversion**	4.2	(5.8)	10.9	(11.8)	9.8	(11.8)			**			
Q7	**Neg cop, modal or aux**	1.3	(0.3)	2.1	(0.3)	0.8	(0.2)			**			
Q8	Y/N Q with S-Aux inversion	0.3	(0.2)	0.4	(0.1)	–0.7	(0.2)	0.2	(0.1)	–0.6	(0.1)	–2.0	(0.1)
Q9	*Why, when, which, whose*	0.4	(0.3)	–1.5	(0.2)	–4.1	(0.5)	0.4	(0.3)	–2.4	(0.2)	–5.9	(0.9)
Q10	**Tag Q**			*						*			
Q11	Neg Q with S-Aux inv.	0.5	(0.4)	–2.5	(0.3)	–4.0	(0.6)	0.3	(0.4)	–3.4	(0.3)	–5.8	(0.8)
S1	**Two words**			*						*			
S2	**Subj + V**			**						**			
S3	V + Object	2.4	(1.9)	7.3	(2.9)	5.8	(2.4)	1.2	(0.3)	3.1	(0.4)	1.9	(0.3)
S4	S + V + O	3.3	(1.3)	7.2	(2.4)	5.6	(1.7)	1.4	(0.3)	2.6	(0.3)	1.4	(0.2)
S5	Any conjunction	2.9	(1.1)	5.3	(1.1)	4.2	(0.9)	2.5	(0.3)	2.4	(0.2)	1.2	(0.2)
S6	Any 2 V's	3.6	(1.5)	8.1	(2.3)	6.6	(1.9)	2.5	(0.4)	3.4	(0.3)	2.3	(0.3)
S7	Conjoined phrases	1.5	(0.3)	2.0	(0.3)	0.7	(0.2)	1.7	(0.2)	0.1	(0.2)	–1.2	(0.2)
S8	Infinitive	1.9	(0.3)	1.1	(0.3)	0.0	(0.3)	1.4	(0.2)	–0.1	(0.1)	–1.3	(0.2)
S9	*Let/Make/Help/Watch*	–0.1	(0.2)	–0.1	(0.1)	–1.5	(0.2)	0.1	(0.1)	–1.1	(0.1)	–2.7	(0.2)
S10	Subordinating conj.	3.7	(0.7)	1.6	(0.4)	0.1	(0.4)	3.3	(0.4)	–0.7	(0.3)	–2.3	(0.3)
S11	Mental state V	1.2	(0.2)	–0.9	(0.2)	–2.6	(0.3)	1.4	(0.2)	–2.0	(0.2)	–4.0	(0.3)
S12	Conjoined clauses	2.9	(0.4)	1.0	(0.3)	–0.7	(0.3)	3.7	(0.4)	–1.8	(0.3)	–3.6	(0.4)
S13	*If* or *wh*-clause	2.3	(0.4)	0.3	(0.3)	–1.2	(0.3)	2.2	(0.3)	–1.4	(0.3)	–3.2	(0.3)

Appendix A (p. 2 of 3)

Revised Items and Statistical Properties of the Proposed IPSyn-C

Table A1. *(Continued).*

Item	Description	Overall scale calibration											
		100-utt sample (*N* = 388)						50-utt sample (*N* = 639)					
		a (*SE*)		c_1 (*SE*)		c_2 (*SE*)		a (*SE*)		c_1 (*SE*)		c_2 (*SE*)	
S14	Bitransitive pred.	0.5	(0.2)	–1.4	(0.2)	–3.3	(0.4)	0.5	(0.2)	–2.3	(0.2)	–4.4	(0.4)
S15	3 or more V's	1.6	(0.2)	0.0	(0.2)	–1.6	(0.2)	2.1	(0.2)	–1.8	(0.2)	–3.7	(0.3)
S16	Relative clause	1.9	(0.3)	0.3	(0.2)	–1.2	(0.3)	1.8	(0.2)	–1.2	(0.2)	–2.8	(0.2)
S17	Infinitival clause	0.8	(0.3)	–1.4	(0.2)	–3.2	(0.3)	0.9	(0.2)	–2.3	(0.2)	–5.5	(0.6)
S18	Gerund	0.3	(0.2)	–1.2	(0.2)	–3.6	(0.4)	0.4	(0.2)	–2.3	(0.2)	–4.4	(0.4)
S19	Left or center-embed clause	1.8	(0.4)	–1.6	(0.3)	–3.4	(0.5)	1.9	(0.3)	–2.9	(0.3)	–5.3	(0.5)
S20	Passive or tag	2.48	(2.6)	–7.96	(5.2)	NA		2.36	(1.7)	–7.86	(3.2)	NA	

