FINAL REPORT- 3/1/98 to 8/29/05

# NIH Contract N01-DC8-2104: Development and Validation of a Language Test for Children Speaking Non-Standard English: A Study of Children Who Speak African American English

**Harry N. Seymour, P. I.** *University of Massachusetts, Amherst*
**Thomas Roeper, Co-collaborator,** *University of Massachusetts, Amherst*
**Jill G. de Villiers, Co-collaborator,** *Smith College*
**Peter A. de Villiers, Consultant,** *Smith College*
**Barbara Z. Pearson, Project Manager,** *University of Massachusetts, Amherst*

**Subcontractor: Harcourt Assessments, Inc.**
**Lois Ciolli, Research Team Project Director**

**Project Officer: Judith Cooper, Director, Division of Scientific Programs**
**Contract Officer: John P. DeCenzo**

This document is the final report of Contract NIH N01-DC8-2104, fulfilling Objective c1 "**Deliver to NIDCD a final report of results of the test development."** The report is comprised of three parts:

I. A summary of the accomplishments of the Contract
II. The Report of the results of test development for the *Diagnostic Evaluation of Language Variation (DELV)*:
    a. Development and Standardization (HAI)
    b. Reliability and Validity, test-retest, CELF-4, NNAT, PLS Articulation Screener (HAI)
    c. Concurrent Validity Based on Comparisons to Language Samples (UMass)
III. A list of the products, datasets, and references produced under the contract.

## I. Summary of Accomplishments under Contract NO1-DC-8-2104

The accomplishments under the contract are accurately summarized by the "Schedule for completion of work" described in the original proposal. The research team and its subcontractors met or exceeded all the terms of the contract in a timely manner.

The purpose of NIH Solicitation DC-97-01 was to develop a fair and equitable testing instrument for children who speak African American English (AAE). In addressing this purpose the research team developed assessment instruments that provide a practical solution to the problem posed in ASHA's position paper on social dialects, articulated 20 years ago and reissued just two years ago (ASHA 1983, 2003). The ASHA paper states that "…. No dialectal variety of English is a disorder or a pathological form of speech or language." Still, speech-language pathologists rely heavily on a single dialect standard, i.e. Mainstream American English (MAE) as the referent of acceptability when assessing the language of children. One very obvious penalty for African American (AA) children, many if not most of whom do not speak MAE, is their disproportionate representation in language services and special education programs throughout the country. In recent figures, little changed from the 1970s (Losen & Orfield, 2002), African American children were three times more likely than MAE speakers to be put into special education programs, where their chances of eventual graduation were reduced to 1 in 4.

The result of the contract is an assessment battery, the *Diagnostic Evaluation of Language Variation (DELV)*, which meets the RFP objectives by employing two strategies: Strategy 1 starts with dialect identification, elicited with highly contrastive structures, (those that differ most consistently between AAE and MAE); Strategy 2 follows with dialect-neutral diagnosis of disorder. To achieve *dialect neutrality*, the *DELV* avoids superficial contrasts between dialects of English, and focuses instead on structures that are non-contrastive. Also, the *DELV* draws upon deep principles of language considered universal across dialects and even across languages. In this way, these tests accomplish the difficult task of distinguishing dialect and normal development from impairment or delay.

The *DELV* instruments, published by the contract investigators and their subcontractor, Harcourt Assessment, Inc. (previously known as The Psychological Corporation), are comprised of three tests appropriate for both AAE and MAE speakers between ages 4 and 9 (Seymour, Roeper & de Villiers, 2003a; 2003b; 2005). The first is a screener, the *DELV-Screening Test (ST)*, with two parts: part one designed to identify Language Variation Status in terms of whether a child is an MAE speaker or not, and part two to screen for children who may be at risk for a disorder (Diagnostic Risk Status). The follow-up test, the *DELV-Criterion Referenced* (*CR*), is longer and provides criterion-referenced cut-off scores for a comprehensive assessment of syntax, semantics, pragmatics, and phonology. The third test is the *DELV-Norm Referenced (NR)*. The *DELV-NR Stimulus Book* is the same as the *DELV-CR*, but instead of cut-off scores, there are norms for using the test. These norms satisfy the requirement in many states that children be tested with a norm-referenced instrument.

