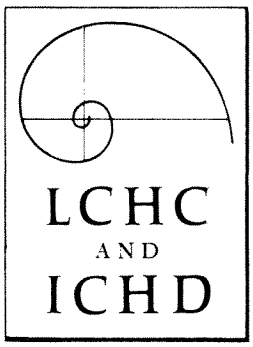

THE ROCKEFELLER UNIVERSITY

LABORATORY OF COMPARATIVE HUMAN COGNITION

AND

THE INSTITUTE FOR COMPARATIVE HUMAN DEVELOPMENT

Working Paper

No. 3

Constraints of Text and Setting on Measurement of Mental Ability

Judith Orasanu

The Rockefeller University

Constraints of Text and Setting on Measurement of Mental Ability

Judith Orasanu, The Rockefeller University

Most people take for granted that "To do well in school you have to be smart. If IQ tests select kids who do well in school, they must select kids who are smart."

The problem with this syllogism is that it has a companion that rarely is considered. For example: "Being smart isn't all it takes to do well in school; you also have to apply yourself," or "It doesn't matter how smart you are; if you miss half the semester, you'll have a hard time passing algebra." In short, common sense tells us that although being smart is important, a person can do well in school and not be too smart or can be smart and not do too well.

The same common sense carries over into the evaluation of IQ test performance; people are willing to assume that smart people generally do well on tests, but there is no one-to-one correspondence between test performance and real ability.

The possibility that a single test might wrongly estimate a child's ability would be of little importance if we could be confident that with repeated testing we could get an accurate assessment of a child's potential. However, the massive fact of population-group differences in both tested and school performances has produced a great deal of pressure for finding ways of determining real potential, stripped of any possible contaminating influences.

The demand for a pure measure of ability is understandable. Such a measure would greatly simplify analyses of several

---

* Alternatives to Standardized Testing. 1976. Symposium held at University of Pittsburgh, PA. Unpublished.

extremely volatile socio-scientific issues. Are there racial differences in intelligence? Is a child's intelligence determined by the age of three years? Can intelligence be modified by proper education? But no pure measure of intelligence exists; we only have measures of performance which summarize people's interaction with a test-giver and the problems that the test-giver poses. Rather than become embroiled in a futile debate over the validity of drawing inferences about "pure ability" from test performance, we will take a different approach.

The whole notion of a standardized method of testing will be reconsidered. If we focus on the fact that behavior in certain rather brief encounters (tests) predicts behavior in other extended encounters (schooling), what can we learn from the variability of children's behavior in the brief encounters that might help us to understand the factors that control their behavior in the domains where it really matters? If we can find factors which modify test performance significantly, might it be possible that similar modifications in the organization of instruction would have a similar effect? In other words, can modifications in standardized test procedures be useful in an analysis of intellectual performance and its relation to educational outcomes?

In order to evaluate these rhetorical questions, I depart rather sharply from two traditional psychometric assumptions. The first is that intelligence is a "thing" which can be measured, much as length can be measured. This implies that it has a relatively permanent, fixed nature which is manifest in

many behaviors. The second assumption is that in order to obtain an accurate measurement, standardized testing procedures must be employed. Thus, individual differences in scores can be attributed to differences in ability, not to inconsistent measuring procedures.

In this chapter, we will not assume that there is a fixed property of individuals corresponding to their intelligence, but rather will focus on the factors which modify performance on a given task. We will not assume that standardized procedures have the same effects on all individuals; rather, we shall pay particular attention to factors which might modify the behavior of one group, while leaving the performance of others relatively unchanged.

Careful consideration of the testing situation has led us to distinguishing three aspects of the interaction between tester and testee that bear on performance: text, setting, and context. (The term interaction is used here in both the active, process sense and in the statistical sense of a subjects-by-condition interaction.) The first aspect --"text"-- concerns the way in which the problem area of concern to the tester is represented to the child; the child must study and respond to a text. The second aspect is the "setting." In discussing this, we will depart from traditional psychological studies to compare behavior in standard test situations with naturalistic observations of abilities supposedly being measured by the tests.

The third interactional aspect, "context," refers to the

influence of motivation on performance, and the role of such variables as the race of the experimenter and the test atmosphere on performance. Discussion of context will be reserved for the next chapter.

The Text Aspect

The four components of text in a testing situation are: (1) familiarity of content; (2) alternative representations of the same problem; (3) the form in which questions are phrased; and (4) dialect.

Familiarity of Content.

Group differences in test performance would appear to be strongly influenced by differential familiarity with the items used in the tests. This issue is much more subtle than most people are willing to admit. Controversy over the issue has a long history and is by no means resolved (c.f. Eells, Davis and Havigurst, 1951, for an extensive, early treatment of this topic). At first it was believed that it would be possible to control for the familiarity of test items by careful selection (to screen out such gems as violin/cello in word analogies). But such simple strategems have not proved to be very successful.

Verbal test items. The problems in this area can be illustrated by considering a typical test item in object classification from the Wechsler Intelligence Scale for Children (WISC) in conjunction with research on the effects of item frequency in adult word associations (Stoltz and Tiffany, 1972).

One of the first items on the WISC similarities subscale is

to tell how an apple and banana are alike. The scoring manual gives us explicit rules for alloting credit to different answers:

APPLE-BANANA

2 points - A response stating they are both fruits.

1 point - Both food...Both round (or similarly shaped)...
Both have skin

0 points - Good for you...Taste alike...Both small...
Same kind of skin...

Why are responses like "good for you" alloted no score at all, while responses based on perceptual similarity are given part credit and those based on class membership full credit? There are two answers. First, empirically, it has been observed that older children are more likely than younger ones to give the responses based on class membership. As a result those responses are defined as more advanced and given greater credit. Also, at a specified age, children who give the high-score responses are more likely to do well in school. Second, theoretical rationales derive from studies of age-related changes in children's verbal behavior. Developmental theorists consistently assert that the course of concept development proceeds from similarity based on co-occurrence or perceptual features to abstract taxonomic relations (Piaget,1926; Werner,1948; Bruner,1957). Many studies have shown that when children of different ages are told to "say the first words that comes to mind" in response to such stimulus words as apple, there is an age-dependent change in the nature of responses. Young children (five or six years of age) are likely to respond

as if they were fitting the words into a phrase or sentence. For example, we might observe the following responses to stimulus words: apple-tree, run-home, or red-balloon. To be sure, we might also encounter sequences such as dog-cat or cow-milk, but the preponderance of young children's responses are of the type identified as phrase constructions.

