

Sections IIc.

Final Report (continued)

The Report of the results of test development for the *Diagnostic Evaluation of Language Variation (DELV)*:
c. Concurrent Validity Based on Comparisons to
Language Samples (UMass)

Section IIc

Evidence Based on Comparisons to Language Samples

Evidence of the DELV's validity was further evaluated by comparing the performance of children on the DELV-Standardization, African American edition with measures of actual language performance in Language Samples (LS) collected within one month of the children's test date. Language samples are often considered the gold standard for accurate observation of children's language ability, and they have been recommended as alternatives to using biased standardized language tests (Lund & Duchan, 1993; Stockman, 1996b). Other authors have suggested that language samples are essential complements to information gained from standardized testing (Wyatt, 2002). Therefore, various DELV scores and subscores were compared to aspects of the language samples to establish the most rigorous concurrent validity for the DELV. It is hypothesized that correlations between DELV Scores and the measures of AAE use and lexical, grammatical, and pragmatic complexity from the language samples will be moderate to high, and diagnostic accuracy for both Language Variation Status (LVS) and language impairment as measured against the LS measures will be over 90%. Further, it is hypothesized that the level of agreement between DELV scores and LS scores will be higher than the level of agreement of LS scores and clinicians' apriori diagnosis of language status.

Characteristics of the Language Samples

Ninety-one audiotapes of language samples of children who had taken the DELV were made within one month of the administration of the DELV by TPC examiners during the AA Norm Referenced standardization January 2003 to March 2004. Of those, 78 were sufficiently intelligible to be reliably transcribed. The demographics of the final sample for whom DELV tests and LS were analyzed are as follows:

Table 1. Subject characteristics.

Demographic Characteristics	Ageband 5:0				Overall Sample
	to 5:5	5:5 to 5:11	6:0 to 6:5	6:6 to 6:11	
n	18	20	20	20	78
Age					
Mean	5;3	5;8	6;3	6;8	6;0
Language status					
Typically developing	14	14	16	14	58
Language Impaired (receiving language services)	4	6	4	6	20
Gender					
Female	11	11	7	6	35
Male	7	9	13	14	43
Female Clinical					5
Male Clinical					15
Race/Ethnicity					
African American					100%
Parent Education					
	3.3	3.8	3.2	3.4	3.4
	(2 = hs diploma; 3 = up to two years of college; 4 = college)				
Region					
NorthCentral					19
Northeast					5
South					30
West					1
N with AA examiner					
	10	8	6	8	32

The Samples

The samples were made following a protocol designed the Principal Investigators which incorporated their own extensive experience and best practices for collecting language samples (protocol in Appendix). The target was 100 child utterances, including some narrative, some exposition, some “problem-solving,” and the rest general conversation.

The mean size of the samples was 175 child utterances, with a range of 49 to 284. There were 12 samples with fewer than 100 utterances. For the whole set of samples, there was a mean of 25% one-word utterances (range 4 to 51%). When one-word utterances were excluded, 17 samples had fewer than 100 utterances. Most analyses are done with the full set of samples but they were also done with the short samples excluded to verify that the effect of the short samples was not statistically significant. Unless noted, the analyses reported below include all samples.

The transcripts:

Each language sample was transcribed by two listeners (10% of them were transcribed independently). A third listener resolved differences between the transcriptions and finalized the typing of the sample in Excel. Portions of utterances with pronunciation that was not captured by standard orthography were written in on the original in IPA, but the IPA was not preserved in the final excel files as phonology was not a focus of the analyses.

The excel files were imported into SALT Student Version 7.0 (Miller et al., 1984-2002) and from there into CP9.5 corpora (Computerized Profiling, Long and Chaney, 2003). So for each child there is a rough draft excel file with handwritten IPA in places,

a finished excel file, a SALT file, a CP corpus, and the associated reports described below.

Coding:

Language Sample (LS) measures

Language Variation Status

The transcribers coded the child utterances for AAE features according to the Wolfram & Fasold list reproduced in Craig & Washington (1994, attached). For variable features like “+ cop” (presence of a present tense *is* or *are* copula), both absences and uses were coded.

Clinical Status

a. The following variables were obtained from by the SALT program:

- 1) # of utterances
- 2) % of one-word utterances
- 3) % of the child’s utterances that were in response to questions; % that were imitations; % that were spontaneous sentences (likely initiations)
- 4) # of different words in the transcript
- 5) # of different words in the first 50 utterances (#words50)
- 6) Mean Length of Utterance in words (mluw)
- 7) Mean Length of Utterance in morphemes (mlum)
- 8) Brown’s stage

b. From CP9.5 the LARSP Profile (Crystal, 1982) provided these additional variables:

- 9) MLU in words and morphemes (as a confirmation)
- 10) Mean Sentence Length, MLU with 1-word utterances excluded (MSL)
- 11) LARSP Phrasal complexity measure (LRSPPHC)
- 12) LARSP Clausal complexity measure (LRSPCLC)
- 13) LARSP % of sentences which are complex (%comp)
- 14) Sentence complexity (per Blake, Quartaro, & Onerati, 1993) (sentcomp)
- 15) IPSyn Total Scores (IPSTot)

c. Hand-coding

Finally, hand-scoring the transcripts yielded a number of further variables. The IPSyn (Scarborough, 1990) Sentence Structure subtotal (IPSynSS) was calculated by hand (and then the hand-scores were used to adjust the IPSyn Total Scores). In addition, 3-point scales were established for several sub-parts of the transcripts, and they were combined in an overall pragmatics composite, 17 LS-prag. For example, how did the child respond when requested to problem-solve, to do something when the materials were not available? (18 HOW) Did the child maintain referential clarity by any means while telling the balloon story, which was embedded in the protocol? (19 CLEAR) Did the child make reference to the mental states (desires and thoughts) of characters in the stories? (20 MNTLST) How much prompting was necessary for the child to tell a story? (21 NARR). (Other passages of the samples were coded for language of causation, temporal devices, and emotion but they were not significantly correlated with the DELV scores nor with age, and so they will not be discussed here.)

Most of the LS measures are considered productive syntax; # of words in 50 utterances a semantic measure; and % of spontaneous sentences and the variables HOW, CLEAR, NARR etc. were included as pragmatics measures.

DELV Scores/ Measures

From the children's DELV the analyses used the following variables: 1) raw scores for each subtest and each domain (Syntot, PragTot, SemanTot, and PhonTot), and 2) a raw total score (DLVTOT), summing raw scores across the three domains included in the DELV standard score, i.e. Syntax, Pragmatics, and Semantics. In addition, 3) a "production DELV"* (DLVProd, composed of raw scores from articles, all pragmatics, verb contrasts, and preposition contrasts rawscores) and 4) a "comprehension DELV" (DLVComp, composed of wh-syntax, passive, quantifiers, and fast-mapping rawscores) were calculated.

For Syntax, Pragmatics, and Semantics scaled domain scores were available from the DELV-NR norming, as well as 5) a standard composite score which encompassed all three (Seymour, Roeper, & de Villiers, 2005). Scaled domain scores have a mean of 10 and a standard deviation of 3; and the total standard score a mean of 100 and standard deviation of 15. From the standard scores a dichotomous variable was created: 6) DLVPass. Following Tomblin et al. (1997) "passing" was set at 1.5 standard deviations below the mean, i.e. 77.5. Above 77 was "passing," 77 and below was "failing." (The Phonology domain was not tested against the Language Samples as only one or two of the LS subjects were being treated for phonological impairment so no statistical testing

was possible. For concurrent validity of the Phonology Domain, see the PLS-Articulation Screener Study in the subcontractor's report.)