An important feature of the *DELV* project was a unique partnership between the principal investigators and a major publishing company, Harcourt Associates, Inc. (HAI, formerly The Psychological Corporation, TPC). The UMass/Smith team designed the tests and the research plans and did the initial piloting, while HAI produced the materials and carried out the nationwide "Tryout" and standardization data collections. The contract was to cover 4- to 6-year-old African American children, with a small comparison group of typically-developing (TD) MAE speakers. Going beyond the stipulated numbers in the contract, HAI extended the Tryout research--at their own expense--to children from 4 to 12 and enlarged the comparison groups of AAE language impaired (LI) speakers and MAE speakers, including a subgroup of LI-MAE speakers. Thanks to this added effort, the *DELV Norm-Referenced*, as well as the *ST* and *CR*, are appropriate for children 4 to 9, which is double the original age range. This report is based on the African American norming carried out as Objective III of the contract. Where possible, comparisons will be made to HAI's data on a general U.S. population (MAE, AAE, and other), but studies reported here are based on data collected under the contract.

Note that since the items on the *DELV* were shown in Tryout to work as well for MAE speakers as AAE speakers, HAI embarked on a second standardization based on a U.S. General population, once again at their own expense. The U.S. general norming provides an external validation of the test in a broader community, at the same time increasing its acceptability to the population it was designed for. This extra step provides a compelling demonstration that that the *DELV* achieves what no other language test can boast, i.e. a rigorous assessment which eliminates the performance gap between Black and white children.

The above accomplishments were achieved in accordance with the three phase objectives of the RFP. The highlights of this three-phase process by which the contract was fulfilled are as follows, (as outlined in the Technical Proposal). This report also follows these three steps:

I. Methodology, choosing the language behaviors to test.
II. Collection of Milestone Data (using items drawn on the behaviors chosen in Phase I)
III. Development and validation of the Language Test (using the most effective items from Phase II, on a norming population of AA speakers).

## PHASE I. Methodology, choosing the language behaviors to test

The choice of language behaviors to test was guided by four strategies:

1. **Select dialect-specific targets for identification; Avoid dialect specific targets for assessment.** Following Seymour & Seymour (1977), language behaviors were analyzed according to whether they were *contrastive* between AAE and MAE or whether they were *non-contrastive*, i.e. realized the same in both dialects. For an example, consider a

feature from phonology. The cluster "st" at the end of a word as in "ghost" is *contrastive* because it is most often produced "ghos" by AAE speakers, whose dialect avoids closing syllables with consonant clusters, whereas MAE speakers, with no such constraint, would most likely say "ghost." The same "st" at the beginning of a word, as in "stove," is *non-contrastive*, since AAE and MAE speakers pronounce it alike.

2. **Probe a deeper level of language knowledge.** Previous language assessment tools focus primarily on broad and low-level descriptive accounts of language. The *DELV* taps into modern linguistic theory for useful insights on universal grammar, which is the foundation of all languages (Chomsky, 1965). What has been uncovered in this work is a schema for languages that is rich and deep, a shared set of properties defining how sentences are made, and principles that the grammars of all human languages obey, no matter how the languages or dialects vary on the surface. We test this abstract language knowledge by looking at, for example, hidden properties of passives, the way "variables" behave in wh-words and quantifiers, and the limits on movement rules (Jackdendoff, 2002). At this deeper level of analysis, disorder is often most obvious.
3. **Focus on pragmatic aspects of language essential for schooling and literacy.** The *DELV* concentrates on evaluating the linguistic bases of academic skills. Although some aspects of pragmatics vary across cultural groups, others are constant. For example, all children need to understand the conditions for speech acts. In narratives, all must use clear referents and events and understand the mental states of characters. In the classroom, all children must be able to ask for the right information.
4. **Avoid focusing on acquired lexical vocabulary.** Acquired vocabulary varies by cultural group, and tests can only evaluate a limited set of words. Rather than counting the number of words from a particular list that a child knows, as many current tests do, the *DELV* assesses instead how well a child learns words *in* context and *from* context, and then how he or she organizes them for efficient retrieval.

