# The Need to Move Beyond Standardized Methods

[A]ny fundamentally new approach to a scientific problem inevitably leads to new methods of investigation and analysis.

Vygotsky, *Mind in Society*

How can we discover what a child knows? Consider two different methods. The traditional approach is to investigate the child's thinking by employing tests or standard tasks. The practitioner administers IQ tests, language tests, achievement tests, and the like in a standardized fashion. All children receive the same set of questions or test items in essentially the same way. Responses are not difficult to score, and test results can easily be computed and compared. Similarly, the researcher gives all subjects the same series of tasks, each presented in a uniform fashion. The resulting behaviors can be coded with adequate reliability, and results can be quantified and analyzed with apparent precision.

The method of standardized testing is extremely influential, dominating the assessment and study of children's thinking. Indeed, this method has virtually come to define what we consider to be "scientific." Use of the standardized method is the dominant procedure in research, as can be confirmed by even a cursory examination of journals like *Child Development* or *Developmental Psychology*, and it is the typical procedure employed in the assessments conducted by practitioners, for whom various psychological tests are indispensable tools.

Consider, however, a radically different approach to understanding what a child knows – an approach that I will argue makes more

sense than the traditional method. In a clinical interview, the examiner begins with some common questions but, in reaction to what the child says, modifies the original queries, asks follow-up questions, challenges the child's response, and asks how the child solved various problems and what was meant by a particular statement or response. The interviewer makes "clinical" judgments concerning the child's motivation and personality and uses these judgments to modify the questioning in appropriate ways, perhaps pressuring one child but treating another with kid gloves. The examiner is constantly engaged in interpreting the child's response so as to follow up on it in an incisive manner.

In employing the clinical interview, the examiner – practitioner or researcher – literally treats each child differently. Indeed, the clinical interview is *deliberately nonstandardized*, thus violating the central tenet of traditional testing. If the traditional method depends on uniformity, the clinical interview thrives on individual variation. In a collection of clinical interviews done for the purposes of either practice or research, it is possible that no two children are treated in an identical fashion and that no two children receive an identical collection of questions. The clinical interview is often considered to be preliminary, sloppy, imprecise, lacking in rigor, "unscientific" – in short, unsuitable for objective practice or rigorous research.

This book is an attempt to show that for certain key purposes the traditional methods of standardized testing are inadequate. Based on outmoded theory, standardized methods often fail to provide adequate insight into cognitive function; they are not effective techniques for understanding the processes of thought. By contrast, the clinical interview method offers a useful and seldom understood alternative to standardized testing. Although deliberately nonstandardized and difficult to employ, the clinical interview method can provide both researcher and practitioner with deep insights into children's thinking. Indeed, this nonstandardized, nontraditional, and, in some circles, controversial method is the "scientific" method of choice for the purpose of entering the child's mind.

Because the method of standardized administration has been so influential, it deserves a serious critique. This chapter therefore describes the method of standardized administration, explicates the as-

2

sumptions underlying it, and finally points out flaws which weaken the method's usefulness for examining conceptual frameworks and strategies. Although useful for some purposes, the method of standardized administration is not effective for others. It should not be allowed to provide the exclusive definition of what is "scientific" in research and practice.

## THE METHOD OF STANDARDIZED ADMINISTRATION

Consider first the basic procedures involved in the method of standardized administration, and second its underlying rationale.

### The Basics of the Method

Suppose that I want to gain insight into such aspects of the child's mind as changes in moral judgment from childhood to adolescence, a student's conceptual abilities in mathematics, the differences between boys and girls in logical reasoning, or the comprehension abilities of a child diagnosed as suffering from "learning disability." To answer questions like these, the traditional method is to employ a *task*,[1] or a series of tasks, which I will call a *test*, with or without norms. For example, if as a practitioner I am interested in examining processes of logical reasoning in an individual child, or if as a researcher I want to study age differences in logical reasoning in groups of 4- and 6-year-olds, the traditional procedure would be to

- develop a test involving a series of logical-reasoning problems or tasks ("Jane rides a bicycle. Bicycle riders are athletic. Is Jane athletic?");
- develop instructions which make clear what needs to be done ("I want you to tell me whether the answer to the question is yes or no");
- use procedures for establishing rapport and motivating the children ("We're going to have a lot of fun playing some games today. Listen carefully to the stories I tell you.");
- present the problems in as attractive a way as possible (perhaps use a colorful picture of Jane on a bicycle as part of the procedure);

3

- make sure that the physical surroundings are comfortable and conducive to diligent work;
- use a checking procedure to determine whether the subject pays attention and seems to understand the basic information in the problems ("What does Jane do?");
- administer the entire procedure in a standardized fashion, perhaps with several such problems or tasks given in the same order to all subjects;
- limit the children to responses that can easily be scored in an objective fashion, such as simple "yes" or "no" responses (so that independent judges would easily agree on the scoring);
- if the responses must be longer (as in the response to a moral-judgment dilemma), develop clear criteria for coding the responses; and
- carefully record the child's behavior, responses, and overall performance.

This example of logical thinking illustrates several basic features of traditional methods. First, it involves a *task* – a situation contrived by the psychologist to elicit behavior (activities, verbalizations, etc.) that will provide information concerning the "dependent variable" of interest, in this case, the child's logical thinking. In general, the task is of limited scope, focused on a particular topic, like the problem concerning Jane and the bicycle. The task elicits a fairly limited response, like the yes or no answer to the specific question concerning Jane's athletic prowess. The task is not modified regardless of what happens in the study; once decided upon, the task does not change.

Second, the task or series of tasks (the test) is administered in a uniform fashion to all subjects. This kind of *control* and *standardization*[2] is the essence of the traditional methodology. Both researcher and practitioner attempt to hold constant the test "stimuli," to make sure that all subjects receive the same problems in the same way. The controlled conditions must involve precise and uniform directions for administering the test, including use of the same questions, materials, tasks, instructions, and even the same tone of voice and pace in talking to different subjects. The testing environment should be the same for all subjects, including "lighting, ventilation, seating facilities, and

working space" (Anastasi, 1988, p. 34), although this is difficult to achieve in practice (as all of us who have had to test children in the school nurse's office know full well).