Older children (eleven or twelve years of age) engage in relatively little phrase construction. They are much more likely to produce the following response: apple-orange, run-skip, or red-black. These responses belong to the same parts of speech and semantic class as the stimulus words.

This shift in the nature of children's verbal responses is widely believed to reflect a more mature language-processing capacity, which has many counterparts in other areas of the child's intellectual behavior. Such parallels have led developmental theorists to formulate a series of stages, or milestones, of intellectual development, such as Jerome Bruner's idea that children first represent information as part of an action, then as an image, and finally in symbolic form. All of this seems so common-sensical that it is difficult to fault.

We first began to suspect that there may be serious difficulties in the straightforward interpretation of these data when we read a study by Stolz and Tiffany entitled "The Production of Child-like Word Associations by Adults to Unfamiliar Adjectives (1972). Taking standardized norms of word frequency from a variety of sources,

Stoltz and Tiffany constructed two lists of words. One was made up of synonyms of the words on the other list. The only difference between the two was the frequency with which they appeared on the norms. The first list was made up entirely of high-frequency words; the second of low-frequency words. Some examples are: many/myriad; neat/fastidious; clever/ingenious.

One group of college students was presented with the first list, another group with the second. The instructions were the standard "say the first word that comes to mind" version of the free-association experiment.

College students who received the high-frequency list responded in the typical adult pattern. Their response words were from the same grammatical class as the stimulus words (many-few, neat-clean, etc.) But when the low-frequency words were used, responses were preponderantly "childlike." That is, responses were more frequently words which would be appropriate in a phrase containing two words (fastidious-housekeeper).

This result has raised a number of doubts about the usual interpretation of word classification and similarity studies. If the frequency with which a word is encountered controls the nature of adults' responses, would the same rule apply to children? And if the frequency principle applies to children, how can we use such materials to test ideas about mental (in this case, semantic) development independent of experience? We can be pretty certain about only one thing: older children have encountered any particular word more often than have their younger

counterparts.

What about children of the same age, but with different family backgrounds? The same problem applies. We know that children from different subcultural groups are exposed to different vocabularies, but we are not sure what the vocabulary or its frequency of occurrence might be. This makes it virtually impossible to equate frequency of occurrence of items on a test across groups of children.

This issue is irrelevant if all we want to know is a given child's point of progress toward giving responses expected of college adults, but it puts us in an impossible position for teasing out semantic word-processing strategies from frequency factors that affect the probability that those strategies will be used.

Nonverbal test items. Familiarity with content of a test item may function in an altogether different way to determine performance on a complex task. The previous examples have shown how familiarity with vocabulary may influence verbal performance on an IQ test. However, familiarity with nonverbal materials that are used to assess other cognitive skills may have a similar effect. That is, the likelihood that a person will demonstrate a particular skill may depend on the materials used to test that skill, rather than on proficiency at the skill itself. Yet, inferences are drawn about the skill, not about familiarity with the materials.

Cole, Gay, Glick and Sharp (1971) provide an instructive

example from their research with the Kpelle in Liberia, which led to a reinterpretation of the performance of American school children on the same task. They were investigating "inferential" ability in people aged 5 to 22 years who had varied amounts of schooling. Inferential behavior was defined as "spontaneous integration of two separately learned behavior segments to obtain a goal" (p. 204), following the work of Kendler and Kendler (1967), who originally devised the task. In the first of the two subtasks the children learned to obtain a marble or a ball-bearing by pressing buttons on differently colored panels of a metal box. In the second component, they learned which ball (the marble or ball-bearing), when dropped into a hole in a third panel of the box would result in the goal, a piece of candy. After the two subtasks were taught separately, a child was presented with the metal box and told to try to obtain the candy. The Kendlers had found that approximately 50% of American third-graders immediately integrated the components, and American college students almost always did. Among the Kpelle, however, only 25% of the subjects of all ages solved the problem immediately. With prodding and some trial-and-error behavior, approximately 65% of all age groups eventually demonstrated integration.

Working on the hypothesis that something specific to the apparatus was causing the difficulty, Cole and his colleagues ran a series of experiments in which they changed the materials to familiar objects, but maintained the identical procedure. To get the candy, the subjects had to learn to obtain red and

black keys from two differently colored matchboxes, and then to use one of the keys to open a locked box. When using these materials, 70 to 80% of the subjects, even young children, spontaneously integrated the components. To further locate the source of difficulty with the original apparatus, the investigators pitted stages of the first two studies against each other. For example, a key taken from the Kendler apparatus could open the locked box. Or the marble and ball-bearing were obtained from the matchboxes to operate the Kendler box. It was shown that subjects who performend poorly did so because they did not deal effectively with the first link in the problem. Only 10% of the children solved directly when they had to operate the Kendler box first, compared to 80% when the first phase involved the matchboxes. Once they got started on the right track, the inference proceeded adequately, but the need to initiate a solution on an unfamiliar instrument seemed to impede the whole process.

For comparison purposes, groups of 20 American first-graders were also given the series of cross-over experiments just described. The results were identical to those obtained with the Kpelle. As a consequence of the early Kendler work, the conclusion had been drawn that young children lack the ability to abstract from the two components the common element required to solve the problem. Clearly, there is nothing wrong with the mediation or integration abilities of the younger children, but these experiments demonstrate how engagement in the process is tied to the specific materials used. Unfortunately, the experiments do not explain

why familiar materials make a differences to the child's getting started on the right track.