### Ia. Translation of experimental paradigms from theoretical linguistics into test items

During the Phase 1 Methodology, experimental paradigms from theoretical linguistics and from specific research on AAE were translated into workable test items for clinicians. Over 50 language behaviors were chosen to collect milestone data on; 320 items were selected from pilot studies in 14 broad areas (e.g. wh-questions, passive, verb vocabulary). Although many of the items appear similar to items on other language tests, the *DELV* items are drawn from areas of innovative research and have a sharper focus than similar items on other tests. In addition, item choices were based on three influential theories about the nature of specific language impairment (SLI), one suggesting that problems with morphosyntax are fundamental (Leonard, 1998; Rice & Wexler, 1996), a second that highlights children's difficulties in processing incoming speech, as tested by non-word repetition tasks (Campbell et al., 1997, Dollaghan & Campbell, 1998), and a third that focuses on difficulties LI children have with syntactic forms that involve movement rules (van der Lely, 2003). These theories are represented in different parts of the assessments.

Item types were organized according to the four traditional domains of language study: **syntax, semantics, pragmatics, and phonology,** with subcategories for **morphosyntax** and **non-word repetition**. All four domains aimed at satisfying 3 goals:

1) to capture abilities developing during this age range;
2) to show no performance differences between AAE and MAE speakers, and
3) to discriminate clearly between typically developing and impaired children.

In addition, each domain provides a different perspective on the child's basic language functioning.

**SYNTAX:** Items were created in 3 subdomains: wh-comprehension, passives, and articles, which probe 1) essential properties of questions, 2) implicit grammatical relations and 3) discourse linking, as discussed below.

**Wh-comprehension**:

Wh-comprehension tells us whether children know the rules and restrictions for moving wh-elements within sentences. Since the 1970s, wh-phenomena, especially the movement of wh-elements in and out of clauses, have been studied in many different languages (Ross, 1967; Chomsky, 1977). Child language researchers quickly understood the deep significance of wh-questions in developing grammars (Otsu, 1981; de Villiers, Roeper, & Vainikka, 1990), and a 1992 study by Seymour and colleagues extended the analyses to AAE, which works essentially like MAE in this respect. Wh-questions prove to be a refined way to see what the child's grammar contains or lacks.

What is wh-movement?

We know that in languages like English the wh-word moves to the front of the sentence to make a simple question: You ate something ==> You ate what? ==> *What did you eat (...) ?* The site from which it moves, and the gap it leaves, can be several clauses away and we can still recover the meaning: *What did you say you ate (...)? What did Jim say he saw you eat (...)?* and so on.

However, there are some places the wh-word cannot move from, for example, from inside a relative clause, as in: *John said he saw a man who ate a snake.* A question without movement, like *John said he saw a man who ate what?* is an interpretable question; however, **What did John say he saw a man who ate (...)?* is not. So, there is a clear syntactic limit on how wh-elements can move (de Villiers & Roeper, 1995a; de Villiers & Roeper, 1995b; Roeper & de Villiers, 1994).

To test children's knowledge of such limits, or "barriers," the DELV presents brief stories about pictured events and asks key test questions about some aspect of the events. In one kind of scenario, for example, we ask whether children can retrieve the place where a wh-word came from if it is two clauses away and embedded in a false clause. Ex. A mother buys a birthday cake, but to keep it a surprise she tells her child that she bought paper towels. The test question, *What did the mom say she bought*? requires the child to interpret both clauses and the relationship between them. She or he cannot just answer what the mother bought, but must answer what she *said* she bought.

The *DELV* also probes other subtle properties of wh-words, especially when asking two questions at once, which brings out the quantifier "hidden" in wh-words, e.g. *who* = "who-everyone." If we ask, *Who bought what*? the question calls for two answers (for "who" and for "what"). The answers must refer to all the members in 2 sets in an ordered relation, that is Person 1 bought Thing 1, Person 2 bought Thing 2, etc.

**Passive**

The passive is another construction found on the *DELV* where some information is totally unspoken but still known to any competent speaker. Passive comprehension items test children's understanding of movement and also implicit relationships, that is, hidden information which is implied by the grammar of the sentence, but not stated in words (Roeper, 1987).