Third, standardized administration often contains devices for promoting rapport, attention, and motivation. Obviously, both researcher and practitioner want children to take the test seriously, cooperate with the examiner, work hard, concentrate, and in general do as well as possible. As a result, we normally stress having the children attend to the test, we minimize distractions, and we encourage the children to work hard. Sometimes we build into the test techniques for checking our assumptions about rapport, attention, and motivation.

Fourth, tasks and tests may or may not be based on norms. I may develop the logical-thinking problems myself, perhaps relying only on informal pilot work. Or I may use already existing tasks about which various amounts of information may be available. If I borrow the problems from an existing research study, I will at least know how that researcher's subjects performed on the test (and other subjects in similar studies). If I use a test with extensive norms concerning age, ethnicity, social class, etc., then I know a good deal about how children "typically" respond to the test. In research, we tend to use tests which do not involve extensive norms; in assessment practice, we tend to use normed tests. But whether the tests are original or borrowed, normed or not normed, their essence is standardized administration as described above.

### Two Rationales for Standardized Administration

Standardized administration can be justified on both scientific and ethical grounds.

#### SCIENCE

The basic scientific justification for standardized administration originated in the 19th century, when experimental psychologists, concerned mainly with sensation and psychophysics, required rigorous control over the conditions of testing. Suppose that you as an experimental psychologist want to investigate the ability to discriminate

different colors. You want to know, for example, whether people can see the difference between red and green, or between one reddish color and another not so red. To find out, you need to make sure that all subjects are presented with the same two colors – in the first case, the same shades of red and green, or in the second, the same "reddish" color and the same "not so red" color. (It would help tremendously if you could define these colors in precise, physical terms, rather than just describing them as "reddish" and "not so red.") You would not obtain useful information were some subjects to see two very similar colors and other subjects two extremely different ones.

Similarly, you need to make sure that the lighting conditions are the same for all subjects. It would be a mistake for you to arrange conditions so that illumination was "bright" for some subjects and "dim" for others. (Here again, precise, physical measurement is helpful: What are "bright" and "dim"?) You also need to use the same instructions for all subjects. Thus it would be a mistake if some subjects were told to indicate when they "thought" one color was different from the other, whereas the other subjects were required to say when they were "sure" that the colors differed.

This then was the model of scientific procedure as developed by 19th-century experimental psychology. And it makes a good deal of sense: if anything is to be learned from investigations like these, the experimenter must have control over the stimuli and must be able to ensure that procedures are fully standardized.

In an effort to provide "scientific" measurement, this logic of control and standardization was then applied to the psychological testing developed in the late 19th and early 20th centuries. The argument was as follows: The goal of testing is to measure variation in some trait across individuals. We want to know whether A is more intelligent, aggressive, or depressed than B. We may want this information in order to compare groups of individuals (are boys more aggressive than girls?), to compare traits within individuals (is intelligence related to lower amounts of depression?), or to compare the effects of independent variables (does increasing stress heighten aggression?). In all of these cases – the comparative study of groups, the study of individual differences, or experimental research – the aim is to get reliable and accurate measures of psychological traits.

How to do this? The basic rationale is again control and standardization. According to Anastasi (1988), who has brought a good deal of sound judgment to bear on the subject of psychological testing: "If the scores obtained by different persons are to be comparable, testing conditions must obviously be the same for all" (p. 25). All subjects must receive the same test stimuli – whether this be a list of nonsense syllables or a paragraph to read or an IQ test – in the same manner. To do otherwise, the argument goes, would be to render the data uninterpretable. If subjects received different test stimuli, then the tester could not confidently attribute subsequent variations in performance to individual differences in a particular trait.[3] Consequently, strict uniformity of administration is required. "Such a requirement is only a special application of the need for controlled conditions in all scientific observations" (Anastasi, 1988, p. 25). This scientific argument then leads to the whole apparatus of standardized administration.

ETHICS

Standardized administration can also be justified on ethical grounds, particularly in terms of a specific kind of fairness. As far back as 1845, Horace Mann offered several reasons for introducing standardized testing to the schools (Wainer, 1992):

> They are impartial.
> They are just to the pupils.
> They prevent the officious interference of the teacher.
> They take away all possibility of favoritism.
> They make the information obtained available to all.
> They enable all to appraise the ease or difficulty of the questions.
> (p. 15)

In one way or another, these reasons revolve around *fairness* or impartiality. One justification is that standard tests prevent the teacher from favoring some children over others (perhaps by giving some children easier questions than others) or from interfering with the process of testing. Another justification is that the tests make the process public, so that an outside observer can judge whether the

questions are too hard or too easy. In this way, testing can be fair and "just."

The originators of psychological testing were also vitally concerned with fairness. In the France of the early 1900s, "alienists" (assessment specialists) used subjective methods to perform diagnoses of mental retardation in schoolchildren. Those children identified as retarded were pulled out of regular classes and assigned to special instruction. Although the alienists' goal was laudable – to identify children "unable to profit . . . from the instruction given in ordinary schools" (Binet & Simon, 1916, p. 9) – the accuracy of their diagnoses was by no means guaranteed, with the result that children were often mislabeled and then denied a mainstream education. Binet and Simon (1916) were particularly critical of the unstandardized nature of the alienists' testing, which seemed to involve

> haphazard decisions according to impressions which are subjective, and consequently uncontrolled. Such impressions . . . have at all times too much the nature of the arbitrary, of caprice, of indifference. Such a condition is quite unfortunate because the interests of the child demand a more careful method. To be a member of a special class can never be a mark of distinction, and such as do not merit it should be spared the record. . . . [T]he precision and exactness of science should be introduced into our practice whenever possible. (pp. 9–10)[4]

Binet and Simon felt that fairness demanded standardized, uniform administration. Their goal in developing the standardized intelligence test was to be *fair* to the child, whose "best interests" demand a precise and exact diagnosis. Haphazard, subjective assessment can result in the child's being labeled as retarded, which is never a "mark of distinction" and which can cause the child to miss out on normal instruction and be consigned to the special classroom. It is a great irony that Binet and Simon's motivation underlying the creation of the intelligence test, which is today so severely criticized for its bias and lack of fairness, was to avoid inaccurate and unreliable diagnoses of schoolchildren.