Thus, we are faced with the dilemma of attempting to assess cognitive skills or processes when we don't know whether the materials used to measure them are equally familiar to all people tested. Familiarity of meaningfulness may operate in different ways on various tasks for different groups of people, completely invalidating the notion of standardization.

With respect to IQ tests themselves, an attempt to get around the problem of content familiarity was the development of "culture-fair" tests (e.g., Raven Progressive Matrices, Cattell Culture-Fair Test, Davis-Eells Games). Most of these are based on perception of relations among abstract forms. The idea was that the contest would be equally unfamiliar to all children, thereby making the tests more fair.

If differences between ethnic or socioeconomic groups on standard IQ tests are the result of differences in familiarity with the content, one might predict diminished differences between groups on the culture-fair tests. However, the differences have not disappeared (Higgins and Sivers, 1958; Jensen, 1970). Interpretation of these findings is problematic, however, since we have little understanding of what is being measured by these abstract tests. Based on the examples cited above, in which familiarity of contests was varied (Stoltz and Tiffany's word association study and the integration task among the Kpelle), it is clear that performance on tasks using unfamiliar materials precludes conclusions about any

underlying cognitive processes or skills.

However, group differences were maintained, even when abstract culture-fair tests were used, suggesting that factors separate from knowledge about the "textual" aspect of the tests contribute to group differences in standard IQ test performance. This topic will be elaborated in the later discussion of "setting."

Alternative Representations of the Same Problem

One frequently-encountered class of problems on intelligence tests requires that children categorize items, or choose among a set of items to find subsets that "belong together." These items discriminate among children according to age, and have been used extensively in studies of the retarded and brain-damaged. The finding usually reported is that younger children are likely to respond to such tasks by choosing items that look alike in some way, whereas older children are more likely to choose items that share some common function or are part of the same taxonomic category (see, for example, Bruner, Olver, and Greenfield, 1966). Retarded children, or adults with brain damage, respond in the same way as younger children.

These results are commonly interpreted to reflect a different organization of word meanings in young children, who are presumed to lack the taxonomic or functional categories by which the items could be organized. As a consequence, it is said that young children rely on the preceptually given, concrete features of problems in their thinking.

Birch and Bortner (1970) were dissatisfied with this interpretation of the standard developmental and normal-abnormal differences. They noted that, in the typical study, the child was presented with an array of objects which could be grouped according to a variety of criteria (perceptual, functional, etc.). It seemed possible that, when young children were presented with such an array and told only that they should choose objects that "go together," they might be able to base their choices on functional or taxonomic relationships, but that they would not do so, either because of their interpretation of the instruction or because the physical similarity was simply more compelling.

Birch and Bortner devised a clever experiment which tested their notion that children who did not choose on the basis of functional or taxonomic relationships nevertheless had the capacity to do so. They first presented a child with a target item and then contrasted it with three alternatives, one of which "belonged" with the target. For one group of children, the matching of target and alternatives could be accomplished on the basis of either functional or perceptual properties. For example, a red button might be presented as the target, with a spool of white thread, a red lipstick case, and a blue poker chip as alternatives. Consistent with previous findings, older children tended to pick the spool of thread to match with the button (a functional pairing), whereas younger children chose either the lipstick case (color) or the poker chip (shape).

A second group was presented exactly the same target item,

but with a set of alternatives that precluded obvious physical matches. In the example used above, the red button remained the target, but the alternatives were a blue nut, a spool of thread, and a cup. Under these conditions, there was a marked increase in the number of choices of objects based on a functional/taxonomic principle; the children not only increased their choice of the spool of thread, but justified their choices in appropriate terms (e.g., "You use them to sew with").

Birch and Bortner note that "This variation in the alternatives presented to the child to assess his classificatory skills markedly shifts our interpretation of children's choices in our initial test. It no longer seems appropriate to conclude that they don't 'have' functional or taxonomic categories where older children clearly do; which immediately poses a new question--why?"

There are several possible answers to that question. It might be that there is something inherently "attracting" in the physical relationships which induces the children to base their judgments on them, given a choice. It might be that physically similar objects actually appear more similar to young children than do stimuli related by function. Or it might be that the children simply interpret the instruction "pick those that belong together" differently from adults.

Birch and Bortner, who favored the first interpretation, did not follow up these various possibilities in a systematic fashion, but we can get some information about the possibilities from

other research that varied the components of what is conceived of as a single problem.

A recent study by Cole (1976) is relevant here. Working with 3 to 5-year-old Black children in a Head Start center, Cole used geometric blocks or wooden dolls in a study of concept learning and transfer. A salient aspect of his results, in the light of Birch and Bortner's findings, was that the older children learned the transfer problem more quickly than did the younger children when geometric blocks were used (the standard finding), but there were no age differences in rate of transfer when the wooden dolls were the objects. These and other results of Cole's study strongly suggested that age differences in conceptual transfer were not the result of any generalized deficit in the younger children, but in the greater vulnerability of the concepts they had learned when they were placed in a situation that made them confront conflicting information. Under such conditions, being allowed to work with stimuli that represented meaningful, as opposed to arbitrary, classifications, seemed to help them maintain their use of a recently learned conceptual scheme.

A similar point is made in a study conducted by Abramyan (reported in Luria, 1961), whose results implicate differences in the meaning of the task as a source of performance differences between younger and older children. Abramyan instructed children 3 to 7 years old to squeeze a bulb in response to pictures presented to them. If a red circle on a gray background

appeared, the child was to squeeze with the right hand; if a green circle was shown on a yellow background, with the left hand. After the child had shown understanding of the instructions by squeezing appropriately for a series of trials, Abramyan presented test trials, on which circles were shown on the opposite backgrounds. All of the children made their responses on the basis of the color of the circles, ignoring the backgrounds.

Abramyan then instructed each child to try to respond in accordance with the background color, not the color of the circle. These instructions worked adequately for the 5 to 7-year-olds but,the three- to four-year-olds were inconsistent. It would appear that they were unable to overcome some strong perceptual pull of the figure, a difficulty that appears analogous to making functional choices in the face of competing physical-choice possibilities.

