DOCUMENT RESUME

ED 141 410 95 TM 006 376

AUTHOR Dyer, Henry S.
TITLE The Interview as a Measuring Device in Education.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO ERIC-TM-56
PUB DATE Dec 76
CONTRACT 400-75-0015
NOTE 17p.; Some parts may be marginally legible due to small print of the original document

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Education; *Educational Diagnosis; *Interviews; *Measurement Techniques; *Surveys
IDENTIFIERS Diagnostic Interviews; Open Interviews; Standardized Interviews

ABSTRACT

The interview is conceptualized as a dyadic process the purpose of which is to obtain usable information either about the cognitive and noncognitive attributes of the person interviewed or about attributes of educational institutions with which the interviewee is associated. Interviews are seen as falling into two broad categories: standardized interviews and open interviews. The standardized type is further divided into two subcategories: the diagnostic type of interview and the survey type. The techniques involved in the several types of interviewing are adduced from a number of illustrative projects involving interviews conducted in various educational settings. The author suggests that there are five concepts that are basic to any form of measurement and that, insofar as the interviewing process incorporates these concepts, it qualifies as a measuring device. Generally speaking, the standardized type of interview tends to meet these criteria more readily than does the open type, but the latter is seen as more likely to uncover new dimensions to be measured. The practicality of both types of interviews is briefly assessed. (Author)

ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08540

TM REPORT 56

DECEMBER 1976

THE INTERVIEW AS A MEASURING DEVICE IN EDUCATION

Henry S. Dyer

ABSTRACT

The interview is conceptualized as a dyadic process the purpose of which is to obtain usable information either about the cognitive and noncognitive attributes of the person interviewed or about attributes of educational institutions with which the interviewee is associated. Interviews are seen as falling into two broad categories: standardized interviews and open interviews. The standardized type is further divided into two subcategories: the diagnostic type of interview and the survey type. The techniques involved in the several types of interviewing are adduced from a number of illustrative projects involving interviews conducted in various educational settings.

The author suggests that there are five concepts that are basic to any form of measurement and that, insofar as the interviewing process incorporates these concepts, it qualifies as a measuring device. Generally speaking, the standardized type of interview tends to meet these criteria more readily than does the open type, but the latter is seen as more likely to uncover new dimensions to be measured. The practicality of both types of interviews is briefly assessed.

INTRODUCTION

For the purposes of this monograph, an interview is thought of as a conversation between two people in which the aim is to generate information either about the person being interviewed (the respondent*) or about other matters with which the respondent is presumably familiar. All such interviews are to some extent structured. Those that tend to be highly structured I call standardized interviews; those that tend to be unstructured I call open interviews. In some cases, an interview may employ both highly structured and open-ended questions.

The *standardized interview* is typically conducted by interviewers specially trained to follow a *standard set of procedures* to insure as far as possible that the answers given by all respondents to all interviewers can be readily compared. There are many kinds of standardized interviews. The examiner administering an individual intelligence test to a child is an example of one kind; the pollster checking off on a clipboard a householder's answers in an opinion survey is an example of another.

In the *open interview*, the respondent is encouraged to talk freely and at length on topics that may be variously worded and ordered by the interviewer to suit the occasion. It is the sort of interview that seeks to explore the respondent's thinking and experience in considerably greater depth than is possible within the more rigid framework of the strictly standardized interview. It may be used for working up case studies of persons, programs, or institutions. It requires special expertise of the sort one expects to find in an experienced clinical psychologist or social anthropologist.

It may seem obvious from the foregoing sketch that the strictly standardized interview is more likely than the free-wheeling open interview to meet the requirements of what we ordinarily think of as *measurement*. Nevertheless, there are certain ideas associated with the theory and practice of measurement that can be applied to some degree in any kind of interview. I shall discuss the applicability of some of these ideas as we consider examples of each type of interview below. But to lay the groundwork for this approach, let us first consider some of the things we mean by the term *measurement* in any context.

*Throughout this monograph, the terms *respondent* and *interviewee* are used interchangeably to denote the person interviewed.

MEANINGS OF MEASUREMENT

During a long history stretching back to ancient times, the term *measurement* has acquired a variety of meanings depending on the kinds of phenomena with which it deals and the purposes it is intended to serve. Measuring the width of a table to see if it will fit through a doorway is rather different from measuring the attitudes of children to see how they feel about going to school. Despite such differences, however, there are at least five concepts that seem to be basic to any and all forms of measurement:

1. Attributes. Measurement always has to do with the attributes of whatever we choose to observe. As one writer (10) on the nature of measurement puts it:

> . . .what is measured is not an *object* but a *property* or *attribute* of an object. One does not measure a table, but one may measure a table's length (or width, height, weight, light reflecting property, etc.). One does not measure a student, but one may measure his weight or his achievement in arithmetic.

Similarly, one does not measure a school; one measures such attributes of a school as its average daily attendance, the mobility of its student body, its social climate, and the like.