First of all, as in wh-questions, the child must be able to follow the movement of grammatical elements to different parts of the sentence and still keep track of the original relationships, for example: *John kissed Mary -> Mary was kissed.*

There are also other pieces of information encoded in the passive. One can visualize the hidden extra information provided by comparing these sentences: a) *The player dropped; b) The player was dropped;* and *c) The player was being dropped.* Sentence (a) does not indicate how the player dropped. She could be doing it herself, someone could be forcing her down, or it could be one of those things that just happen. Sentence (b) tells us someone or something did it, but gives no indication of when. Sentence (c) tells us that someone or something else is doing something to the player and it was on-going at the time the sentence was being spoken. The child's ability to recognize the unstated information comes not through any additions to the content words of the sentence (like *player* or *drop*), but from interpreting the grammatical bits of the sentence (*ed, ing,* and the forms of *be*).

**Articles**

The last item-type in the Syntax domain, Articles, explores children's ability to link information across sentences. Articles feel very simple and automatic, but there are strict conditions on when a speaker may use the definite article *the* as opposed to the indefinite *a*. What one is referring to with *the* must be something already known to the listener. So the speaker has to calculate the listener's knowledge. In most cases that means keeping track of what was said in a previous sentence to know what is "old" or "given" information.

In testing the child's knowledge of what is new or old information, the *DELV* avoids picture stimuli. As pointed out by Schafer and de Villiers (2000), even a first mention of one of the items or people in the picture may elicit *the*, since the picture puts it into the speaker's and listener's shared context. To shift the burden for creating a context back to the discourse, we present brief stories about the items to be questioned and then ask a question whose answer contains an article. For example: A bird and a snake were sitting on a rock. They were friends. One of them flew away. *Which one?* Here the answer is *the bird*, not *a bird*.

**PRAGMATICS:** The Pragmatics subtests show a different view of children's language abilities. These subtests try to characterize children's growing communicative competence (Hymes, 1972; Snyder & Silverstein, 1988) rather than focusing on the structural forms (syntax) or content (semantics) of their language. Three subdomains test children's functional language skills: 1) question-asking; 2) communicative role taking, and 3) short narrative. (A fourth subdomain, identifying referents, was explored through the Tryout phase, but was not carried into the *DELV* for reasons of length.)

The key features of the elicitation materials and procedures in the Pragmatics tests are 1) they provide specific referential support and pragmatic motivation for the language forms and content to be produced by the child, so they greatly increase the likelihood that those forms and functions will be sampled in the assessment; 2) the pictured materials and the elicitation prompts constrain the range of appropriate utterances, so the children's productions are much more easily scored than a more open-ended spontaneous speech sample; however, 3) the procedures retain communicative naturalness rather than resorting to unnatural imitation procedures to elicit the forms; and finally, 4) all of the materials are picture-based so they require minimal technology and can be administered and scored by a single clinician interacting with the child.

**Question Asking**

This set of probes presents a picture with an ambiguously shaped blank white space in it. The "game" calls for the child to ask "the right question" to discover what is happening in the white space of the picture. The items elicit *what, who, where, why*, and *how* questions, and also a double wh-question form (*Who is eating what*?) that indicates whether the child understands the set properties of complex wh-questions (cf. wh-comprehension in Syntax). The challenge for the child is to produce a semantically and pragmatically appropriate wh-question, but the exact syntactic form produced can vary and still be acceptable for that item. Thus, *What she paintin'*? in AAE is as appropriate in pragmatic terms as the MAE *What is she painting*? but *What's that*? or *That's a car* would not be adequate responses.

**Communicative Role Taking**

Children's ability to take the perspective of another speaker and to understand what speech act they were producing is tested in a communicative role-taking task. For each trial the child is shown a sequence of two pictures. In the first, a character participates in an event. In a second picture, the same character is either gesturing and clearly saying something to another person, or is clearly being spoken to by the newly introduced person. Depending on the particular sequence, the child is asked by the tester what the speaking character in the second picture was "telling," "asking," or "saying to" the other person. E.g. If the tester says *What is the character ASKING his dad?* the child's answer can be either a direct question, *Can I go out?* or a report of a question, *if he can go out* (or *aksing can he go out*). But the answer cannot be *Look at the boy*. Thus, this subdomain tests the children's ability to understand the communicative role of the speaking character and also their sensitivity to the pragmatic constraints placed on their response by the tester's prompt.