Similarly, modern testers are also concerned with the issue of fairness, particularly in the case of minority students. Anastasi (1988) argues that when "prejudice may distort interpersonal evaluations,

tests provide a safeguard against favoritism and arbitrary or capricious decisions" (p. 67). She maintains that tests provide an opportunity for bright students of minority background to reveal their abilities in a forum not subject to the bias of teachers. The tests are a level playing ground on which all may compete equally. She argues that if standardized tests were eliminated, we would have to fall back on subjective judgment, which is subject to "unreliability, subjectivity, and potential bias" (p. 68), much as Binet and Simon had earlier maintained. In her typically wise summation, Anastasi claims that "tests can indeed be misused in testing cultural minorities – as in testing anyone else. When properly used, however, they serve an important function in preventing irrelevant and unfair discrimination" (p. 68).

I have tried to show that standardized testing springs from the noble motive of ensuring fairness. As articulated by Mann, by Binet and Simon, and by Anastasi, fairness has the goal of eliminating bias, discrimination against minorities, stereotyping, haphazard judgments, "officious" (or other) interference, and distortions of judgment.

The method for achieving this kind of fairness is to treat all people alike, to give everyone the same conditions for running the race. From this point of view, no one should be given easier questions than anyone else; no person should be judged on a different basis from another. A wrong answer should be a wrong answer for all. The same solution should not be judged incorrect if the child is African-American and correct if she is White. This kind of fairness is *colorblind*: It treats all children in the same way and ignores irrelevant attributes.[5]

The arguments in favor of the method of standardized administration are both scientific and ethical. The method aims at accurate measurement, unconfounded by variation in test administration, and it aims at fairness in the sense of impartiality.

## Some Flaws in the Method

I claim that for certain purposes, the method of standardized administration (which I will sometimes designate as "traditional") often

falls short on both ethical and scientific grounds. There is a sense in which the method is often distinctly unfair and can provide us with only limited information concerning children's thinking. The method of standardized administration suffers from several basic flaws, each of which will be described in some detail below:

- Despite standardized administration, subjects nevertheless interpret tests in idiosyncratic ways, in ways not intended by the examiner.
- It is often not clear what cognitive activities are actually measured by standardized tests.
- Standardized achievement tests are usually based on outmoded assumptions about cognition.
- Standardized methods are not suitable instruments for studying complex thinking.
- Standardized methods are not suitable instruments for studying dynamic change.
- Standardized procedures cannot effectively motivate all children.
- Traditional methods are often not suitable for tapping the competence of those who are in some way "different" by virtue of culture, ethnicity, health status (e.g., the hearing impaired), or other reasons.
- Traditional methods are inadequate for the kind of exploration often required in research and practice.

## CHILDREN MAY NOT INTERPRET THE SAME TESTS IN THE SAME WAYS

As we have seen, traditional psychological research uses the method of standardized administration for a critical reason: to hold test stimuli constant; to ensure that all subjects receive the "same" problems. Although this method may work for the psychophysical experiments in which it originated, it often fails when other situations, including cognitive problems, are involved.

It fails because different people interpret the same *objective* stimuli in different *personal* ways. Given the "same" stimulus, we see different things in it. In Piagetian terms, we always *assimilate* objective reality into our own schemas. We interpret the world, and we respond more to the subjective interpretation than we do to the objective reality. As Shweder (1991) puts it: "'stimuli' are not external to or

independent of our understanding of them" (p. 79). So a stimulus is not a stimulus is not a stimulus. It is what we make of it.

The effect is most apparent in cross-cultural research. Suppose we give Americans and unschooled Africans the same logical-thinking test of the type described earlier (Jane's bicycle). Let's even suppose that both groups are equally familiar with bicycles. It turns out that Americans and Africans answer such questions quite differently. The Americans give answers of the type we expect, but the Africans do not (Cole & Scribner, 1974). Africans say essentially that they cannot answer because they do not know Jane. Does this mean that the Africans cannot reason logically? Not necessarily; we simply can't tell from the results. The problem is that the Africans may have interpreted the question in an unexpected way. For whatever reasons, they convert the question into one about a particular person, whom they don't know, so that the problem cannot be solved. But when the same people deal with questions involving people they do know, they are able to reason logically in the way demanded by the test.

The view that individuals "construct" their realities has important implications for testing as well. It means that different children may interpret in idiosyncratic ways the very same task, the very same instructions. For example, once I asked a child to count some toys. I said, "Count the toys out loud." She proceeded to say, "Ball, block . . ." and to name all of the objects in front of her. But when I said instead, "How many toys are there altogether?" she had no difficulty in counting.[6] If I had limited myself to standardized instructions ("Count the toys out loud"), I would have learned only that she did not understand those particular words; it would have been wrong for me to claim that she could not in fact count.

We do not know how often this sort of misinterpretation of standardized instructions (from the examiner's point of view, not the child's) occurs. Obviously, the less clear the instructions, the more likely is misinterpretation to take place. (It would be interesting to conduct some research on the extent to which children develop idiosyncratic interpretations of problems and instructions on popular tests in the areas of "intelligence," cognitive abilities, and achievement.) But we do know that misinterpretation does occur with some frequency and that researchers and practitioners cannot ignore it.

Indeed, a useful approach to interpreting a "wrong" answer is to discover the question to which the child's answer is correct. In other words, the child's wrong answer to *your* question may be a correct answer to a different question – the one that the *child* was posing.