2. Comparative judgment. An extension of the foregoing concept is that an attribute is defined by the operations we use to measure it. The operations we call measurement vary widely in accordance with the sort of phenomena out of which any given attribute is constructed. However, all such operations have one thing in common: They all involve comparative judgment of one kind or another. An observer compares tables and doorways to see which tables can be shoved through which doorways and thereby arrives at the construct of *width*. An observer compares what one child says about himself with what others say of themselves and arrives at some such construct as *self-esteem*.

3. Index numbers. It is further characteristic of all measurement that the results of comparative judgments may be indexed by one or another of several kinds of numbers such as simple counts, averages, percents, percentiles, ratios, or numerically labeled positions in a rank order. Some kinds of numbers are more amenable than others to rigorous mathematical treatment. The numbers associated with the physical sciences tend to be more mathematically rigorous than those associated with the social and behavioral sciences. But this fact need not mean that the attributes of people and their institutions are inherently less measurable than the attributes of inanimate objects.

4. Validity. One of the primary concerns of all measurement has to do with the validity of the constructs and numbers that emerge from our comparative judgments. A measure is said to be valid to the extent it is credible, communicable, and useful for some designated purpose. If we report that a table is 36 inches wide, we should expect that people who have business with tables: 1) will be justified in believing that we have actually made the indicated observation and have not fudged the data; 2) will recognize what we mean by the term *width;* and 3) will find the observation useful in predicting such things as whether the table will fit through a particular doorway. Similarly, if we report that 36 percent of a high school senior class admits to having cheated on examinations, we should expect that the school staff: 1) will want some indication of the truthfulness of the answers we got from the students we interviewed, 2) will have some common understanding of what we do or do not mean by *cheating;* and 3) will find the information useful in deciding whether to change the examination system.

5. Estimation of error. Not least among the problems that have to be faced in any kind of measurement is that of estimating how wide of the mark we are likely to be, for we have to recognize that the numbers of measurement are always less than perfectly precise. Every kind of measurement is embedded in error. An eminent physicist, Percy W. Bridgman (1), having in mind the highly sophisticated measures in his own field, states the case for all types of measurement:

> [A]ll results of measurement are only approximate. That such is true is evident after the most superficial examination of any measuring process. . . . we never have clean-cut knowledge of anything. . . all our experience is surrounded by a twilight zone, a penumbra of uncertainty. . .

If the "penumbra of uncertainty" surrounding any measure is large, we say that the measure is of low reliability or that it has such a large standard error of measurement that our comparative judgments may be little better than random. A concept helpful in interpreting the degree of randomness in a set of observations is that of the 95 percent confidence interval which defines a band of random errors in such a way that we estimate the odds to be 95 to 100 that the "true" value of an observation lies somewhere between x and y.* To take a hypothetical case, we might say something like this:

> Mary's observed percentile rank on the attribute of self-esteem is 75. Our estimate of the 95/100 error band surrounding a percentile rank of 75 is that it extends from a percentile rank of 60 to a percentile rank of 85. Our best guess, therefore, is that there are 95 chances in 100 that Mary's "true" percentile rank is somewhere between 60 and 85.**

It should be noted that the error band (or indeed any indicator of the reliability of a measure, such as the reliability coefficient) *is itself never more than an approximation* based on such evidence as we can assemble for the purpose. This is to say that the boundaries of the "penumbra of uncertainty" are themselves always fuzzy and uncertain. Nevertheless, we have various methods for estimating where the boundaries may be, and the making of such estimates is essential in any process that we may, in good conscience, call measurement. Care in making such estimates has much to do with the degree to which an interview may qualify as a measuring device.

*In statistical jargon, the 95 percent confidence interval, or error band, is that which extends from 1.96 standard errors below to 1.96 standard errors above the observed value.

**The hypothetical 95/100 error band in this case is not symmetrical around the hypothetical percentile rank of 75 because percentile scales have the characteristic that as one moves away from a percentile rank of 50 toward the high or low end of the scale, the units become smaller and smaller.

STANDARDIZED INTERVIEWS

Standardized interviews serve two general purposes. If the primary purpose of the interview is to measure attributes of the person being interviewed, I call it a *diagnostic interview*. If the primary purpose is to measure the attributes of collectivities such as groups of respondents or educational programs and institutions, I call it a *survey interview*. Although the procedures and problems associated with both types of interviews have much in common, the differences between them are sufficiently great to consider them separately.

The Standardized Diagnostic Interview—Cognitive Attributes

The diagnostic interview that, over the years, has become the most highly standardized is the interview we associate with the administration of an individual intelligence test, such as *The Wechsler Intelligence Scale for Children—Revised* (18). The Wechsler can indeed be regarded as a kind of paradigm of standardized interviews in general and diagnostic interviews in particular. We shall, therefore, consider it in some detail.

This interview involves a series of intensely human transactions between the interviewer and the respondent in which the interviewer plays four roles almost simultaneously:

- A *stimulator* of responses on the part of the interviewee by means of questions that may or may not be accompanied by the presentation of tasks
- An *observer* of the responses so stimulated
- An *evaluator* of each response as it occurs
- A *recorder* of the evaluation (or score) assigned to each response

These four roles, as we shall see, enter into all types of interviewing, though in some cases, one or another of them may be played by persons other than the actual interviewer.