**Narrative**

Narrative is the first genre of reading and writing that children do, so the acquisition of good narrative skills is crucial for early literacy development. The Short Narrative subdomain gives the child a format in which to create an extended discourse based on a picture sequence. Narrators are free to choose different perspectives on events, but well-formed stories of all types have both thematic coherence on the macro-level of the structure or organization of the events, and linguistic cohesion at the micro-level of referents and clauses (Berman & Slobin, 1994; Halliday & Hasan, 1976).

Research studies suggest that AAE-speaking children produce a wider range of different story structures when given an open ended story-telling task (Champion, 1998) so the elements of coherence will differ across cultures, but the requirements for cohesion are more similar. Cohesion is also easier to measure, so the *DELV* concentrates on children's mastery of linguistic cohesion**.** Stories are evaluated "on-line," without the burdensome necessity of being recorded and transcribed. The scoring takes into account contrastive specification of referents (telling the listener who they are referring to as each action and event is described), and how the child links together the events of the story in time, features considered to be revealing of developmental growth or language delay in children (Berman, 1988; de Villiers, 1988, 1991; Liles, 1985) Also, more mature stories make reference to a "Theory of Mind," the meaning of the events for the characters (the "inside view" or "landscape of consciousness" rather than just the "landscape of action" (Bruner, 1986)). The *DELV* picture sequence encourages the expression of the mental states of the characters, and the children are asked specifically about thoughts and motives in follow-up questions. The stories, then, reveal the children's Theory of Mind and their ability to use language about emotions, desires, and cognitions to explain the characters' actions.

**SEMANTICS:** Semantics adds another important dimension to the diagnostic evaluation. Word-learning is crucial for a language user, and it needs to be fast and efficient. In this domain the *DELV* evaluates several aspects of semantic functioning that are neglected in existing tests, and presents alternative strategies for distinguishing language problems. The Semantics subdomains are 1) fast-mapping, 2) verb and preposition contrasts, and 3) quantifiers. There are three properties of semantic functioning that the *DELV* taps into, all very different and unique to our work. The Semantics subdomains look at:

1) process: Can the child learn a new word easily from context?
2) lexical organization/retrieval, which may be more significant than size of vocabulary, and
3) two of the logical properties of the word *every*.

**Fast-mapping**

The idea underlying this subtest is that a typically developing child may have an impoverished or different vocabulary but, regardless of dialect, will still be capable of quick learning of a new word when given the opportunity. By contrast, children with language impairment seem to have special difficulties in the process of fast mapping new meanings from verbal context (Rice, Buhr, & Nemeth, 1990).

Verbs, as opposed to nouns, are especially revealing of the child's ability to use syntactic cues to learn their meaning (Gleitman, 1990). Previous work has shown that even three-year-olds will distinguish transitive from intransitive actions depending on the sentence context within which they are presented (Naigles, 1990; Fisher, 1996). For example, Fisher showed children two scenes, one with one actor, e.g. a boy spinning slowly on a stool, and another with two actors, a person spinning the boy on the stool. When the children were told, *He is mooping,* they were more likely to match the sentence to the picture with one actor (intransitive). When they were told, *He is mooping him,* they were more likely to pick the scene with two actors (transitive). Work in the UMass lab (Johnson, 2001) expanded the procedure to include transfer verbs like *handed* (*The mailman handed the letter to the boy*) and complement verbs, like *ask* (*The policeman asked the woman to stop the car*. )

On the *DELV*, children are trained first with real verbs and familiar actions so they will learn the general idea of the task. Then nonsense verbs are used in these sentence frames to describe pictures with novel actions. The pictures support two or more ways to interpret the activity in them, depending on the sentence used to describe it. The child then answers questions about the novel verb and its subjects and/or objects. For example, the child might hear, *The man is temming the ball.* Then the actors and props from the action are shown separately and the child is asked questions like *Which one is the temmer? Which one was temming? Which one got temmed?* In this way, the child shows which action she has associated with the verb and what kind of verb she thinks it is.