However, Abramyan added a condition to her experiment. The circles were replaced with crude outline figures representing airplanes. The instructions were also slightly modified. The child was told that she/he should squeeze with the right hand when the red airplane was on the yellow background "because the plane can fly when the sun is shining and the sky is yellow"; analogously, the child was told to squeeze with the left hand when the green plane was presented on the gray background "because when it's rainy the plane can't fly and has to be stopped." Under these conditions, the three- and four-year-olds

could be induced to respond to the background as well as to the figure. Whereas the initial observations, if used as a standardized test, could easily have led to the conclusion that young children are rigid and have difficulty in switching attention from dominant to less-dominant features of a stimulus array, the fuller experiment strongly suggested that this is not the difficulty. Rather, the minimal instructions which are adequate for the older children and which seem reasonable to the adult do not provide the young child with an equivalent understanding of the task.

All of these results suggest important ways in which identical stimulus sets are not interpreted identically by children of different ages. So far as we can discern, the items used are familiar to all of the children; they can all be named. But they do not function in equivalent ways that fit well with standard interpretations of differences in ability. Rather, it seems that special care must be taken to insure that the tasks are interpreted and responded to in comparable ways by different subjects before we can consider issues of differential ability.

The form of questions

A very different example, this one directed at the child's interpretation of the tester's questions, concerns the relation between problem-solving and verbal explanations in young children. Many studies with young children (c.f. Stevenson, 1971, for a review) have suggested that they experience difficulty

when asked to explain the basis for their responses in an elementary concept learning study, even when their choices indicate mastery of the problem.

Marion Blank (1975) has convincingly demonstrated that this apparent verbal deficit can result from a systematic difference in the child's interpretation of such questions as "Why did you choose that block and not this one?" in reference to, say, One red and one blue block. She surmised that children interpret this question the same way as an adult would interpret "Why did you sit down on the chair?" That is, it would become an occasion to explain the motivation for the choice (" Cause I liked it"), not an occasion for a physical description of the class of correct choices. In the above case, an appropriate answer would require stating the color of the block.

By varying the form of her question and the conditions under which the question was asked, Blank both laid bare the nature of the task from the child's point of view and demonstrated the existence of heretofore-doubted verbal abilities in young children.

Her procedure was the soul of simplicity. Four groups of preschoolers were taught an elementary visual discrimination -- between a circle and a triangle, for example. Half the children were asked about the reasons for their choices with the objects present, the others with the blocks removed. Half of each of these groups were asked "How did you know which block was correct?"

The others were asked the more explicit question, "Which block was correct?"

If the children were asked the "how" question, virtually all answered in terms of an internal state ("I wanted to") or said they did not know. If they were asked which block was correct, they tended to point at the object if it was present, but gave correct verbal descriptions if it was absent.

Blank's interpretation of her results is particularly germane to our concern:

> The seemingly poor performance on the "how" and "why" questions does not indicate inadequacy on the children's part. Their responses to these questions were not random, but rather were systematically different from their responses to the "which" question (i.e., the latter led to pointing and attribute description, the [former] to qualities or actions of the child). This differentiation suggests that "why" and "how" hold a definite meaning for children, albeit a meaning different from that held by adults. Interestingly, the children's interpretation of "why" and "how" seems quite reasonable. The initial responses of adults in this setting might well be similar to those of the children. Adults, however, would probably recognize that issues of motivation and skill are trivial in this context. Therefore, they would reinterpret the question so that it better represents what they believe the experimenter "must be driving at." In other words, they take what is basically an unreasonable question and turn it into something more appropriate to the task at hand.

Little systematic attention has been directed to the problem of how children of different ages or from different backgrounds interpret questions or instructions found on

standardized tests. It may be instructive to do so.

Dialect effects

An aspect of the situation which relates both to how the questions are asked and familiarity of the materials is the dialect used in test administration. Phonological, syntactic, and vocabulary differences may increase the information-processing load for a child who does not speak standard English. A second consideration, which will be developed in the next section of this chapter, is the affective or motivational effects produced when the child is tested by someone who does or does not speak the same dialect. Sociolinguistic differences in the use of language, particularly question-answer sequences, may also cause difficulties. Dialect is a topic that has aroused much concern among educators, and deserves to be reviewed here.

The foregoing discussion has focused on the effect of familiarity with the content of test items, but concern has also been expressed in many quarters over the possible effects of the test being administered in a dialect different from the child's own. If, because of speaking a dialect different from that of the test administrator, one child interprets instructions or questions differently from another child whose dialect is the same as the tester's, the answers of the first child may be systematically different and scored as wrong. Mercer (1973) has shown that bilingual Spanish children in the Southwest obtained higher scores when the tests were administered in Spanish, than in English. But this is not always the case

for children who are bilingual or bidialectical. Darcy (1963) reports that bilingual Spanish children did not always improve when they were tested in Spanish. The discrepancy is not easy to explain. Because they were bilingual, the children may have been exposed to a culture that differs significantly from the normative one on which the tests were originally standardized. The children may not have had the opportunity to learn the specific verbal content of the tests. That is, they may know English well enough to navigate activities and to understand the tasks, but might not be familiar with the content required to do well on similarities, analogies, vocabulary, or comprehension items.