The danger that children may not interpret tests as we would like is particularly acute when we are attempting to compare children of different cultures, age-groups, social classes, and the like.[7] *Instructions and questions* may be misunderstood or understood differently by children varying in age. The "typical" 8-year-old may understand the request "Tell me how these pictures are alike" differently from the typical 5-year-old. *Procedures* may also vary in familiarity for members of various groups. Suppose, for example, the investigator asks middle- and lower-class children to write down their dreams from the preceding night. The middle-class children, who typically attend better schools and are more skilled at writing, are likely to produce more numerous and vivid (written) dreams than their lower-class peers. But do the poor children really dream less often or less vividly? Probably the groups differ mainly in their facility with the procedure employed in the research. As Cole and Means (1981) maintain:

> Simply following the same procedure or using the same instructions and materials with two groups does not automatically constitute equivalent treatment. . . . In general, there is reason to expect that most standard experimental procedures are neither equally motivating for, nor equally well understood by [all groups of subjects]. . . . Every aspect of an experimenter's method – the materials, instructions, and procedure – can be a source of treatment inequivalence for different comparison groups. (pp. 43–44)

In brief, standardized administration does not automatically result in presenting children with the "same" tasks and tests. Objective equivalence may differ from subjective equivalence, which should be our main concern. We must always be alert to the possibility that different children interpret the same task in different ways.

At the same time, we should recognize that for most practical purposes, standardized administration sometimes comes close enough to establishing subjective equivalence among some children. Thus, if we are testing two children or groups of children from similar,

mainstream backgrounds, we need not be overly concerned about possible widespread idiosyncratic interpretations of test materials. Most middle-class American children probably would interpret the logical-thinking problem of Jane's bicycle in similar ways.

But without further checks, we cannot be sure. And we should be even less sanguine concerning our efforts to "hold the stimuli constant" if the children are from different age-groups, and especially if they are from different cultures, social classes, and the like – that is, if they are *different* from the mainstream.

What can we do to make sure that tests are interpreted in the same way by all who take them? As we shall see, the only solution is to create *objectively different* tests that have *equivalent subjective* meaning for the people in question. Standardized administration won't work. We have to find alternatives.

## TESTS MAY NOT MEASURE WHAT THEY ARE INTENDED TO MEASURE

We use tests to tell us something about underlying thinking. We are not interested in subjects' responses to the test items themselves but in what subjects' behavior indicates about their psychological processes, like logical thinking. But how do we know whether the test measures what we think it does? The usual procedure in developing a test of this type is straightforward. It begins with a task analysis. We invent or borrow items that seem clearly to require a certain cognitive process for their solution. If we want to measure logical thinking, we may create problems that seem to involve syllogisms. We are careful to construct the items in such a way that other methods of solution will not be effective; only use of the syllogism will work. If we want to measure multiplication ability, we create multiplication tasks. We assume that if subjects get the right answer they must have used the process that the task was designed to measure. Then we administer the tasks to the subjects under investigation and make inferences from performance on the test (and related tests) to underlying cognitive process.

But matters are not as simple as they may initially seem: We should not be so quick to assume that tests measure what we think they measure. "Perhaps the greatest pitfall in all experimental work about

thinking processes arises from a confusion between what we think the task entails and what subjects who enter these tasks actually deal with in their efforts to comply with the demands placed on them" (Cole & Means, 1981, pp. 136–137).

Suppose that you want to measure children's knowledge of number facts. You simply want to know whether children can remember the elementary addition combinations, like $3 + 4 = 7$. What could be simpler? You just pick a few problems of this type, administer them with an instruction such as "I'm going to give you a number fact problem like $2 + 2$. What is $2 + 2$? That's right, it's 4. Now I'll give you some more problems, and I'd like you to tell me the answers as quickly as you can." We then observe a certain level of performance on items such as these and conclude that the child has a "good memory for number facts" or a poor one.

Often such inferences may be correct. But unfortunately they can be dead wrong too. Here is an example. When asked, "How much is $7 - 4$?" a first-grade student, Becky, said that the answer is 2. We might assume from the nature of the test that she had simply misremembered the number fact that her teacher tried to convey. If she were given a series of such items on a test, we would add up her correct answers and, if there were few of them (Becky actually got only one of three correct), conclude that she has poor memory for number facts. But suppose we go beyond the standard test procedure. When asked, "How did you get that answer?" Becky replied: "I knew that 7 take away 4 is 2 because I know that 4 plus 2 is 7. And if 4 plus 2 is 7, then 7 take away 2 must be 4."

This example shows how things were more complicated than the test seemed to indicate. First, Becky began with a mistaken idea, namely that "4 plus 2 is 7." This was her basic premise; probably it was a faulty memorization of a number fact, and this was indeed what the test was intended to measure. But there was more to her cognitive process than rote memory. Indeed this second ingredient in the cognitive stew was much more interesting and significant than the first, the faulty memory. She introduced the correct idea that if $4 + 2 = 7$, then it must be true that $7 - 4 = 2$. Her underlying and unspoken assumption was that subtraction nullifies addition: What addition produces, subtraction can undo. She then combined these two ideas

by means of a classic syllogism: If it is true that $4 + 2 = 7$, then it must be true that $7 - 4 = 2$.

So the tester's initial inference that the incorrect answer was a result of faulty memory was correct only in an accidental and trivial way. It is true that Becky misremembered a number fact. But the faulty memory involved $4 + 2$, not $7 - 4$. And the faulty memory was trivial when compared with her creative use of the syllogism.

We see then that it is dangerous to make inferences from test performance to underlying cognitive process, especially when all we usually observe is a right or wrong answer to a test item. In general, test items are deliberately designed to elicit very little behavior from subjects. In particular, they usually do not elicit subjects' verbalizations concerning strategies and methods of solution – verbalizations which might provide considerable insight into children's thinking. No doubt the standard impoverishment of subjects' behavior simplifies scoring and thereby helps ensure that raters can agree on how to score the items (high "inter-rater reliability"). But there may be a serious price to pay for this apparent "objectivity": incorrect inferences concerning what most interests us, children's thinking. Tests allow us to be objective and wrong – that is, to come up with incorrect conclusions – at the same time!