**Verb and Preposition Contrasts**

To tap into how well the children have their words organized, we took advantage of the different levels of relationships in verb "neighborhoods"; for example *walking* and *crawling* are both manners of *moving*. We asked whether the child could provide an appropriate contrast at the appropriate "level" in the hierarchy, in naming some common actions in flexible ways. The Verb Contrast task in the *DELV* is a verb "antonym" task, modeled after a Waxman and Hatch (1992) study. For example, for a picture of a girl licking a popsicle, we might say *This girl is not chewing the popsicle , she's…….*". An appropriate response is *licking*, but not *eating*, even though that is also true of the picture. However, given the prompt, *She's not drinking the popsicle, she's…….*" then *eating* is a better response than *licking*. In order to succeed at this task, children need to have some minimum number of verbs in their vocabulary, but more importantly, they must have those verbs organized into appropriate sub-categories and contrasts.

A similar format works to elicit prepositions as well, once again with the requirement that the child provide a contrasting, but parallel preposition to the one in the prompt. For example, if the child is shown a picture of a girl riding on a horse, the prompts might be: *She's not riding to a horse, she's riding ……* Expect: (*on a horse*). For the second prompt, one might have, *She's not sitting behind the saddle……She's sitting….*

Expect: (*on the saddle*).

**Quantifiers**

How children learn kinds of reference that are not unique lies at the core of language

knowledge, and it is therefore an important dimension of language evaluation. The *DELV* first explores whether the child can recognize a situation that exemplifies *every*. Shown a picture of 3 women in boats and a 4th woman on the beach, will the child say "no" (as he should) when asked whether *every girl is riding a boat*. (and "yes" if there no "extra" women). Another set of quantifier questions probes whether children know that *every* blocks a pronoun in a subsequent sentence from referring to its noun. For example, the "he" in *The baby watched the man. He played the piano* is ambiguous. "He" could be either the man or the baby, but in *The baby watched every man. He played the piano*, "he" can now refer only to "the baby," not to any or even all of the men in the picture. These are subtle facts about how quantity words work (Philip, 1995) that we all learn without being aware of it. The Developmental Milestone data from Phase II show us that typically developing children in this age range appear to learn them easily, but language impaired children take longer to understand them, so they add to the diagnostic power of the Semantics domain.

**PHONOLOGY:** The contrastive-noncontrastive model developed by Seymour & Seymour (1977) was applied to the development of a dialect-sensitive phonology assessment which uses a single scoring and test format, regardless of a child's dialect. Using pictured stimuli for memory support, the child repeats sentences containing the target words. Through extensive field research, based especially on analyses of the archive of AAE child speech gathered during a previous grant to Seymour and Roeper (NIH R01 DC 02172-04), candidate stimulus items were found that respect the phonotactics of AAE; i.e. consonants or consonant clusters are tested only in the initial or medial position of a word. (See also Haynes & Moran, 1989; Stockman, 1993, 1996.) The *DELV* Phonology domain draws upon the predictable and systematic nature of phoneme acquisition, but uses relatively difficult stimulus items to enhance the differences observable between typically developing and phonologically impaired children across the whole age range, 4 to 9 years (Stoel-Gammon & Dunn, 1985; Vihman, 1998).

**Non-word repetition**

A subsection of phonology, non-word repetition, combines phonetic segments common to AAE and MAE into nonsense syllables, which the child repeats after the examiner. This task highlights children's difficulty in processing the purely phonetic component of incoming speech, and has shown its usefulness in contributing to the identification of SLI in a process dependent, non-biased way (Campbell et al, 1997).

**PHASE II. Establishment of developmental milestones for AAE speakers on a range of both innovative and traditional language tasks** (and the establishment of developmental milestones for MAE speakers on the innovative tasks).

In Phase 2, we achieved the clear statement of the ages of expected mastery for a large number of deep aspects of grammar and of selected AAE features in AAE child speech. The selected behaviors, those discussed above under Phase 1, are consistent with developmental trends identified by theoretical linguistics and traditional clinical practice. For the collection of milestone data, 320 items, organized in 14 subtests, were professionally

drawn and published in an interim edition called the *Dialect Sensitive Language Test (DSLT)*. Dialect-screener materials developed under research grant NIH R01 DC 02172-04 (to H. Seymour, P. I.) were incorporated into the new stimulus manual and so comparable developmental milestones could be determined for the screening items as well. In a 10-month period, these materials were tested by 477 testers in all parts of the United States on 1258 children, 560 more than called for in the sampling design.