In all four roles, the interviewer is expected to stick closely to a set of printed standard procedures while at the same time handling the interview situation in a sufficiently flexible way to enlist maximum cooperation on the part of the respondent. Some excerpts from the manual of the *Wechsler Intelligence Scale for Children-Revised* (WISC-R) serve to illustrate the demands on the interviewer for preserving just the right balance between rigor and flexibility in conducting the interview (18):

> The WISC-R should be administered and scored by a competent, trained examiner [who] must carefully follow the directions in this manual. . . . The examiner must not change the phrasing of a test item, spell a word, or provide assistance beyond permissible bounds. Time limits must be strictly observed. . . . Adherence to standardized procedures does not mean that the battery must be administered in a rigid and unnatural way. The words used to introduce test items should be spoken in a natural, conversational tone. The experienced clinician will interject appropriate comments to promote the child's interest in the tasks, to reinforce his effort when this is needed. . . [p. 53] . . . Making the testing experience satisfying to both child and examiner places great demands on the examiner's clinical skills. . . . There is no magic formula for "reaching" a child; approaches that succeed with some children may antagonize others. [p. 55]

Although the mandatory procedures for conducting this kind of diagnostic interview are spelled out in great detail, it is nevertheless clear that the interviewer as stimulator and observer bears a considerable responsibility for exercising his own best judgment in deciding how and when to "interject appropriate comments" without going beyond "permissible bounds." For example, in administering the Vocabulary section of the test, the interviewer is instructed to use "the local pronunciation of each word or the pronunciation you believe to be familiar to the child" (p. 89). Moreover, throughout the battery, there are places where the interviewer is told to *probe* for an answer if the response seems ambiguous. In the Comprehension section, for instance, there are the following instructions:

> If the child is hesitant, encourage him with such remarks as "Yes" or "Go ahead." If the response is unclear or ambiguous, you may say, "Explain what you mean" or Tell me more about it." [p. 96]

Clearly, the intent of instructions like these is to put strict limits on the kind and amount of permissible probing, but clearly also the limitations of language in the communication of the instructions are such that some interviewers might well suppose that they have more leeway for probing than is intended, and others might suppose they have less.

Similarly, in evaluating responses as they occur, the interviewer is expected to adhere scrupulously to criteria spelled out in the manual in the form of actual responses that illustrate those that should receive full credit, partial credit, or no credit. But again, the interviewer's judgment is frequently and necessarily called into play, especially when probes are needed to secure an interpretable response. An example from the scoring criteria for Question 5 in the Comprehension section shows something of the kind of judgmental problem the interviewer is up against. The question is: "What is the thing to do if you lose a ball that belongs to one of your friends?" A full-credit response is one which, in the interviewer's opinion, indicates that the child has grasped the concept of replacing a loss for which he or she is responsible. Sample responses intended to guide the interviewer in making this determination are as follows:

> *2 points*—Give him (her) one of mine. . . . Try to get it back or replace it. . . . Pay for it. . . . Buy her a new one. . . . Buy another one if I can't find it.
>
> *1 point*—Try to find it (Q)* then tell my mother (teacher), she'd look. . . . Tell him and let him decide (Q). . . . Try and help her find the ball (Q). . . . Look all over for it (Q).
>
> *0 points*—I guess I'd just cry. . . . Tell him you're sorry. . . . Tell him to find it. . . . Call him up. . . . I'll get in trouble. . . . Tell your friend. [p. 178]

*The "Q" indicates a point where the interviewer probes for a scorable answer.

As exemplified by the diagnostic interview, four things about standardized interviews in general may be noted at this point. First, the reliance on the interviewer's subjective judgments within the context of an elaborate set of specified procedures is usually justified on the ground that, by adapting the conditions of the interview to the varying conditions of respondents, the comparability of their responses is maximized. This is to say that the standardization of the interview is enhanced if the interviewer is given a degree of flexibility in her or his several roles and that rigid adherence to absolutely uniform procedures is less rather than more likely to result in the level of comparability essential to sound measurement.

Second, there is implicit in the notion of standardization that the interviewer shall have undergone rigorous training in the roles to be performed—training that involves, among other things, a goodly amount of supervised practice with a variety of respondents. In the absence of such practice, it is unlikely that the interviewer will be capable of staying within "permissible bounds" of procedure while exercising good judgment about when and how to vary her or his behavior within those bounds.

Third, the maintenance of just the right balance between freedom and control in the conduct of the interview is regarded as one condition for ensuring that each response will be such as to contribute to the *validity* with which the attribute in question is measured. Once aspect of validity has to do with making as sure as possible 1) that the respondent is sufficiently attentive to hear each question as it is asked; 2) that the respondent is making an effort to understand the purport of the questions; and 3) that the respondent is also making a genuine effort to answer each question as he or she understands it. In respect to each of these matters, the validity of the interview is impaired if the interviewer goes either too far or not far enough in helping the interviewee cope with any question.