## STANDARDIZED ACHIEVEMENT TESTS ARE BASED ON OUTMODED ASSUMPTIONS ABOUT COGNITION

Test theorists themselves are beginning to advance similar criticisms of standardized tests, particularly achievement tests. For example, some argue that standardized tests are based on outmoded psychological theory (Mislevy, Yamamoto, & Anacker, 1992). In a sense, tests use a rather elaborate and sophisticated form of 20th-century technology (think of how many books have been written about the details of testing and testing statistics!) to implement archaic 19th-century psychological theory. In this view, most standardized achievement tests are based on an incorrect psychological theory which views learning as the passive accumulation of facts provided by instruction. The job of the test then is to count up the number of such facts or skills, which can be done by scoring the subject's answers as "right" or "wrong." No doubt some learning is rote and

mechanical. But this is not the only phenomenon achievement tests should focus on, even though it is often encouraged by our educational system. Traditional achievement tests ignore students' construction of psychological knowledge, their thinking, and their underlying conceptions. As anyone who has navigated the educational system knows, it is quite possible for students to perform well on standard achievement tests, even through the college level, and yet understand very little and indeed continue to hold basic misconceptions concerning the subject matter (see, e.g., Clement, 1982, who cites the example of physics).

Others argue that standardized achievement tests do not provide information useful for teaching. The tests are "more likely to confirm what the teachers already know about the student than provide them with new insights or clear indications of how best to help the student. The global score simply does not reveal anything about the causes of the [student's] problem or provide any direct indications of what instructional strategies would be most effective" (Linn, 1986, p. 72).

Dissatisfaction with standardized testing is even more rampant among those who use tests than it is among those who create them. There is widespread consensus in education, particularly in mathematics education, that standardized testing provides little useful information concerning students' knowledge and learning and in fact has created more problems than it has solved. For example, it distorts the curriculum because teachers often teach to the rather trivial demands of the tests (Madaus, West, Harmon, Lomax, & Viator, 1992).[8]

### STANDARDIZED METHODS ARE NOT SUITABLE FOR STUDYING COMPLEX THINKING

There has been a "cognitive revolution" in psychology and other fields (Gardner, 1985), and as the 21st century approaches, psychologists are more vigorously and variously than ever engaged in cognitive research and practice. Researchers study topics as diverse as children's understanding of distinctions between body and mind (Inagaki & Hatano, 1993), their memories of abuse (Ceci & Bruck, 1993), their understanding of self (Damon & Hart, 1992), their "internal working models" of others (Greenberg & Mitchell, 1983), their racial

stereotypes (Bigler & Liben, 1993), and their academic knowledge (Ginsburg, 1989a).

Practitioners are also vitally concerned with understanding children's minds. The diagnostician evaluates the child's cognitive abilities. The clinical psychologist attempts to learn how the child conceptualizes the self and conceives of others, particularly parents, siblings, and peers. The school psychologist investigates the cognitive processes underlying the child's difficulties with school learning. The pediatrician inquires into the child's understanding of his or her illness and of the steps proposed to ameliorate it.

Of course, our common concern with entering the child's mind does not guarantee that we will agree on what we find there. The cognitive revolution has hardly resulted in unanimity! As researchers and practitioners, psychologists are as contentious a lot as they have always been, employing theories almost as diverse as the phenomena they investigate. Psychologists have produced a proliferation of theories and points of view. They may be information processors or Piagetians or Self theorists or followers of Vygotsky or cognitive scientists. Psychologists may even hold to one of several brands of constructivism: simple, social, or radical. Nevertheless, regardless of orientation, researcher and practitioner alike agree that a major aim of psychology is the understanding of children's minds.

Methods are based on theories and need to keep up with them (T. Kuhn, 1962). The question then is whether the traditional method of standardized testing is adequate to the new theoretical task of understanding the complexity of children's thinking and its development. I think the answer is no, for several reasons.

First, the tests simply do not allow for the child to employ much rich thought. Interesting, complex thinking generally does not take place in a short period of time in response to a narrowly focused test item. Thinking is more like a stream – shifting its course and ever changing – than it is like a static pebble in the stream. Tests are more appropriate for picking up pebbles than for capturing the flow of the stream.

Indeed, children learn *not* to think – or not to think deeply – in response to test items. Too much thought might lead you to believe

that none of the four multiple-choice answers is really correct, that the question was not very good in the first place, and that another issue is really involved. And if you try to work all this out, you will not have enough time to get the (wrong) right answer that will get you credit on the test. I have heard that in training sessions designed to promote successful test taking, children are taught not to spend too much time on individual items because doing so will hurt their overall test score. In other words, they learn not to think too much about something that might be challenging.

Second, standard methods are not suitable for depicting complex cognitive *systems*. Researchers and practitioners are interested in gaining insight into such matters as children's "conceptual frameworks" – their relatively complex theories or belief systems concerning the physical, social, and psychological worlds. For example, researchers have examined how children's conceptualizations of teachers and schooling affect academic motivation, and ultimately achievement (Weinstein, 1993). This research arises from a perspective which sees children's motivation as stemming not simply from teachers' instructions, behavior, constraints, and rewards and punishments but from children's *interpretations* of the social context of schooling. More important than the "objective reality" of teachers and schools is the child's "construction" of that reality, the ways in which the child interprets and comprehends the educational experience.

Weinstein finds that children are well aware of the different ways in which teachers treat children seen as high or low achievers. For example, "It's a way that [teachers] talk to you . . . about your grades. . . . A very soft voice lets you know you are doing well and a hard tone of voice – they shout and scream to let you know that you are not doing well" (1993, p. 206). Children are also sensitive to the relations among children considered "smart" and "not so smart." "Not so smart girls can't play with the smart, smart girls because smart girls just act like they ignore them." Children also have ideas about what it means to be not so smart. "When the kids [who] don't pay attention, when they see they have a sad face [a mark of failure given by the teacher] on their paper, they say 'I'm not smart.' . . . They be mad, then they start kicking the college kids" (p. 207).

Understanding children's conceptual systems concerning teachers

and schooling is vital for both researchers and practitioners. As a researcher, I cannot explain children's achievement without understanding what schooling means to them, and as a practitioner, I may not be able to help the individual child who is failing unless I understand how he or she interprets the teacher's behavior and what happens in school. Yet simple standardized tests are inadequate to capture the complexity of cognitive systems like these.