The DELV also yields 7) a Screener Diagnostic Risk Status raw score and 8) Diagnostic Risk label (*Highest, MediumHigh, Lowest, etc.*), and 9) Language Variation Status number of AAE responses, number of MAE responses, and 10) label (*No difference from MAE, Some difference from MAE, and Strong difference from MAE.*)

Methodological notes:

1. Source of the scaled scores: The scaled scores were calculated from the General Population standardization done by HAI in 2004. Rough standard scores were available from the AA-norm referenced data (April 2004), but more complete psychometric analyses were done following the second standardization (January 2005). Since the adequacy of the norms as they will be published is the more important consideration, the analyses for this report used the NR norms. A characteristic of the DELV-NR scoring is a Parent-Education (PED) Level adjustment (See DELV-NR Manual). The adjustment was not used in the comparisons to the LS measures because no parallel PED level adjustment is available for the other term of the correlations: that is, LS measures. Instead, PED level was entered into the analyses of variance as a covariate, so it could apply equally to both DELV scores and LS measures.

2. The Production-DELV score: The Production measure was also calculated using the Non-Word (NW) Repetition and several productive morphosyntactic items from the Screening Test. That score is correlated at .96 ($p < .01$) with the Production

measure without those items. Since the phonology measure was not being examined in this study, the measure without NW Repetition was used.

3. Ethnicity of the examiner: Ethnicity of the examiner was tested as a factor in the AA children's performance on the *DELV* and the Language Samples. Although less than 5% of the Speech Language Pathologists in the U.S. are African American, 32, or 41% of the Language Samples were collected by African American examiners. Ethnicity of the Examiner (AA or White) was entered into a MANOVA along with age and clinical status as independent variables and *DELV* raw scores, mean length of utterance, IPSyn sentence score, LRSP Clausal Complexity, and for Language Variation Status, the raw number of AAE responses and the percent of utterances with AA features as the dependent variables. No interactions were significant and only age was a significant main effect (for the *DELV* measures, not the Language Sample measures consistent with the results reported in Table 7 of this report). For ethnicity of the examiner, $F(1,76)$ ranged from .722 for IPSyn to 1.9 for MLU, all values with probability levels well above .05 ($p = .178$ to $.398$). Even the number of AAE tokens in the Language Samples was not significantly higher for the AA examiners. Ethnicity of the examiner did not, therefore, appear to affect children's performance in these tasks and was not considered further in the analyses.

1. Concurrent Validity of the DELV Language Variation Status and Language Samples

Correlations

The percent of AAE, or “dialect density,” observed in the Language Samples (LS-AAE) was calculated by dividing the number of dialect pattern tokens by the total number of utterances for each child. (This is “Token Method 3” according to Oetting & MacDonald [2002] which demonstrates empirically the equivalence of various methods for calculating dialect density found in the literature.) The correlation between LS-AAE and the number of AAE responses on the DELV Screening Test Items 1-15 was .560 ($p = .01$) for the total group; .57 ($p = .01$) for just the TD children. When the correlation was calculated within Language Variation label (*No difference from MAE*, *Some difference*, and *Strong difference*), the correlation was .674 ($p = .01$) for the 17 children with *No difference*, but much lower, $r = .427$, n.s. for the 13 children with *some difference*, and .132, n.s for the 48 individuals with *Strong difference*. This pattern indicates that above a threshold of AAE responses, the rank ordering by LS-AAE may not be significantly differentiated.

Category agreement

It was perhaps more important to confirm the concordance of the DELV category assignments: How well did the LS-AAE measure reflect the different LVS labels assigned by the DELV Screener? Since non-MAE responses are ambiguous between MAE and AAE disordered and AAE dialect responses for AA children, the analysis was done separately by language status for those who passed the DELV (DLV-TD) and those who failed it (DLV-LI). The steps of the calculations were as follows:

Typically developing children (DELV-Passers, n = 57)

Table 2 shows that 12 of 16 “No difference” children and only 1 “strong difference” child had LS-AAE at .05 or below, so “greater than .05” was used as the criterion to match *Some* or *Strong difference from MAE*.

Table 2. LS-AAE Ratio by DELV LVS Labels

LS-AAE ratio	DELV LVS Label		
	No Diff	Some Diff	Strong Diff
0-.035	7		1
.036-.059	5		
.060-.1	1	1	4
.101-.150	1	5	8
.151-.20	1	1	10
.201-.299	1		10
.3-.5			1
Total	16	7	34

The five discrepancies noted above were examined more closely to see if a principled reason could be found for LS-AAE to disagree with the DELV label. There may, indeed, be no discrepancy for the four children with higher LS-AAE measures than the DELV would predict. It is perfectly consistent with the literature for children to use more AAE features in a conversational situation than in one they might perceive as a test. In fact, two of the four children with a higher LS-AAE than their DELV had African American examiners, which could clearly encourage more AAE production even if the examiner did not use AAE. (The third child was just very chatty.) When the transcript of the 4th child in that category was examined, almost half of his AAE tokens came from “sposeta,” which seems less specific to AAE than the companion expression “fitna.” If one were to discount those tokens, since they are not corroborated as AAE by the rest of

the sample, then he no longer shows a discrepancy between his LS-AAE and his DELV. The one “Strong difference” child per the DELV who used very few AAE tokens in the language sample, nonetheless, used *invariant be* (“he be silly”), which is very specific for AAE with just a token or two. So in 4 of the 5 children, the discrepancy may be unremarkable.

Language impaired children (DELV failers, n = 21)

Among the DELV failers, as expected, almost all children had LS-AAE greater than .05. The two exceptions, DELV-failers who had low LS-AAE, had very short samples, with many one-word utterances (99 and 70 utterances, each with about 30 one-word responses). Their language samples then may not be sufficiently representative, and one might have more confidence in the DELV score than the LS-AAE.

Counting the seven “exceptions” (5 TD children and 2 LI), the diagnostic accuracy was 71 of 78, or 91%; and 97% if the five equivocal cases are not counted as disagreements.

2 - Concurrent Validity of the DELV Language Proficiency Measures and Language Samples

Correlations

Full DELV Correlations

The correlations between DELV raw score and the various LS measures are as follows:

Table 3. Correlations between DELV Raw Score and LS measures (N = 78)

DELV RawScore Total	<i>r</i> =
x	
No Word 50	.514***
Ipsyn Sentences	.306**
LRSPMLUM	.387***
LRSP_MSL	.404***
LRSP_% complex sentences	.339**
Pragmatics Composite (n=40)	.649***

- *p < .05; ** p < .01, ***p < .001

Thus, there are significant but moderate correlations between the DELV Total and most of the productive lexical and semantic measures. The selected pragmatics measures are as follows:

**Table 4. Correlations between DELV Raw Score and LS pragmatics measures
(N = 40)**

DELV RawScore Total	<i>r</i> =
by	
Problem-solving (HOW)	.487**
Narrative (NARR)	.546***
Referential Clarity (CLEAR)	.357*
Temporal Adverbs (TEMP)	.236
Mental State References	.595***
Pragmatics Composite	.620***

* $p = .05$; ** $p = .01$; *** $p \sim .001$

DELV Subscore Correlations

To better tap the productive aspects of the DELV responses, correlations were also done with the subscores of the DELV. The comprehension versus production emphasis of the different parts of the DELV is shown in Table 5. These correlations of the Domain Scores with the Production and Comprehension measure clearly reflect the different proportions of expressive versus receptive responses of the different DELV domains.