By the same token, the validity of the diagnostic interview—*or indeed any kind of interview*—depends heavily on the degree of *rapport* which the interviewer is able to establish and maintain during the course of the interview. Rapport is a subtle quality. It has to do with the nature of the relationship between the interviewer and the respondent. If the rapport is good, the respondent feels at home in the interview situation, has confidence in the interviewer, and is ready and willing to cooperate. If the rapport is bad, the interviewee may feel threatened or uncomfortable during the proceeding, may have deep suspicions about the motives of the interviewer or the purpose of the interview, and, as a consequence, may answer questions reluctantly, untruthfully, or not at all.

Once the raw response data from the diagnostic interview are in hand, the next step is to summarize them in such a way as to produce a numerical index that constitutes a measure of the attribute in question. This step is one that may or may not be performed by the person who has conducted the interview. In the case of the WISC-R, the measure that results from the summarization is the so-called deviation IQ, which, simply stated, is an index of how the respondent's overall cognitive performance, as sampled in the interview, compares with that of a sample of other respondents of approximately the same age.*

Finally, there comes the question of the range of error that must be taken into account when interpreting the results of the interview. In the case of the WISC-R, the data on this matter are provided in terms of reliability coefficients and standard errors of measurement. Wechsler reports that the average standard error of measurement for all age groups $6\frac{1}{2}$ to $16\frac{1}{2}$ is 3.19 IQ points (Wechsler, Table 10, p. 30). By using this information, we may estimate that the 95/100 error band in terms of IQ is about $\pm$ 6.25 IQ points—that is, $\pm$ (1.96 X 3.19). This means that if a student's observed IQ is 100, the chances are 95 in 100 that his or her "true" IQ probably lies somewhere between 94 and 106. Translated into percentile ranks, this says that, if a student is at the 50th percentile of his or her age group, one may infer that her or his "true" percentile rank on the WISC-R lies somewhere between a PR of 34 (IQ equivalent 94) and a PR of 66 (IQ equivalent 106). If his or her PR is near the low or high end of the percentile scale, the 95/100 error band is narrower. Thus, if the observed PR is 25 (IQ equivalent 90), the chances are probably 95 in 100 that the "true" PR lies somewhere between 14 and 38.

The foregoing estimated error bands, in terms of percentile rank, may strike the reader as indicating an unexpectedly large amount of uncertainty in the assessment of cognitive attributes. But the degree of uncertainty one finds in the WISC-R results is likely to be the *least* one can expect from any standardized diagnostic interview, or indeed from any method of measurement that relies on interactions between persons, however scrupulously controlled the interactions may be. Moreover, one needs to bear in mind that the error bands we have shown above for the WISC-R are based on reliability data obtained from a series of interviews conducted by presumably well-trained examiners under presumably optimum conditions. Whenever the conditions of interviewing are something less than optimum, the 95/100 error bands will, of course, be wider. And, by the same token, if the conditions of interviewing are *unknown*—as they sometimes are in run-of-the-mill situations—we are up against a situation where we can have no idea whatever of how much confidence we can put in the numbers.

The Diagnostic Interview—Noncognitive Attributes

The use of the diagnostic interview as a measure of the cognitive attributes of children has benefited from research and experience that goes back to the turn of the century. Attempts to use similarly standardized techniques for measuring the noncognitive attributes of school children are more recent, more scanty, and less well-developed. One recent effort along this line is the *Thomas Self-Concept Values Test* TSCVT (16) for use with children age four to nine. Here the purpose of the interview is to evoke from the child a series of brief verbal responses from which one may infer how the child pictures himself or herself and how he or she thinks others—mother, teacher, other children—

*In view of the fact that the deviation IQ is in no sense a quotient and in view of the interminable confusions and controversies that have grown up around the term IQ, it would be helpful if the term could be wholly expunged from the vocabulary of measurement and replaced by an index less vulnerable to misinterpretation or over-interpretation. One such index of overall cognitive performance could just as well be the percentile rank (PR) which states directly and unambiguously—*though of course, as always, only approximately*—where the respondent stands on the attribute in comparison with others of her or his age group. To say of an individual that, in the cognitive performance defined by her or his responses, she or he stands at about the 75th percentile of ten-year-olds ought to be more communicative and less confusing than to say that she or he "has" an IQ of 110, as though one possesses such a number in the same sense that one "has" a tongue or brown eyes or a liver!

think of her or him: whether happy or sad, smart or not very smart, scared of people or not scared, and so on. An interesting innovative feature of this particular technique is that, while the child is answering the interviewer's questions, he or she has his or her attention focused on a Polaroid snapshot of himself or herself.

As with any standardized interview, the manual of the Self-Concept Values Test prescribes the procedures for eliciting, observing, evaluating, and recording the responses. In respect to these matters, however, a good many details have still to be worked out through experience. For instance, one does not find in the manual instructions to the interviewer that are comparable in elaborateness to those the authors of the WISC-R have found through long experience to be necessary. Little is said about the matter of rapport or about the amount of probing that is allowable. No examples are given of the ways children actually respond—examples an interviewer might need for deciding in doubtful cases whether to probe or how a response should be scored when it does not exactly match the words given in the key. Some effort has been made to estimate the reliability of the scores obtained from the interview. From the data presently available, one may very tentatively infer that the 95/100 error band around a total self-concept score at the 50th percentile runs from a PR of 21 to a PR of 79.