On the other hand, being bilingual or bidialectical may result in an access or "production deficiency." That is, information of skills may be available which are not demonstrated because of a mismatch between the testing language and the child's preferred dialect. Williams and Rivers (1972) have hypothesized that when a test is administered in a mismatching dialect it can be considered to contain a high proportion of "noise." This precludes adequate activation of the child's linguistic-cognitive system, which is required if appropriate answers are to be produced. To test this hypothesis, they translated the Boehm Test and Basic Concepts, which assesses children's knowledge of time, space, and quantity, into nonstandard Black dialect. Examples of standard and nonstandard versions are as follows:

	Standard Version	Nonstandard Version
(1)	Space:	
	Mark the toy that is behind the sofa.	Mark the toy that is in back of the couch.
(2)	Quantity:	
	Mark the apple that is whole.	Mark the apple that is still all there.
(3)	Time:	
	Mark the boy who is beginning to climb the tree.	Mark the boy who is starting to climb the tree. (Variations may be used, as: about to, getting ready to)

Children were required to point to one of four pictures that corresponded to the verbal description. Both standard and nonstandard versions of the test were administered to kindergarten, first- and second-grade, poor, Black children. Performance for all these groups was significantly higher on the nonstandard version and was equivalent to scores obtained by middle-class or upper-class children who were given standard instructions. The results indicate that the children understood the concepts but had difficulty with the vocabulary. However, it is not possible to tell from the experiment whether the language changes altered the items to be dialect-specific or were simply easier for all children. Wolfram (1974), among others, has criticized the language of intelligence tests as being "formal" and far removed from speech used in everyday talk. Difficulty with the test

items would be expected to increase with the difference between test language and everyday language. Johnson (1974) has shown that there is a higher correlation between language used in spontaneous, casual conversation and test situations for White children than for Black children, suggesting that the test language would cause greater difficulties for Black children. However, Williams and Rivers' experiment lacks the appropriate White control group, which would permit an evaluation of whether the scores of all children increase with the nonstandard version, or whether the improvement is restricted to the poor, Black children.

If performance on the Boehm test is dialect-specific, speakers of standard English would be expected to do more poorly on the nonstandard version. Such a result has been obtained by Weener (1969), Baratz (1969), and Hall, Cole, and Reder (1975) in tasks that required subjects to imitate sentences or retell stories which were presented in a dialect that did or did not match their own. The typical finding is that more information is correctly recalled by Black children when the sentences or stories are presented in nonstandard Black English (regardless of the race of the speaker), and that White children recall more when the materials are spoken in standard English. To the extent that the sentence or story recalled depends on comprehension and encoding of the meaning of the stimuli, a familiar linguistic structure may facilitate performance, as Williams and Rivers suggest.

Facilitation by matching dialect is a far from universal finding, however. Quay (1971, 1972, 1974) has administered nonstandard English versions of the Stanford-Binet to Black children ranging from kindergarten through sixth grade, and has found no improvement in test scores compared to those obtained with standard English administration. An item analysis indicated no superiority for the dialect version, even on those items which were most language-dependent, such as similarities, vocabulary, or comprehension. However, the information required to answer the question and the actual content of the items were not changed, so the dialect in which the instructions were presented and the questions asked may not have been important.

Quay's results suggest that the poorer performance by Black children on IQ tests is not, in general, an information-processing problem; that is, the Black children do not have difficulty understanding the questions. This conclusion is supported by Hall, Cole, and Reder's (1975) study of story-recall, in which children's comprehension and retention of information from the stories were not affected by the dialect in which they were read, even though free recall was. Probe questions were asked after recall, and all children demonstrated that they knew more than their recall indicated. Their comprehension did not depend on a matching dialect. Peisach (1965) reached the same conclusion when she tested Black children's comprehension of a White teacher's standard English speech. She found that the children understood the

sentences perfectly well, as judged by their ability to produce a single word which completed sentences appropriately. Levy (1972) has also concluded that the Black children he tested who speak nonstandard English were aurally bidialectical.

Thus, children who speak a nonstandard dialect, but who are exposed to standard English in the da ssroom (or on television), may be equally able to comprehend both standard and nonstandard speech. However, their verbal production may vary as a function of the dialect of other speakers. The reasons for this variability are probably social, rather than linguistic. For example, if a child's speech is systematically criticized as wrong and "put down" in the classroom, the child may be reluctant to participate in a verbal test. Or the child's use of language, not the phonological of syntactic structure, may differ from that required to do well on the test. As Cazden (1970) has observed, "Sociolinguistic interference from constrasting communicative demands outside and in school are almost certainly more important than grammatical interference."

That is not to say that dialect does not play an important role in certain classroom activities, such as learning to read, where both comprehension and production factors are involved. Phonemic and syntactic "mismatches" are probably compounded by sociolinguistic factors.

The evidence suggests that most "language" problems in the test situation are sociolinguistic, not purely linguistic, except for clear cases of children who do not speak English. Questions

of language use in testing will be further explored under "The Setting."

Summary

The use of standard tests to diagnose intellectual skill or ability has been considered here. Typically, conclusions about individual differences in ability are drawn from such test scores. The assumption implicit in the use of tests for this purpose is that differences in performance on standard tests, administered under standard conditions, reflect differences in underlying ability. However, a review of research in which "text" factors have been varied indicates that it is impossible to draw such conclusions from performance on a single test. Variations in familiarity with the materials, problem configuration, and the way in which questions are asked have been shown to influence performance on tasks, frequently providing evidence of an ability that was not thought to be "there." These factors may influence the child's understanding of the task or the kind of answer the child thinks is appropriate.

However, in other cases, changing some textual aspect of the interaction has not reduced differences as expected, e.g., the shift to culture-fair tests or Black dialect in administering IQ tests. These failures to reduce differences between subcultural groups lead us to believe that the source of the differences cannot be located exclusively in the test. Some

other aspect of the social interaction between the child, the test, and the tester may operate differentially for various groups.

The Setting

In the previous sections, we have shown how intellectual performance can vary with changes in the test. This section will address the issue of the setting from which the sample of behavior that constitutes the test is taken. The term setting, as used here, needs further specification. We are concerned with the situation as it is defined by the child, that is, what kind of task the child perceives her(him) self to be engaged in. The major distinction is between assessment carried out in special circumstances (like schoolrooms) using predesigned instruments (like the WISC), and assessment that samples children's behavior in nonschool circumstnaces (like at home), using nonreactive, observational methods for data collection.

It is important for us to consider both aspects of the setting -- its formality and the type of observation or measuring instrument used -- because changes in one or the other alone may not be sufficient to provoke a change in behavior, whereas changes in both aspects may.