Table 5. Relations between Domain Scores and Comprehension and Production

Subscores

	Syntax Tot	Prag Tot	Seman Tot	Phon Tot	Sem Prod	Scrner	ALLDLV
DLVProd	.847***	.979***	.711***	.552***	.710***	-.710***	.928***
DVComp	.923***	.769***	.872***	.455***	.732***	-.720***	.954***

* p = .05; ** p = .01; *** p ~ .001

One can see in Table 5 that DELV Syntax is more Comprehension than Production; DELV Pragmatics is more Production than Comprehension; Phonology is more Production than Comprehension. Semantics is also more Comprehension than production, but only marginally so when we take out only the Verb and Preposition Contrast subtests, the production tasks (SemProd). The Screener (Scrner) is about equally Production and Comprehension, whereas the correlation of the Screener to the Full DELV (not shown) is somewhat higher, $r = -.765^{***}$ perhaps because there are more items and more variability counting the whole test together.

When compared to the DELV subscores, the correlations to LS measures were about the same magnitude or a little higher for the Productive measure, and a little lower for the Comprehension measure than to the full DELV. (The pattern is shown with a few representative measures in Table 6).

Table 6. Correlations between DELV subscores and LS measures (N = 78)

	SyntaxTot	PragTot	SemanTot	PhonTot	DLVProd	DLVComp
MSL	.340**	.465**	.313**	.310**	.458***	.317**
Wds 50	.481**	.498**	.418**	.310**	.505***	.466***
% Spon. Sents	.195	.144	.149	.227*	.156	.210
	(N = 40 for narrative measures)					
Narr	.408*	.605**	.527**	.398*	.599***	.443**
Mental State	.463*	.664**	.540**	.424*	.644***	.510***
Pragmatics Composite	.542**	.630**	.546**	.493**	.639***	.559***

* p = .05; ** p = .01; *** p ~ .001

The Semantic Production Only subparts (Verb contrasts and Preposition contrasts) do in fact show some more correlation with Words 50 than the full Semantics, .470** v. .418** but less change wrt to MSL, .338** vs. .313**

Correlations with age and Parent Education (PED) level

By contrast to the DELV, the LS measures showed no correlation with age. In this age range, the LS measures are not sensitive to age differences. (Indeed, if the LS measures had this useful property, their correlation with the DELV might have been even higher.)

Table 7. Correlations between DELV and LS Measures and Age (in Months)

Age x	r =
No Words 50	.129, p > .262
Ipsyn Sentence	.116, p > .311
LRSP_MSL	.074, p > .521
DELV Raw Score	.267, p = .018*

Note: PED level is a factor in the Norm-Referenced scoring, but with this group of 78, no PED level correlations were significant. PED Level was, nonetheless, entered into the ANOVAs below as a covariate.

Category Agreement (DELV Passers versus DELV Failers)

Concurrent validity of the DELV and the LS measures is also established by the significant differences between DELV passers and DELV failers on the LS measures. (We report selected LS measures which are representative of the related measures.)

Table 8. Analysis of Variance LS Measures by DELV status

	Passer (n=57)	Failer (n=21)	F	p	η^2
# of words50	106.9 (3.1)	81.6 (5.1)	17.36	.000***	.19
IPSyn					
Sentences	29.4 (.756)	25.96 (1.248)	4.95	.029	.064
MLU-m	4.86 (.157)	3.81 (.259)	9.52	.003***	.115
LRSP-MSL	6.07 (1.69)	4.96 (2.8)	9.94	.002	.12

Straightforward category agreement is hampered by the lack of clear categories for LS measures. Therefore, we used the scores of the set of children who were non-clinical according to both the clinicians' *a priori* judgments and the DELV ("double-passers") to establish the expected range of typical performance for this group. For each measure we established by ageband the mean and standard deviation to use in making z-scores for each measure.

As indicated by the lack of correlation, age group is not a significant variable for any of the LS measures. One can see from the means and the F-values in Table 9 that there is no clear age progression for the scores, either by 6-month age bands or by 12-month groupings, 5-year-olds versus 6-year-olds. [still to do]

Table 9. LS measures Mean (and Standard Error) by age group

(Means for double passers, N = 50)

	Age group				F	p
	5;0-5;5 N = 11	5;6-5;11 N = 13	6;0-6;5 N = 13	6;6-6;11 N = 13		
# of words50	104.3 (16.2)	112.6 (27)	112 (22.3)	108.1 (30)	.380	.768
IPSyn Sentences	28.5 (4.7)	30.1 (5.9)	29.2 (5.5)	28.9 (5.8)	.193	.901
MLU-m	4.61 (.90)	5.4 (1.6)	5.06 (1.4)	4.9 (1.6)	.154	.927
LRSP-MSL	5.67 (1.02)	6.4 (1.7)	6.2 (1.3)	5.9 (1.5)	.130	.924

By contrast, there are significant age effects for the DELV raw score variables, as shown in Table 10.

Table 10. DELV measure by age group (Mean for double passers, N = 50)

	Age group 3				<u>F</u>	<i>p</i>
	5;0-5;5 N = 11	5;6-5;11 N = 13	6;0-6;5 N = 13	6;6-6;11 N = 13		
DELV Raw Score	94.3 (12.6)	94.3 (13.7)	101.9 (10.1)	102.6 (8.0)	5.2	.026*

The patterns of significance in Analysis of Variance Tables 9 and 10 support the correlations above.

3. Comparison to *a priori* clinician judgments

The *a priori* clinicians' labels and the DELV outcomes differ by 1. That is, the clinicians identified 20 children as language impaired and those children were all receiving services at the time they took the DELV. According to the DELV Norm-Referenced standard scores, 21 children failed the DELV (scored > 1.5 standard deviation below the mean). In fact, though, the clinicians and the DELV identified different children: 7 *a priori* "clinical" children passed the DELV and 8 *a priori* "non-clinical" children failed the DELV. Thus, the two criteria disagreed 19% of the time.

Using z-scores for the LS measures--#words50, IPSyn, MSL, Narrative, etc.--a profile of LS measures was made for each discrepant child, shown in Figures 1-15. As one can see in Figures 1 to 15, the LS measures clearly bear out the DELV over the clinicians 10 of 15 times (67%), are equivocal 3 of 15 (20%), and disagree with the DELV in two cases (13%). Ultimately, one would say that the DELV and the LS measures were discrepant 2/78, or 3% of the time and inconclusive, 3/78 or 4% of the time.

Clinical cases who passed the DELV.

(In Figures 1-15, the variables on the x-axis are from left to right: the number of different words in 50 utterances, IPSyn Sentences, IPSyn Total, MLU in morphemes, Mean Sentence Length (not counting one word utterances), Blake & Quartaro syntactic complexity measure, and the hand-scored Pragmatics Composite. The right most bar represents the average of the 7 measures.)

- For 6 of the 7 of these children, the LS measures clearly support the DELV scores. (See Figures 1-6) Two children's LS measures were generally above of the average score for TD children of their age. Three others were within one-half of a standard point below the mean on average and had no z-score lower than -1.5 . The 6th child similarly had only one score lower than -2 and an average z-score of -1 .
- Only one clinical child (Figure 7) who passed the DELV (SS 86) had poor LS measures across the board (average LS z-score -2.5). Note that he has an articulation problem (PH-ss = 5), so perhaps the transcribers were not able to understand some of the complexity in his language. Note also that pragmatics and semantics were average (10 and 8), so he may have profited by the more explicit goals of those tests, as opposed to just general conversation.