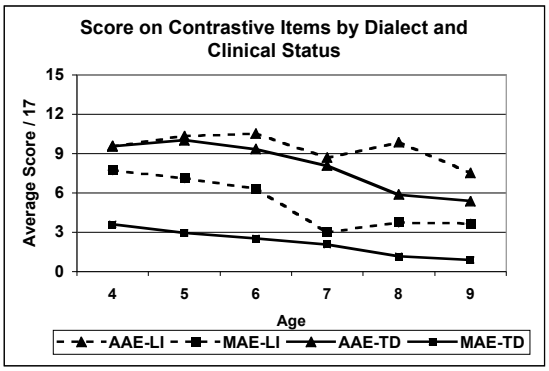
Three key elements of the data collection at this phase were that

1) it oversampled low-SES children (87% had parents with a high school education or less),
2) it oversampled language impaired children (33%), and
3) both dialect groups had both typically developing and impaired children, so all four subgroups were large enough to permit meaningful statistical comparisons.

Findings from the resulting database of responses were summarized in the Developmental Milestone report to NIH (UMass Working Groups, April, 2002). These analyses reinforced the appropriateness of the items for AAE speakers, and the items were found to be effective for MAE speakers as well. In particular, this database and the 120 charts developed from them for the Milestone Report provided graphic support for the strong claim underlying the contract (following Seymour & Seymour, 1977): that there exist a number of *contrastive* item types where the typically developing dialect groups showed radically different performance. More importantly, there were many *non-contrastive* items where typically developing children of both dialect groups performed the SAME, and both TD groups performed differently from the comparable Language Impaired groups. Also as part of the Tryout phase, an inter-rater reliability test was carried out with 25 AA children, tested by both an AA examiner and a White examiner, in counterbalanced order approximately two weeks apart. The results were analyzed for "decision consistency" between administrations, which averaged 77%. No systematic bias by ethnicity of the examiner was observed. This analysis was done after the 2002 Phase II report was submitted, but it is detailed in the *DELV-CR Manual* (pp. 84-86).

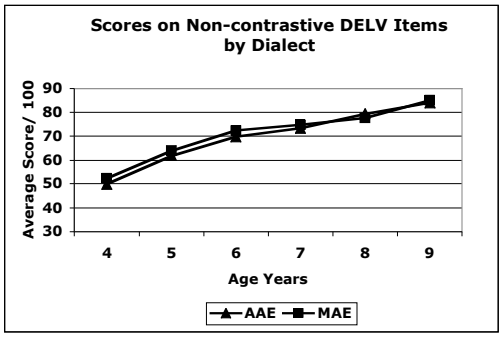
Thus, the developmental milestones for AAE child speech determined from the fieldtesting results showed us which features can be used to differentiate the groups by dialect and which <u>other</u> features can be used, for both dialect groups, to differentiate typically developing from language impaired. Examples of the charts from the milestone report (Figures 1 to 3) show the strikingly different patterns observed for contrastive and non-contrastive features. In Figure 1 of contrastive Identifier Item Responses from the Screener-Part 1, the TD AAE and TD MAE children, represented by the two solid lines, are significantly different across the age range on these AAE features. Note also how the <u>TD</u>-AAE and <u>LI</u>-MAE children (the two middle lines) pattern similarly on these measures, and both groups of AAE children show similar values for them. This shows how a high score on identifiers (AAE features) is ambiguous for the AA child: such features may be present because of dialect or because of disorder. These identifier items cannot make that diagnostic distinction.

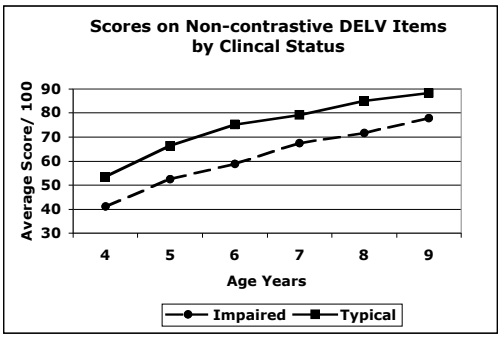
Figure 1.




Figures 2 and 3 are graphs representative of the many diagnostic non-contrastive variables detailed in the milestone report. TD children of both dialect groups are not different from each other, but they are significantly different from LI children of both dialect groups.

Figures 2 and 3.