Thus, although one might say that the TSCVT in its present stage of development has a minimum claim to be a measuring device, one can hardly regard it as ready for routine use in the measurement of children's self-concepts. The same can be said of diagnostic interviews designed to measure such attributes as students' attitudes toward school and learning. In other words, we seem to have a long way to go before we have enough information about the error component in standardized interviews for assessing pupils' affective attributes to enable us to exercise appropriate precautions in interpreting the numbers they yield. This is not to say, however, that the goal is forever unattainable, nor are we suggesting that other techniques, such as the self-administered questionnaire, yield numbers that can be interpreted with any greater confidence. To the contrary, the self-administered questionnaire, though apparently more efficient, simply buries many of the problems that come to the surface in the face-to-face interview.

How Meaningful Are the Responses?

In the previous discussion of the validity of the interview, I mentioned the credibility aspect of the validity question—that is, procedures for ensuring as far as possible that each response shall reflect what the interviewee really knows or feels. There is another somewhat different aspect which may be put as a question: Assuming that the particular response to a particular question does indeed reflect some part of what the respondent really knows or feels, how reasonable is it to infer that this piece of information helps to define an attribute that carries the same meaning for different people? One way in which this question has been dealt with is to ask a group of judges to rate or rank a set of responses to a given question to see how closely the judges agree in their ratings. If the agreement is high, then one can infer that the responses do indeed help to define a recognizable attribute, and thus a communicable construct.

A number of years ago, for example, this writer wondered to what extent the recorded responses of a group of ninth grade children could be differentiated by a mixed group of 20 adult judges with respect to the attribute *prejudice*. One of the questions the interviewers had asked of the students was intended to evoke a response indicating the degree to which the respondent might harbor prejudice against social classes different from his or her own. The question was in two parts as follows:

1) Would you rather have as your friends boys and girls who are twice as rich or half as rich as you? (Why?)
2) Why wouldn't you want friends who are twice (half) as rich as you?

In this case, the judges were able to agree only moderately well in ranking the responses from most to least prejudiced. The average intercorrelation among the ranks was .64 (5). However, there was fairly good evidence that they could agree quite well about responses that they rated at one extreme or the other. For example, 19 of them ranked the following response either as the most extremely prejudiced or close to being so:

> I'd want them half as rich. (Why?) These rich people are conceited. They're high hat. They send their children out all slicked up which more or less disgraces *us*. [p. 220]

Similarly, 19 percent of the judges ranked the following response as most or next-most *un*prejudiced:

> It wouldn't matter to me. (Why?) It's what they are, not the money they have, that counts.

But the judges were in far less agreement about the following response:

> Oh, I don't know. It depends on what they're like. If a person's rich, sometimes they're not so nice.

Three of the judges rated this response as close to most prejudiced; 5 rated it as close to least prejudiced; the rest of the ratings were scattered through the middle range.

These results suggest that the attribute *social-class prejudice* is recognizable in responses at the extremes, but that in between the extremes there are many responses that cannot be validly recognized as belonging under the rubric of prejudice. A somewhat similar kind of judgmental analysis has been used with responses evoked in interviews employing projective techniques. One investigator, for instance, attempted to measure the attitudes of young children toward their school by showing them pictures of various classroom situations and asking in each case what the respondent thought was happening (3). Four judges, working independently, were able to agree rather well whether any given response reflected a positive or negative attitude toward schooling (pp. 70-71).

This writer is unaware of any attempts to apply this kind of validity test to the responses obtained from standardized interviews in the cognitive domain. It would be useful to know, for example, to what extent one might find agreement that the responses to the WISC-R question illustrated above were seen as representing varying degrees of *comprehension* or whether they might be more readily recognized as showing individual differences in respect to some other attribute such as *moral judgment*. Similarly, in the TSCVT mentioned above, there might be some dispute over the way certain responses should be evaluated. One of the

questions, for example, asks the child whether he or she perceives him- or herself as "strong" or "weak." If the respondent is a boy, the response "strong" is scored as indicating positive self-concept; but if the respondent is a girl, the same response is scored as indicative of negative self-concept—on the theory, one imagines, that females are supposed to perceive themselves as weak. It seems to this writer that if this way of evaluating the response were submitted to a random sample of men and women, there might be some disagreement about a conception of self-concept that makes this kind of distinction between the sexes.

As matters now stand, one has to conclude that in nearly all types of diagnostic interviews, however scrupulously standardized with respect to procedure, the amount of public consensus concerning the categorization and evaluation of individual responses is largely unknown. And this leaves in doubt some of the claims that are made for the construct validity of the response material.

The Standardized Survey Interview

Probably the most familiar example of the standardized survey interview is the public opinion poll designed to predict the voting behavior of the electorate from a small but representative sample of potential voters. The same type of survey interview has, of course, a multitude of other uses as well. It has been used to measure such things as the delivery of social services, the buying habits of consumers, and the attitudes of different segments of the public toward their schools and other social institutions.