For example, recent attempts have been made to assess the effects of changing the setting in which a standardized test is given by carrying the test to the child's home (Mehan, 1973; Roth, 1974). Generally, such changes seem to have little effect on children's test performances. In contrast, Mishler (1975)

observed children's speech in the classroom, contrasting their performance when engaged in conversation with peers and with adults. Informal peer conversations were considerably more complex than were the teacher-dominated "instructional dialogues." In this instance, the setting remained constant, but the participants (and presumably the child's definition of the task) varied; so does the evaluation of the child's linguistic ability.

A firm grasp on the role of situational factors in test performance is a clear prerequisite for evaluating the generalizability of test performance beyond the rather narrow confines of schooling. It is also central to the question of precisely what the tests measure. If it can be shown that, outside of the testing situation, children possess and use abilities which they are assumed to lack on the basis of their test performance, our diagnosis and prescriptions for their future school experiences ought to be modified appropriately.

Unfortunately, research on situational variability in the application of intellectual skills is exceedingly scanty. One reason for the lack of data is the great difficulty, outside of tests, in specifying exactly what intellectual work is being done. For this very reason, psychologists at the turn of the century resorted to experimental methods for the study of thinking. What little data we do have comes mostly from anthropologists and linguists, or from linguistically

oriented psychologists. This work is relevant to our concern with intelligence testing, because tests of language ability generally correlate quite handsomely with IQ tests. As the following discussion should make clear, it is also relevant because many of the linguistic behaviors observed in formal language testing would be described as typifying subnormal intellectual performance according to standardized intelligence-test criteria.

One of the most provocative examples of situational variability in linguistic performance is provided by William Labov (1969). The following transcript of an interview with a Black boy from New York City is typical of the verbal output obtained in formal interviews.

> The boy enters a room where there is a large, friendly, White interviewer, who puts on the table in front of him a toy and says: "Tell me everything you can about this." (The interviewer's further remarks are in parentheses.)
>
> (12 seconds of silence)
> (What would you say it looks like?)
> (8 seconds of silence)
> A space ship.
> (Hmmmm.)
> (13 seconds of silence)
> Like a je-et.
> (12 seconds of silence)
> Like a plane.
> (20 seconds of silence)
> (What color is it?)
> Orange (2 seconds) an' whi-te.
> (2 seconds) an' green.
> (6 seconds of silence)
> (And what could you use it for?)
> (8 seconds of silence)
> A je-et.
> (6 seconds of silence)
> (If you had two of them, what would you do with them?)

(6 seconds of silence)
Give one to some-body.
(Hmmm. Who do you think would like to have it?)
(10 seconds of silence)
Cla-rence.
(Hmm. Where do you think we could get another one of these?)
At the store.
(Oh ka-ay!)

The length and pacing of this child's responses make him appear to be extremely dull. Labov assumed, however, that the child was working very hard to avoid saying anything that could get him into trouble, while providing those minimal responses needed to satisfy the interviewer. Repeating this exercise with a Black interviewer brought no change, nor did a change to an "exciting" topic. However, Labov reports that enormous changes took place in the child's speech when the interview was transformed into a partylike situation; the Black interviewer sat on the floor with the child, his friend, and a bag of potato chips. Taboo words and topics were introduced into the conversation. In this social context, the boy who previously had responded in monosyllables entered eagerly into the conversation. Rather than using language in a minimal, defensive way, he now employed it to compete actively with his friend for the floor, to defend his reputation, and to set the record straight regarding a fight he had been involved in.

Labov believes that children's speech is controlled by their perceptions of the power relations in the situation. The child in the example had probably experienced negative consequences for saying the wrong thing when being interrogated by an adult

in school situations. For our discussion, the central point is the extreme nature of the changes needed to discover the circumstances in which a "non-verbal" child will display his skill.

Another example of variability in the complexity of speech, this time with very young children, is provided from research by Dowley (see Hall, Cole, Reder, and Dowley, 1976). Dowley took three- and four-year-olds enrolled in a Head Start program in New York City on trips (accompanied by a tape recorder) to the local supermarket, where they discussed the food they saw. Upon returning to the classroom they were asked to tell their teacher about the trip. Speech in the supermarket was compared to their retelling of the event in school.

In the informal, supermarket setting the average number of words was greater, the percentage of questions attended to was greater, and the average number of words in response to a question was higher. Despite quantitative differences, language used in the two situations was qualitatively similar in several respects: neither the form of utterances (questions, commands, statements, or assertions) nor the content they expressed (want/need, family-related, love/like) differed drastically across the two situations.

Other lines of work that rely heavily on observation of naturally occurring conversation bolster Labov's point. Houston (1970) has identified two "registers' or styles of speech regularly employed in school and nonschool situations by rural Black children in Florida. Linguistically, the "nonschool

register" is more complex because of the children's employment of a greater number of options. It appears that the children appreciate that the norm for school speech is standard English, but in trying to produce it they restrict the range of options they actually use, yielding what sounds like deficient speech. Non school speech is the preferred mode for these children. They have not yet achieved the degree of proficiency in the two registers which would permit them to switch at will. Not only their syntax, but also their manner of speaking varies with the situation. Speech in the "school register" is slower, differently pitched and stressed, and relatively emotionless, compared to the non school register. Obviously, conclusions about the level of the speech development of these children will vary, depending on which register is sampled.

Labov and Houston's examples highlight the fact that how children talk depends on whether they perceive themselves to be in a situation where they will be evaluated on the basis of what they say. Evaluation may be done via a formal test, but it is also present in everyday classroom activities when the teacher calls upon a child to answer a question in front of the class. In trying to determine whether tests are equally valid estimaters of ability for all children, we need information concerning subcultural differences in attitudes to evaluation and public performance. Three examples in which children's "normal" behavior is clearly at odds with what is expected of them in the test/classroom are instructive.

