Language Samples as a Gold Standard for an Innovative Test
2011 ASHA Abstract (submitted)

Janice E. Jackson & Barbara Zurer Pearson
University of West Georgia & University of Massachusetts Amherst

This paper uses concurrent language samples as a gold standard to test the sensitivity and specificity of an innovative, dialect-neutral language test.

When children from other linguistic backgrounds who are learning general American English (GAE) as a second dialect take tests standardized on GAE first-dialect speakers, the norms will likely misrepresent standards of risk for impairment for them. Without a test normed specifically for them, culturally and linguistically different (CLD) children may score below their aptitude on tests because of "difference, not deficit." Therefore, caution has been advised in interpreting assessment results for CLD children, such as those who speak African-American English (AAE) (Craig & Washington, 2006).

The same caution is required when evaluating an innovative non-biased test for CLD/AAE-background children. For example, the *Diagnostic Evaluation of Language Variation, Norm-Referenced (DELV-NR*, Seymour, Roeper & deVilliers, 2005) achieved dialect-neutral assessment with the use of innovative item-types and was normed independently on an AAE population and a general American population. However, the very uniqueness of such a test, that intentionally differed from standard assessments, means that previous tests--that were deemed inappropriate, or potentially biased for diagnosing CLD *children*—will also be inappropriate references for evaluating the sensitivity and specificity of the new *test*. Rather, it must be shown that the new test coincides with an *independent* assessment of impairment.

Language samples are often considered the *gold standard* for observations of children's language ability and have been recommended as alternatives to using biased standardized language tests (Lund & Duchan, 1993; Stockman, 1996; Wyatt, 2002). Thus, language-sampling is a logical candidate for concurrent validity for an innovative test such as the *DELV-NR*. In this paper, language samples are used as a reference comparison (Dollaghan & Horner, 2010; Heilmann, Miller & Nockerts, 2010) for *DELV-NR* outcomes. Our research question asked how convergent were the diagnostic decisions based on the *DELV-NR* compared to similar decisions based on language samples? Further, when the innovative test and the child's clinical status based on traditional methods of assessment differed, which diagnosis was more likely to agree with the child's language sample?

Seventy-eight AAE-background children, ages 5;0 to 6;11, were recruited by certified SLPs and categorized as typically-developing (TD) or having language impairment (LI) based on the assessment of professionals in their schools and communities: that is, 20 children were receiving language services and were thus considered as having LI, while 58 were not receiving services and were labeled TD. Participants took the DELV-NR according to the published directions for the test, and a language sample (LS) incorporating narrative, exposition, and free conversation was collected within three weeks of testing. The *DELV-NR* were scored according to the published directions, and -1.5 standard deviations (SD) below the mean (or >77) was set as the criterion for TD, and 77 or below as the criterion for referral for services. LS were transcribed

and analyzed with Computerized Profiling (Long et al., 2003). The following measures were calculated: number of different words (NDW), mean length of utterance (MLU), IPSyn sentences (Scarborough, 1990), a syntactic complexity measure (Blake et al. 1993), a literate language measure (Westby, 1991), and a pragmatics composite score based on the narrative and exposition in the LS prototcol (Magaziner et al., 2008).

To derive a diagnostic category (TD or LI) from the language samples, a *z*-score was calculated for each measure based on the means and standard deviations for four 6-month agebands in the age range. A profile was created for each child, encompassing *z*-scores for the four semantic and syntactic measures and the two pragmatics measures. Children's clinical status was determined using -1 SD or above as "average" and < -1 SD as "below average." Profiles within the average range that included an individual score lower than -2 SD below the mean were considered "mixed," as it was not clear how to characterize overall performance for those children. Using this system, there were 55 children with LS profiles clearly above average, 14 children whose profiles were clearly below average, and 9 were mixed. The 69 clear-cut LS profiles were used as the gold standard for the *DELV-NR* and especially for the 15 of 78, or 19%, of cases where the clinical status from the *DELV-NR* differed from the clinical status based on whether the child was receiving language services or not.

Sensitivity and specificity for their diagnosis determined by the *DELV-NR* were calculated for the 69 children with clear-cut language samples. Overall agreement with the LS profiles was 84%; specificity was .9 and sensitivity .64. That is, there were 91% *true negatives*—"LS-TD" children—among the TD children identified by the *DELV-NR*, and 60% *true positives*, or "LS-LI" children, among those identified as having LI by the *DELV-NR*. LS profile agreement with the pre-existing clinical categories was somewhat lower, at 75%, and specificity and sensitivity were .84 and .43 respectively.

For the 15 cases where the *DELV-NR* and "standard practices" disagreed with each other, three had mixed LS profiles and level of agreement could not be determined. Twelve of the 15 cases of disagreement had clear-cut LS profiles: ten of the LS profiles (83%) agreed with the *DELV-NR* (and not standard practice), and two (17%) agreed with the pre-existing diagnosis (and not the *DELV-NR*).

Therefore, with regard to sensitivity and specificity based on language samples as the gold standard, this innovative dialect-neutral test appeared to be superior to standard practices and showed promise for identifying language impairment in populations frequently at risk for mis-diagnosis. Nonetheless, sensitivity values were lower than is customary. To the extent that language samples are easier than the new test, some children may not have shown their weaknesses in them, since they said only what they could say and may have avoided more difficult constructions. On the other hand, language samples may not have given other children the opportunity to demonstrate their knowledge of more challenging constructions, like long-distance syntactic movement or theory of mind as found in the *DELV-NR*, and thus the LS may have underestimated their abilities. (991)