As we have indicated above, one of the principal differences between the diagnostic and the survey type of standardized interview is that the former focuses on the attributes of the individual respondent while the latter focuses on groups of people or on programs and institutions. This difference has consequences for the nature of the sampling process and the estimation of error due to sampling. In the diagnostic interview, we emphasize the adequacy of sampling across a defined universe of responses in order to measure some attribute of an individual; in the survey interview, we emphasize the adequacy of the sampling across a defined universe of individuals in order to measure some attribute of a group.

Two surveys which, taken together, exemplify many of the procedures, possibilities, and problems that turn up in the actual use of the survey interview as a measuring device are described below. They differ in a number of ways, but primarily in that the first was a survey of people's *opinions* about their schools and the second is one that sought to uncover the objective *facts* about certain program operations.

The Opinion Survey: This kind of survey was conducted by the Opinion Research Corporation for the New Jersey State Board of Education and the New Jersey School Boards Association (12). Both groups wanted to know how the adult population of New Jersey viewed the schools and what should be done to improve them.

The procedures used in the survey may be broken down into four activities which tend to characterize any well-planned survey:

1. Developing the interview schedule. The questions that eventually formed the main body of the interview schedule grew out of a year-long series of statewide, regional, and local conferences under the auspices of a committee of the State Board of Education. Participants in these conferences included school board members, public officials, legislators, parents, school administrators; teachers, students, and other interested citizens. The conferences were freewheeling discussions in which the participants presented and debated their views about the public schools. This exercise produced two sets of statements of educational goals which, in effect, summarized the various concerns that turned up in the course of discussions. One set, called "outcome goals," described the different kinds of benefits that the participants thought students should derive from their school experience (good health habits, mastery of basic skills, understanding of and respect for different racial, ethnic, and cultural groups, and so on). The other set, called "process goals," had to do with school policies and practices such as provision of adequate guidance services, assurance that teachers were of high quality, and provision of programs for prekindergarten children. These two sets of goals were incorporated in a preliminary form of the interview schedule and tried out on a few respondents at various home locations. The tryout not only helped refine the wording of the questions and the procedures for presenting them; it also produced some additional public concerns for inclusion in the final form of the interview schedule.

2. Selecting the sample of respondents. The sampling design called for three groups of people to be interviewed: 1) a probability sample of the general public age 16 and over; 2) a separate subsample of Spanish-speaking residents to be interviewed in Spanish or English according to the wishes of the respondent; 3) a subsample of "knowledgeables"—people likely to have greater than average exposure to students or graduates of New Jersey schools. On the basis of U.S. Census data, the general public sample was so drawn that each person in the whole population of the state had a known probability of being interviewed. This part of the sampling procedure predesignated the districts, the neighborhoods within districts, the households within neighborhoods, and the specific type of individual to be interviewed within each household. The method of sampling ultimately made it possible to estimate from the sample of responses what the responses would have been if the interviews had been conducted with all adults in the state, including all members of the various subpopulations of interest (young-old, men-women, black-white, urban-suburban-rural, and so on).

3. Conducting the interviews. The actual interviewing required seven weeks during which time a corps of pretrained interviewers fanned out across the state to their predesignated locations and questioned 1,105 persons for an average of 56 minutes each. For the most part, the questions were so framed that the interviewer needed only to check off the answers on the interview schedule itself. The question on outcome goals, for example, was handled by giving the respondent a card listing the goals and making the following request verbatim:

> From each of the items on the list, please tell me whether you think it is a very important, fairly important, or not too important goal for New Jersey public schools. If you think something on the list is not a

. . proper goal for New Jersey's public schools, just let me know when we come to it.

The interviewer then read off each item number and recorded the respondent's answer by checking the appropriate space on the interview schedule. Other questions having to do with the respondent's age, educational background, income level, and the like were handled by a similar checking-off technique. In short, with some exceptions, the whole procedure—from locating the person to be interviewed to recording the interviewee's responses—was intended to require an absolute minimum of subjective judgment on the part of the interviewer.

4. Organizing the data. The organization of the response data from a survey of this kind consists of straightforward counts of the responses to each question and the conversion of the counts into percentages of the total sample or of various subsamples. Thus, it was found that 80 percent of the general public sample rated the outcome goal *respect for authority* as "very important." By comparison, only 29 percent rated the schools as doing a good or excellent job in instilling respect for authority. The 95/100 error band for estimating what the responses of the entire state population would be on these two matters was six percentage points. That is, one could say with a good deal of confidence that there were 95 chances in 100 that, if the whole adult population had been interviewed, the percentage figures would lie somewhere close to 80 percent with respect to the *importance* of instilling respect for authority and somewhere between 26 and 32 percent on the question of whether the schools were doing well in this respect. The fact that the two error bands do not overlap strongly supports the inference that, on this point, the public sees a wide gap between what it wants from the schools and what it is getting.

The New Jersey survey is a good illustration of how masses of response data can be combined, summarized, and presented in a way that makes their implications readily interpretable. Figure 1 illustrates the technique. It locates the response data in four quadrants determined by a vertical axis (percent of total public who rate each goal "Very Important") and a horizontal axis (percent who rate public schools "Excellent" or "Good" on each goal). Thus, it can be seen in quadrant I, for example, that upwards of 70 percent of the public thinks "understanding/respect for differences among people" is a very important educational goal, while less than 40 percent thinks the schools are doing a good or excellent job in this respect. Similarly, over 65 percent thinks that giving students a "desire to continue to learn" is an important goal, yet less than 40 percent thinks the schools are doing well in encouraging this desire. By comparing data in this way, one gets a graphic measure of what the public sees as the needs of the educational system.