Susan Phillips (1972) has described the speech of Indian children from the Warm Springs Reservation in a variety of school and community settings. Standard reading scores for these children are generally far behind grade level. When the teacher calls upon one of the children to answer a question in class, she is likely to get no answer and little volunteering of information by other Indian children unlike what frequently happens in white middle-class schoolrooms. However, Phillips observed the same children under four alternate types of classroom organization. She found that they demonstrated cooperation, involvement, initiative, and verbal communication when they were allowed to organize their own activities in small groups. When working together on projects, their language was highly adequate and certainly very different from their responses to the teacher's direct questions. Phillips hypothesized that the children feel threatened when the teacher directs the class from her position of authority, because they are not treated like that in their communities. Outside of school, children are included in adult activities as silent observers and are expected to try their hands at adult tasks, publicly demonstrating their competence only after having first succeeded in private. The teacher's request for public performance as part of the learning experience is not consistent with the rules of participation in their community.

Consequently, the children collectively resist pressure

to demonstrate their skills. Teachers who are most "successful" in terms of getting the children actively involved in classroom learning are those who spend little time in front of the class soliciting unwilling answers from resistent children, but who allow the children to work in small groups where they can determine for themselves when and how they will participate.

Individual competition and striving for leadership positions is inimical to the Warm Springs Indians' value system. "Shining" in class is, therefore, a "privilege" to be avoided. On the other hand, pride in group accomplishment and competition among teams is common and engaged in vigorously.

Boggs (1972) presents a similar pictures of indigenous Hawaiian children. Most of the 14-year-olds he observed were reading at the second-grade level. In class, these children would not respond individually to questions posed by the teacher, but would blurt out the answers as a group without being asked. They were also happy to volunteer information in the form of narratives to an adult whom they perceived to be receptive. That they understood questions was obvious from their production of them in conversation with one another, and also by their responses to questions in casual conversation with adults. Casual conversation apparently was not perceived as an evaluative situation. In contrast, Boggs observed that the children responded with suspicion when singled out for individual attention by an adult. "Why do you want to know?" was the reply to a question Boggs asked early in his observations.

A third example of a mismatch between standard school routines and nonschool language practices is provided by Lein (1975). Lein's work was carried out among migrant farm workers in the northeast United States. In class, many of these children participated little and were barely passing most subjects. Lein observed that both the amount of talk (number of words per minute) and complexity (number of words per utterance) were greatest when the children were at home with their peers and with no adult present. Intermediate levels were produced at home with adults present, and the lowest were produced in the classroom. Speech measures were also obtained in the classroom for students identified by the teachers as the "best" students. Although classroom speech of the migrant children was far below that of the best students, their speech with peers was as good or better than the best students' classroom speech.

One factor that may contribute to the children's low level of speech in the classroom is a discrepancy in the use of language for evaluation in school and at home. Aside from routine threats, such as, "I'll beat you," which carry no real import, very few actual threats or evaluative statements are addressed to children by adults at home. When a child is given a task to perform, no conditions are placed on the outcome in the sense of extrinsic reward or punishment. If performance is adequate, there usually are no consequences at all, except perhaps for a comment acknowledging that the job is completed. However, if the performance is inadequate, the child is simply told to repeat it

until he/she gets it right: for example, "Those dishes ain't washed. Go wash those dishes."

True negative evaluative comments are reserved for instances when the adult wants the child to stop whatever she/he is doing. For example, if a child is rude, the parent might say, "That's not right. Hush now," clearly directing the child not to repeat the behavior. Lein reports that such comments are not used to criticize a child's attempts at new activities. On the other hand, teachers frequently use negative evaluation when they want children to try a new task again. Obviously, the intent of the teacher's evaluative comments is not consistent with the parent's intent, and the child is likely to misinterpret the teacher. The net result is that the child may be discouraged from displaying new skills before the teacher. In school, threats of consequences are supposed to be motivating, but in fact may have the opposite effect for these children.

Conclusions

From the work reviewed, it seems safe to conclude that there is great situational variability in the manifestations of linguistic ability for at least the populations studied. Children who seem grossly deficient in classroom or test situations appear normal and competent elsewhere.

This, taken by itself, ought to urge great caution on the interpretation of test scores which involve face-to-face interaction, even when the content of the test is purported to be "nonverbal." It is clear, even from our bare descriptions of

a few examples, that social, emotional, linguistic, and intellectual factors are inextricably bound together in producing a child's language and IQ test performances.

These observations, by themselves, do not constitute the basis for alternatives to standardized testing, at least not at our current level of understanding and technology. They are consistent with efforts to evaluate children's social competence, but the requirements for testing social competence preclude reliance on standardized tests or even on school settings as an appropriate source of observations.

There are also many fundamental theoretical questions left unanswered in this work. We need to know much more about the relations between school and nonschool language and intellectual demands as they relate to such population characteristics as ethnic group origin and socioecomonic status. It is generally agreed that the "mismatch" between school and home is greater for some groups than others. But little observational work that traces children in a variety of situations, including interactions with significant adults and peers, has included the population comparisons on which such assumptions rest.

We also need to know much more about the intellectual demands of conversation in different settings. It is Blank's (1975) thesis, for example, that educational discourse is fundamentally different from informal conversation. Insofar as this is so, it renders the samples of informal speech provided by Labov and others fundamentally uncomparable with the samples used in

educational and test settings.

This research also urges on us a more complex view of the influence of the factors we have reviewed earlier in this chapter. Such variables as the way questions are asked, the race of the experimenter, friendliness, and dialect all help children to define the situation in which they are asked to perform. The evidence is overwhelming that these definitions are arrived at and responded to in ways that involve much more complex interactions than current psychological analyses can capture. Moreover, the same factors will define different situations for children from different backgrounds.

The weight of these difficulties makes us doubt the possibility that standardized tests are valid indicators of intellectual ability for all children in our culturally heterogeneous society.

Bibliography

Baratz, J. A bi-dialectal task for determining language proficiency in economically disadvantaged Negro children. Child Development, 1969, 40, 889-901.