Note. For more detailed individual item descriptions, please see Altenberg et al. (2018). **Bolded** items were removed after statistical analysis. Table data designators: * = item dropped due to little variability in item responses; ** = item dropped due to unstable model estimation due to quasi-complete separation; # = item dropped due to insignificant slope; NA = items that have only two-category responses. All dropped items for 50-utterance samples are also **bolded**. utt = utterance; a = slope; c_1 = Intercept 1; c_2 = Intercept 2; NP = Noun Phrase; Det + N = determiner + noun; V = verb; 2-wd = two-word; 3-wd = three-word; Adv = adverb; V part or prep = verb particle or preposition; Prep phr = prepositional phrase; cop = copula; Aux/aux = auxiliary; Prog = progressive; 3rd pers sing pres = third-person singular present; Past tns modal = past tense modal; Reg past = regular past tense; Past aux = past tense auxiliary; Neg/neg = negation; *Wh*-Q = *wh*-question; S = subject; S-Aux = subject–auxiliary; Y/N Q = yes/no question; Tag Q = tag question; Neg Q = question with negation; inv. = inversion; Subj = subject; O = object; conj. = conjunction; pred. = predicate.

Table A2. A summed score conversion table for the proposed IPSyn-C in yearly intervals for ages 2–6 years (see Table A1).