The Fact-Finding Survey: An example of the fact-finding survey is to be found in a study conducted in 1972 by Educational Testing Service (7). The purpose of the survey was to find out what was then going on in state educational assessment programs throughout the country by questioning people who were in a position to know. The respondents were 79 individuals known to have official responsibility for the planning and operation of such programs in each of the 50 states and in Puerto Rico, the District of Columbia, and the Virgin Islands.

Although the procedural elements of this survey had a pattern resembling the opinion survey described above, there are, as we shall see, some differences of detail that mark it off as a fact-finding survey.

1. Developing the interview schedule. Like the interview schedule described above, this one was some time in the making. It grew out of an earlier exploration of state educational assessment programs in which loosely structured interviews were conducted at 51 of the same 53 locations (6). The interviewers in the exploratory survey were given a rough guide suggesting the kinds of topics to cover and the kinds of people to see and talk to. But they were left pretty much on their own to identify at each location the kinds of respondents most likely to have useful information. The result was that in total they conversed with 247 people about a wide variety of matters having to do with each state program—its purpose, policy control, funding, operational assignments, and the like. In many ways, the interviewers' methods in this preliminary foray into a subculture of the educational bureaucracy were not unlike those a social anthropologist uses in doing fieldwork.* It was from the descriptive field reports that 56 specific questions to be used in the 1972 survey were formulated. For the most part, the questions finally settled upon were of the check-off variety, with provision for open-ended responses as needed. For example, Question 20 reads:

Which of the following groups initiated the idea for this program?
- a. Governor's office
- b. Independent organization
- c. State Board of Education
- d. State Education Agency
- e. State Legislature
- f. Chief State School Officer
- g. Teachers Association
- h. Other

The option "Other," if chosen, opens up the interview. A few questions were entirely open-ended. For example, Question 51 reads: "What are the major problems related to the program?" Moreover, although this was by and large a fact-finding survey, some questions clearly called for expressions of opinion:

Question 50: In general, how well would you say the program objectives are being achieved?