Third, we know that cognitive competence is often "situated" in various environmental contexts. In the natural environment, problems are complex and people deal with them over time, with various tools, with the help of other people, and with many different kinds of motivation. Thus, children may engage in effective mathematical thinking in connection with selling candy on the streets but may not achieve similar levels of success in their school mathematics (Carraher, Carraher, & Schliemann, 1985). Standardized procedures, however, do not tap rich environmental contexts. Tests are designed to be narrow, asocial, and myopic. The usual testing situation is so devised that subjects must deal with a narrow problem over a short period of time and are allowed to respond only in limited ways. Tests demand impoverished responses because to permit richness would be to make scoring difficult.[9] It is harder to see complex phenomena than simpler ones, so test constructors choose to look at what is easier to see. These constraints of standardized tests limit what subjects can do and what can be learned from what they do.

Because of this decontextualization, tests are often not effective in tapping children's intellectual competence. McDermott describes the everyday problem-solving activities of Adam, a boy considered learning disabled. His behavior was quite variable. In everyday situations, he acted quite competently. In the cooking club, when allowed to work with his friend, he was able to bake cakes. But in testing situations, "Adam stood out from his peers not just by his dismal performance but by the wild guesswork he tried to do" (McDermott, 1993, p. 279). Adam's guessing on tests may be important to know about. For one thing, he may engage in similar behavior in school. But it would be a mistake to think that the tests tell us the whole story about Adam's mind.

## STANDARDIZED METHODS ARE NOT SUITABLE INSTRUMENTS FOR STUDYING DYNAMIC CHANGE

Tests do not allow for the study of the *development* of thinking in the individual. Often, a child's conceptual frameworks and strategies seem to shift even within a short period of time. In the course of an interaction with the psychologist, the child's concept may seem to change, several different strategies may be used, or different aspects of the child's world view may be emphasized, elaborated, or obfuscated. The child believes that the teacher treats "smart" kids differently from others, but in the next breath points out how the teacher behaves in the same way toward all. Or on the first subtraction problem the child incorrectly uses the "subtract the smaller number from the larger" strategy, or "bug," and concludes that $12 - 4 = 12$ because the 2 must be subtracted from the 4 (J. S. Brown & VanLehn, 1982; Ginsburg, 1989a), but on the next problem a different bug is used, or perhaps the strategy used is correct.

Some of the changes may reflect instability in the child's emerging mind: Conceptual frameworks and strategies are not yet firmly established; the structures are not solid. Other changes may reflect learning: The experience of thinking about a topic or using strategies may in itself lead to learning and development. But whether we posit instability or learning or something else, one thing is clear: Conceptual frameworks and strategies are not conveniently static. It is rare that we can say with confidence that at a given point in time the child "has" exactly Concept X or Strategy Y and no other. "[I]t may be more accurate to view change, rather than static states, as the norm" (Siegler & Crowley, 1991, p. 614).

Perhaps tests and tasks can sample relatively static forms of knowledge at discrete times. But they are poor tools to employ in examining the processes of change. Indeed, tests tend to depict change as an undesirable lack of reliability which must be eliminated. Partly because of such limitations of method, developmental psychology has unfortunately devoted insufficient attention to the study of what should be central to the discipline: the study of development itself.

## STANDARDIZED PROCEDURES CANNOT EFFECTIVELY MOTIVATE ALL CHILDREN

Traditional test theory stresses the necessity for establishing rapport and motivation in order to ensure adequate conditions for testing. Some tests suggest specific statements the examiner can make in an effort to establish rapport. More often, before administering a test, psychologists engage in informal conversation with the child, asking about brothers and sisters or favorite games and activities and the like, all in an effort to put the child at ease, to relax the child so that full attention can be given to the test and serious work on it can begin.

In many cases this works. If the child is a good student, a mainstream middle-class child who is confident in school, comfortable with professional adults, even familiar with testing, then he or she is likely not to be too upset and to work reasonably hard at taking the test. Some children even *like* to take tests, seeing them as an opportunity for demonstrating their prowess. Given a standard test, one child, 4 years and 9 months, repeatedly exclaimed with great enthusiasm, "I like these puzzles!"[10] Other children will cooperate with the tester because they like to get out of class. How bad could the "games" be compared with what is happening in class?

Many children, however, do not enter the testing situation with this constellation of motives and attitudes. Some children are frightened of adults; some children see the researcher's games as evaluations or do not know what to make of them; some children think that taking the test will make them look stupid; some children think that their parents send them to a psychologist to punish them or that the psychologist will tell parents their secrets. These children are likely to have negative attitudes toward testing and are not likely to work hard at taking them. In response to a tester's statement that he would enjoy the "game," one child, 3 years 6 months, responded, "This is no game. I know so many games that's better."[11] And as the years go on, children who do poorly in school gradually seem to lose their motivation for taking tests (Paris, Lawton, Turner, & Roth, 1991).

All this is obvious to anyone who has ever worked with children. But the key point is that standardized administration – treating all children the same way – cannot succeed in establishing rapport with

and motivating *all* children. Chitchat about brothers and sisters may work with some children, but it is not enough to reduce the fears of a child who is consistently ridiculed for being "dumb." Describing the tasks as games may convince some children not to worry, but it cannot motivate the child who feels that whatever happens in school is not to be taken seriously. To establish rapport with many children and motivate them, you need to treat them as individuals, which is exactly the opposite of what standardized administration involves. Treating all children the same cannot lead to knowing many children as individuals; and if you do not know them as individuals, you cannot establish "rapport" and create motivation.

### STANDARDIZED METHODS ARE OFTEN NOT SUITABLE FOR TAPPING THE COMPETENCE OF THOSE WHO ARE "DIFFERENT"

Difficulties in standardized administration are particularly evident when it comes to the fair and accurate testing of minority groups. Suppose we wish to understand minority children's intellectual abilities. If we use standardized tests to get information of this type, we need to ensure that minority children are properly motivated to take the tests. As we have seen, many middle-class children seem to have little difficulty with test-taking attitudes or motivation. But the same may not be true of many lower-class and minority children. For example, lower-class African-American children may consider doing well in school, which presumably includes taking tests seriously, to be a sign of being "White" and hence to be undesirable (Fordham & Ogbu, 1986). In some cultures, it is not considered polite for children to offer opinions or even talk with adults. This must certainly interfere with test taking.