Summed			Percentile rank						Summed			Percentile rank					
Score	T score	*SE*	24–72 mos.	24 mos.	36 mos.	48 mos.	60 mos.	72 mos.	Score	T score	*SE*	24–72 mos.	24 mos.	36 mos.	48 mos.	60 mos.	72 mos.
0	14.8	4.3	0	0	**0**	**0**	0	0	43	54.1	2.7	66	100	73	**33**	7	1
1	17.0	4.2	0	0	**0**	**0**	0	0	44	54.8	2.7	68	100	77	**38**	9	1
2	18.7	4.1	0	0	**0**	**0**	0	0	45	55.5	2.7	71	100	80	**43**	11	1
3	20.3	3.9	0	0	**0**	**0**	0	0	46	56.2	2.7	73	100	84	**48**	14	2
4	21.8	3.7	0	0	**0**	**0**	0	0	47	56.9	2.7	75	100	86	52	16	2
5	23.1	3.5	0	1	**0**	**0**	0	0	48	57.7	2.7	78	100	89	58	20	3
6	24.3	3.4	1	1	**0**	**0**	0	0	49	58.4	2.7	80	100	91	63	24	4
7	25.5	3.3	1	2	**0**	**0**	0	0	50	59.1	2.7	82	100	93	67	28	5
8	26.6	3.2	1	3	**0**	**0**	0	0	51	59.8	2.8	84	100	95	71	32	6
9	27.6	3.1	1	5	**0**	**0**	0	0	52	60.6	2.8	86	100	96	76	37	8
10	28.6	3.1	2	7	**0**	**0**	0	0	53	61.3	2.8	87	100	97	79	41	10
11	29.5	3.0	2	9	**0**	**0**	0	0	54	62.1	2.9	89	100	98	83	47	13
12	30.4	3.0	2	12	**0**	**0**	0	0	55	62.8	2.9	90	100	98	86	52	16
13	31.3	2.9	3	15	**0**	**0**	0	0	56	63.6	3.0	91	100	99	89	57	20
14	32.2	2.9	4	19	**0**	**0**	0	0	57	64.4	3.0	93	100	99	91	63	24
15	33.0	2.9	4	23	**0**	**0**	0	0	58	65.3	3.1	94	100	99	94	68	29
16	33.8	2.8	5	28	**0**	**0**	0	0	59	66.1	3.1	95	100	100	95	73	34
17	34.6	2.8	6	32	**0**	**0**	0	0	60	67.0	3.2	96	100	100	96	78	40
18	35.4	2.8	7	37	**0**	**0**	0	0	61	67.9	3.3	96	100	100	98	82	46
19	36.2	2.8	8	43	**1**	**0**	0	0	62	68.8	3.4	97	100	100	98	86	52
20	**37.0**	**2.8**	**10**	**48**	**1**	**0**	**0**	**0**	63	69.7	3.5	98	100	100	99	89	58
21	37.8	2.8	11	54	**1**	**0**	0	0	64	70.7	3.6	98	100	100	99	92	65
22	38.5	2.8	13	59	**2**	**0**	0	0	65	71.7	3.7	98	100	100	100	94	71
23	39.3	2.8	14	64	**3**	**0**	0	0	66	72.8	3.8	99	100	100	100	96	77
24	40.1	2.8	16	68	**3**	**0**	0	0	67	73.8	3.9	99	100	100	100	97	82
25	40.8	2.8	18	73	**5**	**0**	0	0	68	75.0	4.0	99	100	100	100	98	87
26	41.6	2.8	20	78	**6**	**0**	0	0	69	76.1	4.1	100	100	100	100	99	91
27	42.3	2.8	22	81	**8**	**1**	0	0	70	77.3	4.2	100	100	100	100	99	94
28	43.1	2.8	25	85	**10**	**1**	0	0	71	78.6	4.3	100	100	100	100	100	96
29	43.9	2.8	27	87	**12**	**1**	0	0	72	79.8	4.5	100	100	100	100	100	97
30	44.6	2.8	29	90	**15**	**2**	0	0	73	81.1	4.6	100	100	100	100	100	99
31	45.4	2.8	32	92	**18**	**3**	0	0	74	82.4	4.7	100	100	100	100	100	99
32	46.1	2.8	35	94	**22**	**3**	0	0	75	83.7	4.8	100	100	100	100	100	100
33	46.8	2.8	37	95	**26**	**5**	0	0	76	84.9	4.9	100	100	100	100	100	100
34	47.6	2.8	41	96	**30**	**6**	0	0	77	86.1	5.0	100	100	100	100	100	100
35	48.3	2.8	43	97	**35**	**8**	1	0	78	87.3	5.1	100	100	100	100	100	100
36	49.1	2.8	46	98	39	**9**	1	0	79	88.3	5.2	100	100	100	100	100	100
37	49.8	2.8	49	99	45	**12**	1	0	80	89.3	5.2	100	100	100	100	100	100
38	50.5	2.8	52	99	50	**15**	2	0	81	90.3	5.3	100	100	100	100	100	100
39	51.2	2.8	55	99	54	**18**	2	0	82	91.2	5.4	100	100	100	100	100	100
40	52.0	2.7	58	99	60	**21**	3	0	83	92.1	5.5	100	100	100	100	100	100
41	52.7	2.7	61	100	64	**25**	4	0	84	92.9	5.6	100	100	100	100	100	100
42	53.4	2.7	63	100	69	**29**	6	0	85	93.7	5.6	100	100	100	100	100	100
									86	94.4	5.7	100	100	100	100	100	100

Note. The T score is a standardized score with a mean of 50 and an *SD* of 10 that is transformed from item response theory scaled scores. If a child received a summed score of 20 (see the bold number line in the table), the corresponding T score is 37.0, and the standard error of the T score is 2.8, which means the 95% confidence interval for the child's score would be [37 – 1.96 × 2.8, 37 + 1.96 × 2.8]. With respect to the population of typically developing children between 24 and 72 months of age, this child's score is at the 10th percentile. If the child is 24 months old, the percentile rank is 48. If this child is 36 months old, the score falls at the first percentile, and the child can be flagged as a late talker with only 2% of likely error. The bold and underlined scores in the 36th and 48th month columns indicate that the children with those scores are highly likely to be late talkers with only 2% false positive rate. If a score is bold, the child can be flagged as atypical with 80% correct probability, but also with a larger probability (40%–50%) of false diagnosis. mos. = months.