In this framework, creating a test for AAE learners was not an exercise in "lowering the bar" so AAE speakers who failed traditional tests could succeed. In fact, the bar was raised for everyone. AAE and MAE learners who differed with respect to many elements of morphosyntax and phonology, nonetheless showed their equivalence in a series of subtle semantic, pragmatic, and syntactic constructions.

**PHASE III. Development and validation of the standardized (norm-referenced) test.**

At Phase 3, the assessment battery was reduced to the 150 most effective items from the four principal domains. Some items were eliminated because they were too easy or too hard; others because they showed bias against the AAE groups. One whole subtest was eliminated because it was less practical to administer than the others. Only 3 new parallel items had to be created. The end result is two tests, the Screener and the full Diagnostic test, that can be used independently or together.

For the standardization, the two *DELV* tests were bound in one book and new Record Forms and Administration Directions were produced. HAI reactivated their nationwide network of testers and testing sites, and in January 2003 the second data collection, for the standardization and validation of the *DELV* began, and it continued through March 2004.

In contrast to the Phase 2 data collection, which intentionally oversampled specific demographic groups, the population for the Phase 3 data collection aimed as closely as possible to match the categories of the U.S. Census (2000) for the general U.S. African American school-age population: that is, 45% mid-SES, 55% low-SES, not more than 7.5% language impaired, and 50% female; 51% from the South, 25% from the Midwest, 15% and 9% from the Northeast and West, respectively. In addition, data were collected for an extensive set of validity and reliability studies. The contract calls for 660 children taking a total of 950 tests, but in fact, 992 subjects took over 1500 tests. In addition, all 992 subjects took the AAE Screening test instead of the 30 stipulated in the contract.

The reliability and validity studies also went beyond the minimum requirements of the terms of the contract. To test the *DELV*'s test-retest reliability, the *DELV* was administered a second time to 100 children between two and four weeks after the first administration. In addition, several statistical analyses of internal reliability were performed. To test validity, several strategies were followed. First, there were two groups of impaired children, 140 language impaired (42 in the main standardization sample plus an additional 75 recruited for this validity study) and 30 phonologically impaired children for detailed comparisons of their performance on the relevant parts of the *DELV*. A second strategy was to compare *DELV* scores with performance on three existing tests: 1) a non-verbal IQ test, the *Naglieri Non-Verbal Abilities Test* (*NNAT*, Naglieri, 2003); 2) a language test, the *Clinical Evaluation of Language Fundamentals 4* (*CELF-4*, Semel et al., 2004); and 3) the articulation screener of the *Preschool Language Scale-4 (PLS-4*, Zimmerman et al., 2002), 82 children for the first test, 100 children each for the latter two tests. The goal of the comparison to the cognitive test was to confirm that the *DELV* is not a cognitive test. The goal for the two language tests was to examine the extent to which the *DELV* scores agreed with them. However, this represents a dilemma for the project. On the one hand, one expects the *DELV* to agree to some extent with the judgments of existing tests. But if the agreement is too high, it will not show the revolutionary nature of the *DELV*: it would be using as a "gold standard" the very tests the *DELV* is designed to improve on.

Therefore, a third study was undertaken to compare *DELV* performance to language samples collected from children within one month of taking the *DELV*. In order to have as full a picture as possible of the language function of children with a wide range of scores on the *DELV,* the number of language samples was increased from 20 as stipulated in the contract to 78. The design of this study also varied the ethnicity of the examiners to test the effect of that factor on both language samples and *DELV* performance. The language samples were collected by HAI examiners, and then transcribed and analyzed by the UMass/Smith team using *SALT (Systematic Analysis of Language Transcripts*, Miller & Nockerts, 1984, 2002) and *Computerized Profiling* (Long, Fey, & Channell, 2003) as well as several measures devised by the authors. In addition to evaluating clinical status, the

language samples offered an opportunity to validate the Language Variation Status scores from Part I of the Screener.

These studies together offer "a comprehensive examination of the reliability and validity of the *DELV* as an effective test for children who speak African-American English"-- as required by the contract. The information, including subject selection, materials, procedures, and results for the standardization process and the related reliability and validity studies for the 600 children ages 4;0 to 6;11, are detailed in Sections IIa and IIb, the subcontractor's Final Report, and IIc, the UMass concurrent validity study of the *DELV* and Language Sampling.