Non-clinical children who failed the DELV

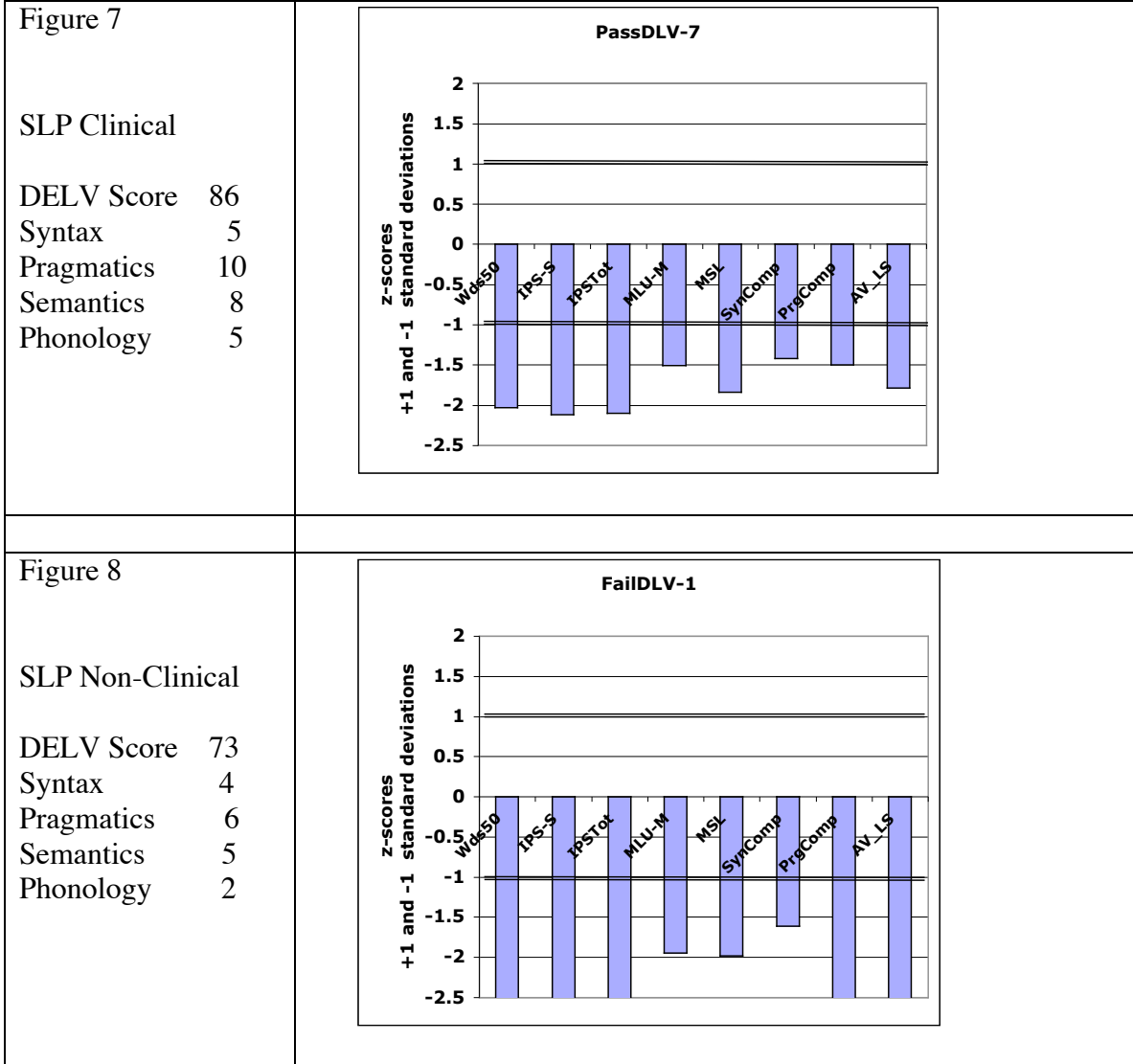
- Four of the 8 children in this group have language samples that clearly support a low DELV score. (See Figures 8-11) All of their LS measures are well below the average of their age group with average z-scores of -2.5 and -1.5 .

Another child (Figure 12) with a borderline DELV (SS 74) has mixed LS measures. His strongest LS syntax measures are at the phrasal level (IPSyn Tot is average but not IPSyn Sentences). Like his DELV scores, the LS measures are just barely below the average for the age group.

- Two children (Figures 13 and 14), however, --both early 6's-- have strong LS measures (except # of words in 50 utterances). They are low in both Comprehension and Production aspects of the DELV. One appears to have an articulation problem, but the other just appears to be quite average.
- The final child (Figure 15) has a DELV of 73 and is well above average in every measure except # of words in 50 utterances, and so we must ask why her DELV is weak and her language sample so strong. She represents the 2nd strong discrepancy between the DELV and the LS measures.

<p>Figure 1</p> <p>SLP Clinical</p> <p>DELV Score 82</p> <p>Syntax 7</p> <p>Pragmatics 8</p> <p>Semantics 6</p> <p>Phonology 10</p>	<table border="1"> <caption>PassDLV-1 z-scores</caption> <thead> <tr> <th>Category</th> <th>z-score</th> </tr> </thead> <tbody> <tr><td>Wds50</td><td>0.4</td></tr> <tr><td>IPS-S</td><td>0.5</td></tr> <tr><td>IPSTot</td><td>0.6</td></tr> <tr><td>MLU-M</td><td>0.8</td></tr> <tr><td>MSL</td><td>0.4</td></tr> <tr><td>SynComp</td><td>0.8</td></tr> <tr><td>PrgComp</td><td>0.7</td></tr> <tr><td>AV_LS</td><td>0.6</td></tr> </tbody> </table>	Category	z-score	Wds50	0.4	IPS-S	0.5	IPSTot	0.6	MLU-M	0.8	MSL	0.4	SynComp	0.8	PrgComp	0.7	AV_LS	0.6
Category	z-score																		
Wds50	0.4																		
IPS-S	0.5																		
IPSTot	0.6																		
MLU-M	0.8																		
MSL	0.4																		
SynComp	0.8																		
PrgComp	0.7																		
AV_LS	0.6																		
<p>Figure 2</p> <p>SLP Clinical</p> <p>DELV Score 89</p> <p>Syntax 8</p> <p>Pragmatics 10</p> <p>Semantics 7</p> <p>Phonology 7</p>	<table border="1"> <caption>PassDLV-2 z-scores</caption> <thead> <tr> <th>Category</th> <th>z-score</th> </tr> </thead> <tbody> <tr><td>Wds50</td><td>-0.2</td></tr> <tr><td>IPS-S</td><td>0.8</td></tr> <tr><td>IPSTot</td><td>0.2</td></tr> <tr><td>MLU-M</td><td>-0.2</td></tr> <tr><td>MSL</td><td>0.3</td></tr> <tr><td>SynComp</td><td>0.4</td></tr> <tr><td>PrgComp</td><td>0.7</td></tr> <tr><td>AV_LS</td><td>0.3</td></tr> </tbody> </table>	Category	z-score	Wds50	-0.2	IPS-S	0.8	IPSTot	0.2	MLU-M	-0.2	MSL	0.3	SynComp	0.4	PrgComp	0.7	AV_LS	0.3
Category	z-score																		
Wds50	-0.2																		
IPS-S	0.8																		
IPSTot	0.2																		
MLU-M	-0.2																		
MSL	0.3																		
SynComp	0.4																		
PrgComp	0.7																		
AV_LS	0.3																		
<p>Figure 3</p> <p>SLP Clinical</p> <p>DELV Score 98</p> <p>Syntax 7</p> <p>Pragmatics 11</p> <p>Semantics 11</p> <p>Phonology 3</p>	<table border="1"> <caption>PassDLV-3 z-scores</caption> <thead> <tr> <th>Category</th> <th>z-score</th> </tr> </thead> <tbody> <tr><td>Wds50</td><td>-1.4</td></tr> <tr><td>IPS-S</td><td>0.1</td></tr> <tr><td>IPSTot</td><td>0.1</td></tr> <tr><td>MLU-M</td><td>-0.5</td></tr> <tr><td>MSL</td><td>-0.1</td></tr> <tr><td>SynComp</td><td>0.5</td></tr> <tr><td>PrgComp</td><td>0.1</td></tr> <tr><td>AV_LS</td><td>-0.2</td></tr> </tbody> </table>	Category	z-score	Wds50	-1.4	IPS-S	0.1	IPSTot	0.1	MLU-M	-0.5	MSL	-0.1	SynComp	0.5	PrgComp	0.1	AV_LS	-0.2
Category	z-score																		
Wds50	-1.4																		
IPS-S	0.1																		
IPSTot	0.1																		
MLU-M	-0.5																		
MSL	-0.1																		
SynComp	0.5																		
PrgComp	0.1																		
AV_LS	-0.2																		