If children do not work hard at taking the test, the results are uninterpretable.[12] A low score on an ability test, for example, might indicate low ability, as it is intended to and is usually interpreted to mean, or it might merely indicate low motivation to perform well on the test and therefore permit no inferences about ability. Of course, the test score will accurately predict that the child is not doing well in school. But we already knew that. The issue is not school performance but underlying ability.

22

Given all this, we must try to identify and reduce factors that might interfere with minority children's display of their competencies on standardized tests. Or conversely, it is incumbent on us to make sure that minority children are well motivated to do their best on tests. What can be done? According to Anastasi (1988), some special precautions can be taken: "A desirable procedure is to provide adequate test-taking orientation and preliminary practice" (p. 64). But can this really be accomplished in a brief period of time with methods of standardized administration? Can our little games overcome deep-seated cultural differences and suspicions?

Anastasi (1988) also recommends avoiding use of materials that are "unfamiliar in a particular cultural milieu" (p. 307) or that may alienate minority children by referring repeatedly to suburban, middle-class family scenes, by picturing only affluent White children, or by stressing artifacts and customs of middle-class culture. (Similar arguments have of course been made about gender bias.)

This logic sometimes leads to the elimination of test items that reflect differences among cultures. The strategy is to keep only those items common to all, yielding a "culture-fair" test.[13] But I believe the strategy is likely to be counterproductive: These "culture-fair" tests are bound to fail because they are really "culturally deprived" tests. Culture-fair tests try to *remove* what is important to individual cultures. They eliminate culturally specific knowledge and practice – precisely the material that might stimulate motivation and allow members of the culture to demonstrate their competence.

The well-intentioned attempt to eliminate bias in testing may result only in empty, lifeless tests that cannot succeed at promoting motivation and otherwise providing insightful information about children's thinking. The solution, as we shall see, is just the reverse of what has been proposed: Play to the culture's strengths, do not try to eliminate them.[14] In brief, if we wish to motivate those who are different, standardized administration is not an effective method. And eliminating culture from tests may only reduce their power to interest and motivate.

23

## TRADITIONAL METHODS ARE INADEQUATE FOR EXPLORATION

One of the most important activities for a researcher or practitioner is to explore, to become familiar with the lives of those who will be studied. Before doing more "rigorous" investigation or assessment, the psychologist needs to get a feeling for the child, to develop an informal understanding of the topic under investigation, to develop some "common sense" about the issues.

How do you explore? You observe children, you talk with them, you work with them as teacher or friend, you spend time with your younger relatives, you engage in informal conversations. Basing your knowledge of children solely on what you read in textbooks or on what you learn from test scores leads to a severe case of intellectual deprivation. Both researchers and practitioners need to draw upon and be constrained by informal knowledge – common sense – about the children they are studying.

Suppose we want to understand the language competence of African-American adolescents. Suppose we know nothing about the topic except for the fact that research shows that these youths perform extremely poorly on standardized tests of language administered in the schools. This limited set of data almost inevitably leads to the conclusion that African-Americans' language is impoverished.

Suppose, however, that we have informal contact with these youths, observing them and talking with them in nonschool situations. This shows immediately that their language, although different in some respects from the mainstream, is far from impoverished (Labov, 1970). Given this informal knowledge, this "common sense" resulting from informal observation and exploration, we would not be likely to fall into the trap of misinterpreting the data (and perhaps we would not even bother to collect standard test data in the first place). The tests may tell us something about performance in the school setting, but they do not provide insight into African-American adolescents' basic linguistic competence. To gain deeper insight into this competence, we must employ techniques other than standardized tests and we must conduct research outside the school setting, as indeed Labov then proceeded to do, producing important results.

So if we lack common sense and familiarity with the local culture, we may be led to faulty conclusions. If we do not interpret the data provided by standard research through the lens of rich informal experience, we may reach conclusions which are in serious error, as was true of many researchers investigating poor children's intellect (as cited in Ginsburg, 1986b).

You cannot do good research or clinical practice unless you use your head. First, you have to explore, and indeed, sometimes you learn more from exploration than from formal, "rigorous" procedures used later.[15] But standardized tests cannot be used effectively for exploration. This should come as no surprise: They are not *designed* for purposes of exploration. Restricted and focused as they are *intended* to be, they cannot give you the lay of the land; they cannot provide you with intuitive knowledge of your subject matter. The point is not to criticize standardized tests for not being able to do something they were not intended to do but to emphasize that if you want to explore, as you *must*, you need other methods. Don't get locked into the narrow, myopic world of standardized tests.

### Conclusions

There are some things that standardized tests can do reasonably well. Often, instructions are reasonably clear to many children, who arrive at more or less common interpretations and are adequately motivated to work hard at the test. If so, a test may tell us something about a child's cognitive ability: for example, a child is skilled in comprehension but weak in productive vocabulary. A cleverly designed test, interpreted with sophistication, may even provide information about children's cognitive processes. From such a test, we may learn that a child seems to have some competence in logical thinking or seems to be in Stage 3 of moral judgment. Unless tests gave us *some* useful information, they probably would have been replaced long ago.

At the same time, we should not ignore serious shortcomings in the method of standardized administration. Different groups of children may interpret the "same" stimuli in different ways, so that the logic of "holding stimuli constant" is inevitably compromised. The tests may

not measure what we think they measure, and achievement tests often attempt to measure the wrong sorts of things. Tests are not useful for studying complex thinking or development. Standard administration is ineffective in motivating many children, particularly those who are in some way different. Indeed, it is a contradiction in terms to believe that standardized administration can motivate all children. And tests are particularly inappropriate for understanding children who are "different" in terms of culture or in other ways.

For all these reasons, the method of standardized administration reveals only the tip of the iceberg with respect to children's thinking. Testing modeled on psychophysics is a poor vehicle for entering the child's mind; standardized methods are not well suited for obtaining a rich and sensitive view of cognitive processes.