Birch, H. & Bortner, M. Cognitive capacity and cognitive competence. American Journal of Mental Deficiency, 1970, 74, 735-744.

Blank, M. Eliciting verbalization from young children in experimental tasks: a methodological note. Child Development, 1975, 46, no. 1, 244-251.

Boggs, S.T. The meaning of questions and narratives to Hawaiian children. In Functions of Language in the Classroom, C.B. Cazden, V.P. John, D. Hymes (Eds.) New York: Teachers College Press, Columbia University, 1972.

Bruner, J. Going beyond the information given.In Contemporary Approaches to Cognition: A Symposium held at the University of Colorado. Cambridge, MA: Harvard University Press, 1957.

Bruner, J., Olver, and Greenfield, Studies in Cognitive Growth, New York: John Wiley, 1966.

Cazden, C. The situation: a neglected source of social class differences in language use. Journal of Social Issues, 1970, 26, 35-59.

Cole, M., Gay, J., Glick, J.A. and Sharp, D.W. The Cultural Context of Learning and Thinking. New York: Basic Books, 1971.

Cole, M. Probe trial procedure for the study of children's discrimination learning and transfer. Journal of Experimental Child Psychology, 1976, 22, no. 3, 499-510.

Eells, K., Davis, A., Havigurst, R.J., Herrick, V. and Tyler, R.W. Intelligence and Cultural Differences: A Study of Cultural Learning and Problem Solving. Chicago: University of Chicago Press, 1951.

Hall, W., Cole, M. and Reder, S. Story recall in young black and white children: effects of racial group membership, race in experimenter, and dialect. Developmental Psychology, 1975, 11, No. 5, 628-634.

Hall, W.S., Cole, M., Reder, S. and Dowley, G. Variation in young children's use of language: Some effects of setting and dialect. In Discourse Production and Comprehension, R.O. Freedle (Ed.). Hillsdale, NJ: L. Erlbaum and Assoc., 1976.

Higgins, C. and Sivers, C.A. A comparison of Stanford-Binet and Colored Raven Progressive Matrices IQs for children in low socioeconomic status. Journal of Consulting Psychology, 1958, 22, 465-468.

Houston, S. Competence and performance in child black English. Language Sciences, 1970, p. 9-14.

Jensen, A.R. Comparison of "Culture-Loaded" and"Culture-Fair" tests. In A.R. Jensen and W.D. Rowher, An Experimental Analysis of Learning Abilities in Culturally Disadvantaged Children. University of California, Berkeley, 1970.

Johnson, D.L. The influences of social class and race on language test performance and spontaneous speech of preschool children. Child Development, 1974, 45, 517-521.

Kendler, T.S. and Kendler, H.H. Experimental analysis of inferential behavior in children. In Advances in Child Development and Behavior, L.P. Lipsitt and C.C. Spiker (Eds.) Vol. 3. New York: Academic Press, 1967.

Labov, W. The logic of non-standard English. Part II. In Language in the Inner City: Studies in the Black English Vernacular. Philadelphia: University of Pennsylvania Press, 1972, pp. 201-255.

Lein, L. "You were talkin' though, oh yes, you was". Black American imigrant children: their speech at home and school. Council on Anthropology and Education Quarterly, Vol. VI, No. 4, 1975.

Levy, B. Dialect proficiency and auditory comprehension in standard and black non-standard English. Paper presented at the Annual meeting of the American Education Research Association, Chicago, April 1972.

Luria, A. Speech and the Regulation of Behavior. New York: Liveright, 1961.

Mehan, H. Assessing children's school performance. In Childhood and Socialization. H.P. Dreitzel (Ed.) New York: Macmillan, 1973, pp. 240-264.

Mercer, J.R. Labelling the Mentally Retarded. Berkeley, CA: University of California Press, 1973.

Mishler, E.G. Studies in dialogue and discourse: II: Types of discourse initiated by and sustained through questioning," Journal of Psycholinguistic Research, 1975, 4:99-121.

Peisach, E.C. Children's comprehension of teacher and peer speech. Child Development, 1965, 36, 467-480.

Phillips, S.U. Participant Structures and Communicative Competence: Warm Springs Children in Community and Classroom. In Functions of Language in the Classroom, C.B. Cazden, V.P. John, D. Hymes (Eds.) New York: New York: Teachers' College Press, Columbia University, 1972.

Piaget, J. The Language and Thought of the Child. London: Routledge and Kegan Paul, 1926.

Quay, L.C. Language dialect, reinforcement and the intelligence test performance of Negro children. Child Development, 1971, 42, 5-15.

Quay, L.C. Negro dialect and Binet performance in severely disadvantaged black four-year -olds. Child Development, 1972, 43, 245-250.

Quay, L.C. Language dialect, age and intelligence-test performance in disadvantaged black children. Child Development, 1974, 45, 463-468.

Roth, D.R. Intelligence testing as a social activity. In Language Use and School Performance. A.V. Cicourel, K.H. Jennings, S.H.M. Jennings, K. Leiter, R. MacKay, H. Mehan, D.R. Roth (Eds.) New York: Academic Press, 1976, pp. 143-218.

Stevenson, H.W., Williams, A.M., and Coleman, E. Interrelations among learning and performance tasks in disadvantaged children. Journal of Educational Psychology, 1971, 62, 179-184.

Stolz, W. and Tiffany, J. The production of "child-like" word associations by adults to unfamiliar adjectives. Journal of Verbal Learning and Verbal Behavior, 1972, 2, No. 1, 38-46.

Weener, P.D. Social dialect differences and the recall of verbal messages. Journal of Educational Psychology, 1969, 60, 194-199.

Werner, H. Comparative psychology of mental development. Chicago: Follett, 1948.

Williams, R.L. and Rivers, L.W. The use of standard versus non-standard English in the administration of group tests to black children. Paper presented at the Annual meeting of the American Psychological Association, Honolulu, September, 1972.

Wolfram, W. Levels of sociolinguistic bias in testing. Mimeograph dated 1974 to appear in Seminar in Black English, Erlbaum Publishers.