<p>Figure 4</p> <p>SLP Clinical</p> <p>DELV Score 79</p> <p>Syntax 5</p> <p>Pragmatics 6</p> <p>Semantics 8</p> <p>Phonology 3</p>	<table border="1"> <caption>PassDLV-4 z-scores</caption> <thead> <tr> <th>Category</th> <th>z-score</th> </tr> </thead> <tbody> <tr> <td>Wds50</td> <td>-1.0</td> </tr> <tr> <td>Ips-S</td> <td>-0.2</td> </tr> <tr> <td>IPSTot</td> <td>-0.2</td> </tr> <tr> <td>MLU-M</td> <td>-0.8</td> </tr> <tr> <td>MSL</td> <td>-0.2</td> </tr> <tr> <td>SynComp</td> <td>-0.2</td> </tr> <tr> <td>PrgComp</td> <td>-1.5</td> </tr> <tr> <td>AV-LS</td> <td>-0.5</td> </tr> </tbody> </table>	Category	z-score	Wds50	-1.0	Ips-S	-0.2	IPSTot	-0.2	MLU-M	-0.8	MSL	-0.2	SynComp	-0.2	PrgComp	-1.5	AV-LS	-0.5
Category	z-score																		
Wds50	-1.0																		
Ips-S	-0.2																		
IPSTot	-0.2																		
MLU-M	-0.8																		
MSL	-0.2																		
SynComp	-0.2																		
PrgComp	-1.5																		
AV-LS	-0.5																		
<p>Figure 5</p> <p>SLP Clinical</p> <p>DELV Score 89</p> <p>Syntax 7</p> <p>Pragmatics 9</p> <p>Semantics 9</p> <p>Phonology 9</p>	<table border="1"> <caption>PassDLV-5 z-scores</caption> <thead> <tr> <th>Category</th> <th>z-score</th> </tr> </thead> <tbody> <tr> <td>Wds50</td> <td>-1.5</td> </tr> <tr> <td>Ips-S</td> <td>0.5</td> </tr> <tr> <td>IPSTot</td> <td>0.5</td> </tr> <tr> <td>MLU-M</td> <td>-0.8</td> </tr> <tr> <td>MSL</td> <td>-0.2</td> </tr> <tr> <td>SynComp</td> <td>-0.2</td> </tr> <tr> <td>PrgComp</td> <td>-0.2</td> </tr> <tr> <td>AV-LS</td> <td>-0.5</td> </tr> </tbody> </table>	Category	z-score	Wds50	-1.5	Ips-S	0.5	IPSTot	0.5	MLU-M	-0.8	MSL	-0.2	SynComp	-0.2	PrgComp	-0.2	AV-LS	-0.5
Category	z-score																		
Wds50	-1.5																		
Ips-S	0.5																		
IPSTot	0.5																		
MLU-M	-0.8																		
MSL	-0.2																		
SynComp	-0.2																		
PrgComp	-0.2																		
AV-LS	-0.5																		
<p>Figure 6</p> <p>SLP Clinical</p> <p>DELV Score 80</p> <p>Syntax 8</p> <p>Pragmatics 7</p> <p>Semantics 5</p> <p>Phonology 1</p>	<table border="1"> <caption>PassDLV-6 z-scores</caption> <thead> <tr> <th>Category</th> <th>z-score</th> </tr> </thead> <tbody> <tr> <td>Wds50</td> <td>-0.2</td> </tr> <tr> <td>Ips-S</td> <td>-1.0</td> </tr> <tr> <td>IPSTot</td> <td>-1.0</td> </tr> <tr> <td>MLU-M</td> <td>-0.8</td> </tr> <tr> <td>MSL</td> <td>-0.2</td> </tr> <tr> <td>SynComp</td> <td>-0.2</td> </tr> <tr> <td>PrgComp</td> <td>-1.5</td> </tr> <tr> <td>AV-LS</td> <td>-0.5</td> </tr> </tbody> </table>	Category	z-score	Wds50	-0.2	Ips-S	-1.0	IPSTot	-1.0	MLU-M	-0.8	MSL	-0.2	SynComp	-0.2	PrgComp	-1.5	AV-LS	-0.5
Category	z-score																		
Wds50	-0.2																		
Ips-S	-1.0																		
IPSTot	-1.0																		
MLU-M	-0.8																		
MSL	-0.2																		
SynComp	-0.2																		
PrgComp	-1.5																		
AV-LS	-0.5																		