## NEW DEVELOPMENTS IN METHOD

I have argued that psychologists have been slow to abandon or at least deemphasize methods rendered obsolete by the new theoretical developments associated with the cognitive revolution. Yet progress can be noted in several important attempts to develop techniques appropriate for the study of mind.

Some of these are modifications of existing methods. In the bastions of psychometrics, attempts are being made to make tests sensitive to the needs of individuals. In "tailored testing," the difficulty of test items is adjusted to the needs or previous responses of the individual (Green, 1983). Thus, in the case of computerized testing of attitudes or personality traits, a statistical algorithm determines a subject's basic tendencies and then uses this information as a decision rule for omitting items which are likely to be irrelevant or redundant. If a subject believes "Abortion is a terrible sin," there is likely to be no point in asking him or her whether "Abortion could be legalized if there is an acceptable definition of when life begins" (Kamakura & Balasubramanian, 1989).

Tailoring has also been applied to educational achievement testing. Here tests are "tailored" to local conditions when they employ only items which are relevant to a particular local curriculum (Rudman, 1987). If Johnny or his class has not studied long division, then items

on that topic are eliminated from the test. In both cases, the goal is to produce an alternative to a standardized administration, in which subjects all receive the same content, regardless of predisposition or local conditions.

Moreover, researchers are currently attempting to develop new testing procedures that can measure underlying understanding, strategies, and the like (Glaser, 1981; Nichols, 1994; Royer, Cisero, & Carlo, 1993). Indeed, I have participated in this effort myself, helping to develop a test intended to measure mathematics concepts and strategies (Ginsburg & Baroody, 1990). No doubt new approaches to testing will help us gain greater insight into cognitive processes, although we do not yet know how substantial these improvements will be. I suspect that the very nature of standardized testing sets limits on what it can accomplish with respect to entering minds.

Other methodological innovations are more far-reaching. Thus, researchers have gone outside the laboratory to investigate memories of everyday life (Pillemer & White, 1989). Investigators have employed "think-aloud methods" to examine complex problem solving (Ericsson & Simon, 1993; Schoenfeld, 1985). The microgenetic method – repeated observations and interviews of individual children working on a set of problems over a relatively long period of time – has been used to examine development (D. Kuhn & Phelps, 1982; Siegler & Crowley, 1991). Others have called for methods that will allow for consideration of meaningful activities in their social context (D. Kuhn, 1992). Ethnography – the detailed observation of persons in their natural environment and culture – is traditional in anthropology and has much to offer research in some areas of psychology (Fetterman, 1989).

Innovations like these have "opened up" our methods, freeing them from traditional constraints. Consider next a powerful method that is relatively "old."

## THE CLINICAL INTERVIEW

This book is about one class of nontraditional methods, the *clinical interview*, originally developed by Jean Piaget. In recent years the clinical interview has achieved some popularity in research and prac-

tice related to developmental psychology. (I often refer to "the clinical interview" in the singular. But remember that the reference is to a *class* of methods, of which there are many, not just one.) Contemporary investigators have developed forms of the method, to be used alone or in conjunction with other procedures, to investigate topics as diverse as moral judgment (Smetana, 1982), mathematical thinking (Gelman, 1980), reading (Garcia, 1991; Wixson, Bosky, Yochum, & Alvermann, 1984), understanding of maps (Liben & Downs, 1991), and understanding of physics (Johansson, Marton, & Svensson, 1985). It is also used in clinical practice (Garbarino, Stott, & Faculty of The Erikson Institute, 1992; Greenspan & Thorndike-Greenspan, 1991; Hughes & Baker, 1990) and in vocational counseling and medicine (Millar, Crute, & Hargie, 1992). Use of clinical interview methods is increasing in mathematics education (Ginsburg, Jacobs, & Lopez, 1993), where the National Council of Teachers of Mathematics encourages its members to conduct flexible interviews in the classroom (National Council of Teachers of Mathematics, 1989, 1995), and in educational research generally. I have even found an example of the clinical interview used by police investigators to enhance the accuracy of eyewitness testimony (Kohnken & Brackmann, 1988). Clearly, use of clinical interview methods is not limited to the topics originally investigated by Piaget.

Despite its emerging popularity, the clinical interview method is not sufficiently understood. This is partly because the method violates the standard paradigm. It directly challenges the traditional point of view concerning scientific method in research and practice. It forces us to think very differently about how we should enter children's minds.

The purpose of this book is to introduce you to the clinical interview and to convince you that it offers a viable alternative, or at least a supplement, to traditional methods. The clinical interview can help you understand how children construct their personal worlds, how they think, how their cognitive processes (at least some of them) operate, how their minds function. I believe that the clinical interview can make important contributions both to basic research and to applications in the clinic, the school, and elsewhere too, like the courts and the physician's office. It can help us understand such phenomena as

problem solving in school, world views and concepts of reality, understanding of one's illness, memories of abuse, stereotypes of others, and basic concepts of the self. It can be useful both for investigators in search of general knowledge concerning children and for clinicians seeking to capture the individual child's uniqueness.

Although the method has attracted increased interest in recent years, particularly in cognitive developmental research and in education, some psychologists continue to hold misconceptions concerning its scientific status or do not fully understand its strengths. Many researchers seem to believe that unstandardized clinical interview methods of the type I will describe are merely "informal" or even sloppy and suitable at best for "pilot work," preparatory to more rigorous research. Clinicians, school psychologists, and others often tend to assume that the more "standardized" the assessment, the more "valid" it is, and that "clinical judgment" may perhaps supplement such standardized procedures but is inferior to them in accuracy and reliability (not to speak of scientific respectability). Indeed, perhaps because of their devaluation of it, some practitioners are not as skilled in "clinical interviewing" as they should be.

If you think this, I hope to broaden your view. No doubt, both basic researchers and those doing applied work can derive some benefits from traditional procedures like standardized tests. These have an important (although more limited than usually assumed) place in our array of research and assessment methods. But I hope to convince you to be more catholic in your approach. Because people's minds are so extraordinarily complex, we must expand our methodology to include the *deliberately nonstandardized* approach of the clinical interview.