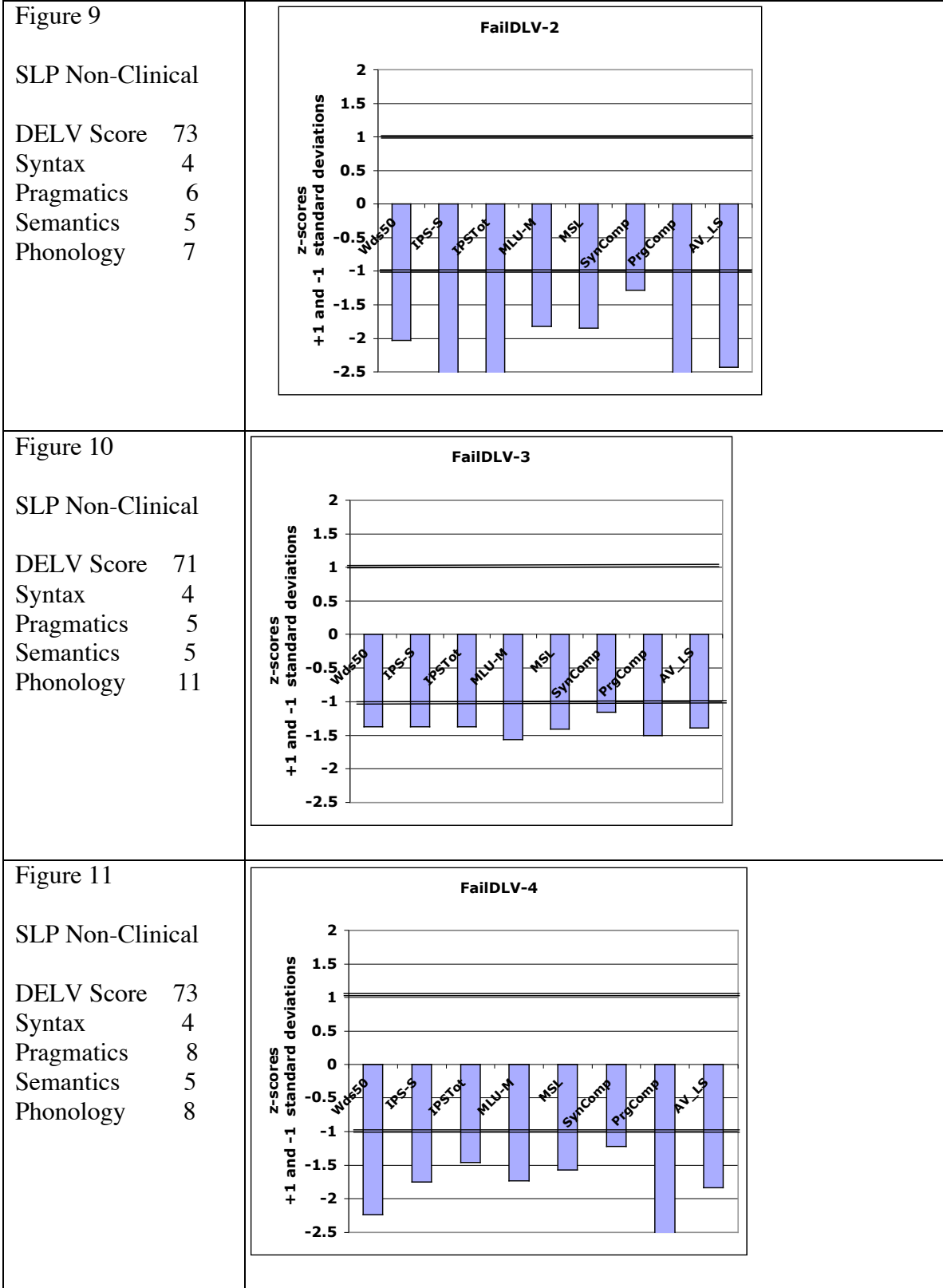


Figure 12

SLP Non-Clinical

DELV Score 74
 Syntax 4
 Pragmatics 6
 Semantics 6
 Phonology 11

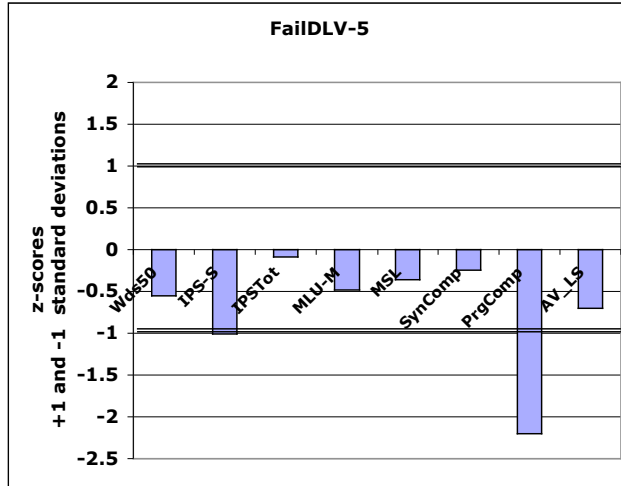


Figure 13

SLP Non-Clinical

DELV Score 68
 Syntax 4
 Pragmatics 4
 Semantics 4
 Phonology 1

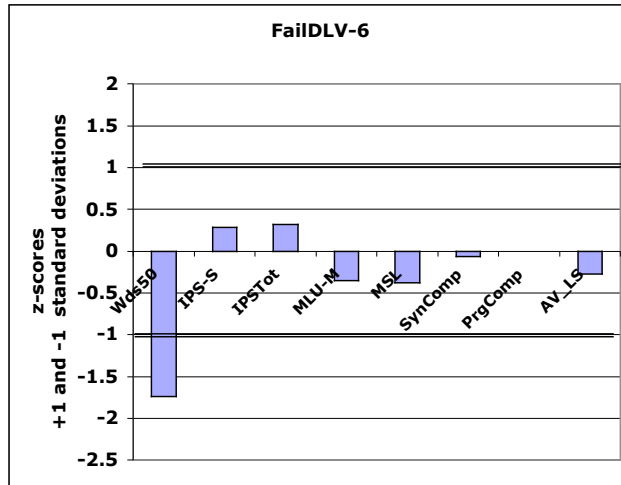
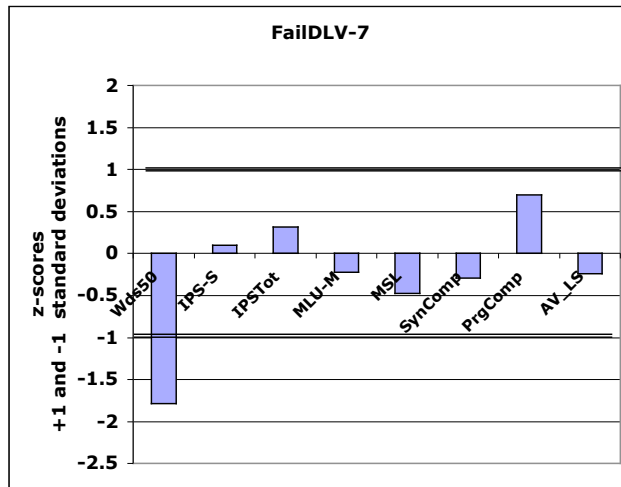
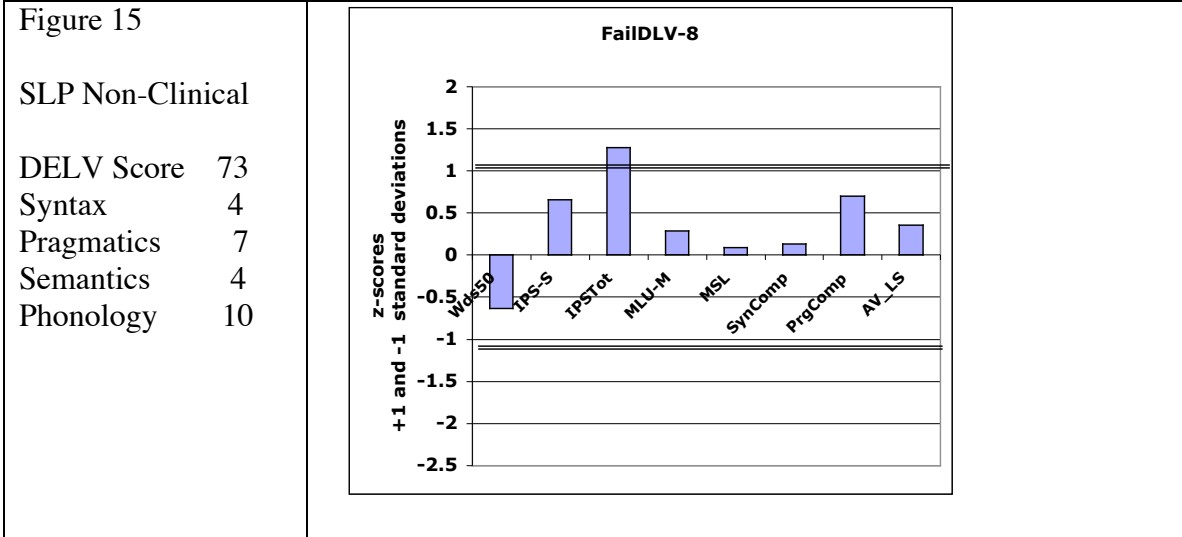


Figure 14

SLP Non-Clinical

DELV Score 73
 Syntax 4
 Pragmatics 6
 Semantics 5
 Phonology 12





A further corroboration of the *DELV* scores over the clinicians' judgments (based on other testing) comes from the comparison of analysis of variance of the LS-variables done with clinical status as the independent variable as opposed to *DELV* passing or failing as the independent variable. (For both analyses, PED level is entered as a co-variate and Age in Years as a second independent variable. In neither analysis is Age in Years or the interaction of "clinical" and Age a significant effect. For the analysis with the clinicians' scores PED level is a significant, but very small effect ($p = .045$, $\eta^2 = .054$). Table 11 shows the *F*-statistic, *p*-value, and eta-square for the two analyses. As one can see in the Table, "clinical" is a significant effect for #words50 and MLU in morphemes, but according to the eta-squared below the level of even a "small" effect. By contrast, when the *DELV* pass or fail is used, #words 50, IPSyn Sentences, MLU in morphemes and MSL are all significant with higher eta-squares.

Table 11. Comparison ANOVAs, "Clinical, Non-clinical" versus DELV Pass or Fail

clinical pass (n = 58)			DELV Pass (n = 57)		
F	p	eta-square	F	p	eta-square

NoWords50	5.132	0.026*	0.066	17.363	< .0001***	0.192
IPSyn Sent	0.131	0.719	0.002	4.958	0.029*	0.064
MLU-m	4.175	0.045*	0.054	9.517	0.003***	0.115
MLS	2.97	0.099	0.037	9.943	0.002***	0.12
SyntCompl	0.135	0.715	0.002	2.887	0.094	0.038

Conclusions

Several clear findings emerged from these analyses. First, the comprehensive DELV test was more sensitive to developmental growth over the age range from 5;0 to 6;11 than the various language sample scores. Significant growth in raw scores for each of the DELV domains (Syntax, Semantics, and Pragmatics) and in the total DELV score was observed across this age range. In addition, age in months was significantly correlated with total raw score. However, there were no significant relationships between age and any of the language sample measures. Nevertheless, there were significant correlations between the children's DELV scores and almost all of the measures of the language samples. These correlations were higher when the DELV profile was divided into subtests involving language production (e.g., Wh-question production, or short narrative production etc.) versus language comprehension.

The DELV Language Variation Status was also corroborated by the children's language in the samples. In addition to the correlations greater than .5 ($p = .01$), there was 91% diagnostic agreement (71/78) between the LVS labels and the number of AAE tokens observed, with only 2 of the 7 disagreements truly discrepant (76/78 or 97%).

Where there were discrepancies between the clinician's categorization of the children's clinical status and their performance on the DELV-NR, the language sample

analyses supported the DELV test results for 67% of the cases and was equivocal for another 20%. Only two cases were truly discrepant. We conclude that the DELV test provides a rich profile of children's language skills in the 5 and 6 year range that is in keeping with their language in spontaneous speech samples. Information from these two sources is usually complementary, though the information from the formal test is much more quickly and easily gathered.

REFERENCES:

- American Speech-Language-Hearing Association. (1983, September). Position paper: Social dialects and implications of the position on social dialects. *Asha*, 25(9), 23–27. Author.
- American Speech-Language-Hearing Association. (2003). Technical Report: American English Dialects. *ASHA Supplement 23*, in press.
- Berman, R. A. & Slobin, D. I. (1994). *Relating events in narrative: A crosslinguistic developmental study*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Berman, R. A. (1988). On the ability to relate events in narrative. *Discourse Processes*, 11, 469–497.
- Blake, J., Quartaro, G. & Onorati, S. (1993). Evaluating quantitative measures of grammatical complexity in spontaneous speech samples. *Journal of Child Language*, 20, 139-152.
- Bruner, J. (1986). *Actual minds, possible worlds*. Cambridge, MA: Harvard University Press.
- Campbell, T., Dollaghan, C., Needleman, H., & Janosky, J. (1997). Reducing bias in language assessment: Processing-dependent measures. *Journal of Speech, Language, and Hearing Research*, 40, 519–525.
- Champion, T. B. (1998). “Tell me somethin’ good”: A description of narrative structures among African American children. *Linguistics and Education*, 9(3), 251–286.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.
- Chomsky, N. (1977). On wh- movement. In P. W. Culicover, T. Wasow, and A. Akmajian (Eds.), *Formal syntax* (pp. 71–132). New York: Academic Press.
- Crystal, D. (1982). *Profiling linguistic disability*. London: Edward Arnold.

- de Villiers, J. G., & Roeper, T. (1995a). Barriers, binding, and the acquisition of the DP-NP distinction. *Language Acquisition*, 4, 73–104.
- de Villiers, J. G., & Roeper, T. (1995b). Relative clauses are barriers to wh- movement for young children. *Journal of Child Language*, 22(2), 389–404.
- de Villiers, J., Roeper, T., & Vainikka, A. (1990). The acquisition of long-distance rules. In L. Frazier & J. de Villiers (Eds.), *Language processing and language acquisition* (pp. 257–297). Boston: Kluwer Academic Publishers.
- de Villiers, P. (1988). Assessing English syntax in hearing-impaired children: Eliciting production in pragmatically-motivated situations. In R. Kretschmer & L. Kretschmer (Eds.), *Communication Assessment of Hearing-Impaired Students* (pp. 41–72). Academy of Rehabilitative Audiology, Monograph Supplement, Vol. 2.
- Dollaghan, C., & Campbell, T. (1998). Non-word repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*, 41, 1136–1146.
- Fisher, C. (1996). Structural limits on verb mapping: The role of analogy in children's interpretations of sentences. *Cognitive Psychology*, 31, 41–81.
- Gleitman, Lila. (1990). The structural sources of verb meanings. *Language Acquisition*. 1(1), 3–55.
- Green, L. (2002). *African American English: A linguistic introduction*. Cambridge, MA: Cambridge University Press.
- Halliday, M. A. K., & Hassan, R. (1976). *Cohesion in English*. London: Longman.
- Haynes, W. O., & Moran, M. J. (1989). A cross-sectional developmental study of final consonant production in southern black children from preschool through third grade. *Language, Speech, and Hearing Services in Schools*, 20(4), 400–406.
- Hymes D. (1972). On communicative competence. In: J. B. Pride & J. Holmes (Eds.) *Sociolinguistics* (pp. 269-293). New York: Penguin.
- Jackendoff, R. (2002). *Foundations of language*. Oxford: Oxford University Press.
- Johnson, V. (2001). *Fast mapping verb meaning from argument structure*. Unpublished Ph.D. thesis. (Department of Communication Disorders), University of Massachusetts, Amherst
- Labov, W. (1972). *Language in the inner city: Studies in the Black English vernacular*. Philadelphia: University of Pennsylvania Press.
- Leonard, L. B. (1998). *Children with specific language impairment*. Cambridge, MA: MIT Press.
- Liles, B. (1985). Cohesion in the narratives of normal and language disordered children. *Journal of Speech and Hearing Research*, 28, 123–133.
- Long, S., Fey, M., & Channell, R. (2003). Computerized Profiling. Software available at URL: www.computerizedprofiling.com. (last accessed: URL 3/8/04).

- Losen, D. J. & Orfield, G. (Eds.) (2002). *Racial inequity in special education, The Civil Rights Project at Harvard University*. Cambridge: Harvard Education Press.
- Lund, N. J. & Duchan, J. F. (1993). *Assessing children's language in naturalistic contexts*. Englewood Cliffs, NJ: Prentice Hall.
- Miller, J., with A. Nockerts. (1984, 2002). SALT (Systematic Analysis of Language Transcripts). Software program, Waisman Center, University of Wisconsin, WI.
- Moran, M. J. (1993). Final consonant deletion in African American children speaking Black English: A closer look. *Language, Speech and Hearing Services in Schools, 24*, 161–166.
- Naglieri, J. A. (2003). *Naglieri Non-Verbal Abilities Test (NNAT)*. San Antonio, TX: Harcourt Assessments.
- Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language, 17*, 357–374.
- Oetting, J. B. & McDonald, J. L. (2002) Methods for characterizing participants' nonmainstream dialect use in child language research. *Journal of Speech, Language, and Hearing Research, 45*, 505-518.
- Otsu, Y. (1981), *Universal Grammar and Syntactic Development in Children: Toward a Theory of Syntactic Development*. Unpublished doctoral dissertation, MIT.
- Philip, W. (1995). *Event quantification in the acquisition of universal quantification*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.
- Rice, M. L. & Wexler, K. (1996). Toward tense as a clinical marker of specific language impairment in English-speaking children. *Journal of Speech and Hearing Research, 39*, 239-257.
- Rice, M. L., Buhr, J. C., & Nemeth, M. (1990). Fast mapping word-learning abilities of language delayed preschoolers. *Journal of Speech and Hearing Disorders, 55*, 33–42.
- Roeper, T. (1987). The acquisition of implicit arguments and the distinction between theory, process, and mechanism. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 309–343). Hillsdale, NJ: L. Erlbaum Associates.
- Roeper, T., & de Villiers, J. G. (1994). Lexical links in the wh-chain. In B. Lust, G. Hermon, & J. Kornfilt (Eds.), *Syntactic theory and first language acquisition: Crosslinguistic perspectives. Vol. II: Binding, dependencies, and learnability* (pp. 357–390). Hillsdale, NJ: L. Erlbaum Associates.
- Ross, J.R. (1967). *Constraints on variables in syntax*. Unpublished doctoral dissertation, MIT.
- Scarborough, H. (1990). Index of productive syntax. *Applied Psycholinguistics, 11*, 1-22.
- Schafer, R. J., & de Villiers, J. (2000). Imagining articles: What a and the can tell us about the emergence of DP. In S. C. Howell, S. A. Fish, and T. Keith-Lucas

- (Eds.), *BUCLD 24: Proceedings of the 24th annual Boston University Conference on Language Development* (Vol. 2, pp. 609–620). Boston, MA: Cascadilla Press.
- Semel, E., Wiig, E. H. & Secord, W. A. (2004). *Clinical Evaluation of Language Fundamentals 4 (CELF-4)*. San Antonio, TX: Harcourt Assessments.
- Seymour, H. N., & Seymour, C. M. (1977). A therapeutic model for communicative disorders among Black English speaking children. *Journal of Speech and Hearing Disorders*, 42, 247–256.
- Seymour, H. N., & Seymour, C. M. (1981). Black English and standard American English contrasts in consonantal development for four and five-year old children, *Journal of Speech and Hearing Disorders*, 46, 274–280.
- Seymour, H. N., Roeper, T. & de Villiers, J. G. (2003). *DEL-ST (Diagnostic Evaluation of Language Variation) Screening Test*. San Antonio TX: The Psychological Corporation.
- Seymour, H. N., Roeper, T. & de Villiers, J. G. (2003). *DELV-CR (Diagnostic Evaluation of Language Variation) Criterion-Referenced Test*. San Antonio TX: The Psychological Corporation.
- Seymour, H. N., Roeper, T. & de Villiers, J. G. (2005). *DELV-NR (Diagnostic Evaluation of Language Variation) Norm-Referenced Test*. San Antonio TX: The Psychological Corporation.
- Snyder, L.S. & Silverstein, J. (1988). Pragmatics and child language disorders. In R. L. Schiefelbusch & L. L. Lloyd (Eds.), *Language perspectives: Acquisition, retardation, and intervention*. 2nd ed. (pp. 189-222). Austin TX: ProEd.
- Stockman, I. J. (1993). Variable word initial and medial consonant relationships in children's speech sound articulation. *Perceptual and Motor Skills*, 76, 675–689.
- Stockman, I.J. (1996a). Phonological development and disorders in African American children. In A. G. Kamhi, K. E. Pollock, & J. L. Harris (Eds.), *Communication development and disorders in African American children: Research, assessment, and intervention* (pp. 117–154). Baltimore: Paul H. Brookes.
- Stockman, I. J. (1996b). The promise and pitfalls of language sample analysis as an assessment tool for linguistic minority children. *Language, Speech, Hearing Services in Schools*, 27, 355-366.
- Stoel-Gammon, C. & Dunn, C. (1985). *Normal and disordered phonology in children*. Austin TX: Pro-Ed.
- Tomblin, B., Records, N. L., Buckwalter, P. R., Zhang, X., Smith, E. & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research*, 40, 1245-1260.
- U.S. Bureau of the Census (2000). *Current Population Survey, October 2000: School Enrollment Supplemental File [CD-ROM]*. Washington, DC: U.S. Bureau of the Census (Producer/ Distributor).

- University of Massachusetts Working Group on African American English. (April, 2002). *Development milestones for AAE speaking 4-, 5-, 6-year-olds*. Amherst, MA: Author.
- van der Lely, H. 2003. Do heterogeneous deficits require heterogeneous theories? SLI subgroups and the RDDR hypothesis. In Y. Levy & J. Schaeffer (eds.) *Language competence across populations: Toward a definition of specific language impairment*. Mahwah, NJ: Erlbaum.
- Vihman, M. M. (1998). Later phonological development. In J. E. Bernthal & N. W. Bankson (Eds.), *Articulation and phonological disorders* (4th ed., pp. 113-147). Boston: Allyn and Bacon.
- Washington, J. A. & Craig, H. K. (1998). Socioeconomic status and gender influences on children's dialectal variations. *Journal of Speech, Language, and Hearing Research, 41*, 618–626.
- Washington, J. A., & Craig, H. K. (1994). Dialectal forms during discourse of poor, urban, African American preschoolers. *Journal of Speech and Hearing Research, 37*, 816–823.
- Waxman, S. R., & Hatch, T. (1992). Beyond the basics: Preschool children label objects flexibly at multiple hierarchical levels. *Journal of Child Language, 19*(1), 153–166.
- Wolfram, W. (1991). *Dialects and American English*. Englewood Cliffs, NJ: Prentice Hall.
- Wyatt, T. A. (2002). Assessing the communicative abilities of clients from diverse cultural and language backgrounds. In D. E. Battle (Ed.), *Communication disorders in multicultural populations* (3rd ed., pp. 415–459). Boston: Butterworth-Heinemann.
- Zimmerman, I. L., Steiner, V. G. & Pond, R. (2002). *Preschool Language Scale-4 (PLS-4)*, San Antonio, TX: Harcourt Assessments.

ATTACHMENTS:

1. Authors' language sample protocol
2. List of AAE features from Washington & Craig, 1994.
3. Pragmatics coding