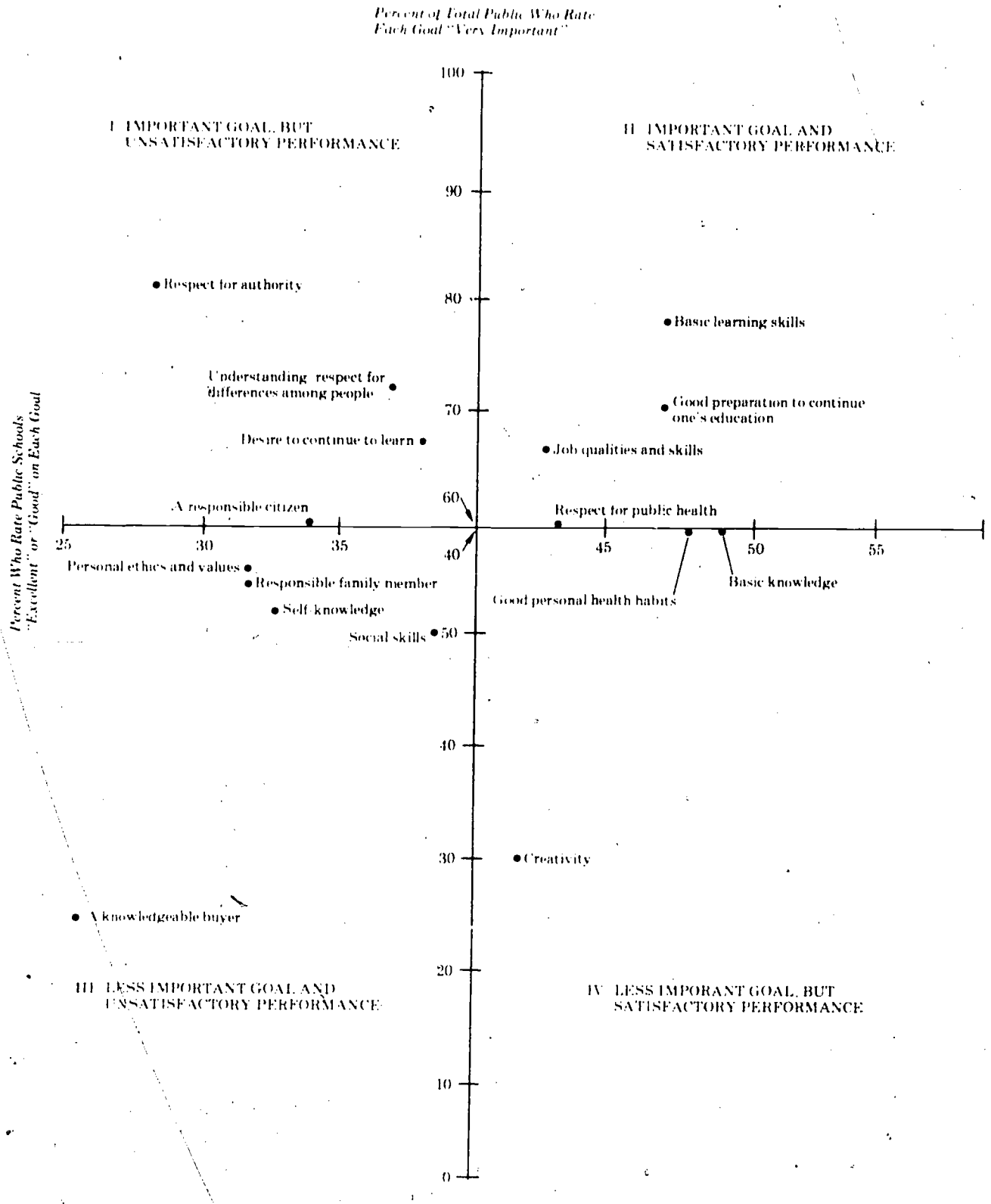
Very poorly ——— Very well

2. Selecting the sample of respondents. In this case, the notion of a probability sample was not applicable, since the aim was to secure the facts from the entire universe of 53 education agencies. The problem was to identify one or a few key people in each agency who would be in command of all the facts or would know where to get them. The

*Although the so-called anthropological approach is outside the scope of this monograph, it is nevertheless one that developers of interview schedules might use to advantage in searching for questions which are most likely to touch on the crucial features of the programs or institutions to be studied. A useful guide for this kind of exploratory work can be found in Wax's book (17) on doing anthropological fieldwork.

FIGURE 1

HOW CITIZENS APPRAISE THEIR SCHOOLS



earlier fieldwork provided most of the clues for deciding who the key respondents should be. Of the 79 chosen, 33 were identically the same persons who had been informants in the earlier fieldwork; 13 held positions the same as, or similar to, those held by other respondents in the first round; and the rest were persons responsible for brand new programs.

3. Conducting the interviews. The setup for interviewing was such that the key respondents would themselves serve, in a sense, as surrogate interviewers within their own agency. That is, they were asked to consult their colleagues as needed to ensure full and accurate answers to all questions. To this end, the interview schedule was mailed in advance to each of the 79 key respondents so that they would have ample time to get together all the requested information. The interview with each key respondent was then conducted by telephone and recorded on tape. A transcript of the full interview was then returned to the state agency for verification of the facts, and alterations were made as necessary.

4. Organizing the data. Although much of the data from the 53 interviews could be—and to some extent was—summarized in numbers, the material was so enriched by the free-response data that a discursive treatment, within a common set of categories, was used to describe each state's program. That is, the main bulk of the data was organized into what amounts to 53 case studies, one for each state. Nonetheless, some numerical comparisons across states were made with respect to certain program attributes. For example, it was found that 17 of the programs were designed primarily for decision making at the state level and 13 for decision making at the local district level; the remainder were unclear on this point since they were still in the process of getting organized. Cross-tabulation of the data on two other program attributes showed something of the measurement possibilities in a straight fact-finding survey. The two attributes are 1) whether or not the state required participation in the program by local school districts and 2) the source of funding for the program. Table 1 shows how the numbers fell.

Table 1
Source of Funds vs. Nature of Participation

		Nature of participation	
Source of funds	*No. of programs*	*Required*	*Voluntary*
State only	13	23%	77%
Federal only	20	25%	75%
State + federal	12	25%	75%
State + federal + local	8	25%	75%
Total	53	24.5%	75.5%

Without applying any fancy statistical tests, one can see from these data that the funding source had no bearing on whether program participation was required or voluntary. In view of the fact that the above measures are derived from the entire universe of state assessment programs, one might infer with some justification that errors due to sampling are nonexistent—that the *reliability* of the data need not concern us. We can, nevertheless, raise some questions regarding the *validity* of these data and of data obtained from any survey interviews.

The Validity Problem in Surveys by Interview

In connection with the fact-finding survey just described, one may legitimately wonder about the *credibility* of the answers supplied by the respondents. Did they know for certain what they were talking about? Were their perceptions of their programs possibly colored by their hopes? Or perhaps by their eagerness to give definite answers even when they were uncertain? Or by a touch of resentment at having been bothered in the first place? If a different group of interviewers and interviewees had been involved, would the results have been different? Such threats to validity are inherent not only in the straight fact-finding survey but in any survey in which a corps of interviewers encounters persons in the field. This is particularly true of the kind of house-to-house survey described previously where the members of the interviewing team, out there working alone and unobserved in the field, may well vary among themselves in the amount of care they give to following prescribed procedures, in the degree of rapport they are able to establish with respondents of differing backgrounds, and in the skill with which they make use of probes when the answers they get are ambiguous.

In recent years, there has been a good deal of research on these and related matters that touch on the validity of the survey interview. One of the classic works is that by Hyman et al. (9), and a review of much of the more recent research can be found in the article by Weiss (20) to which is appended a bibliography of some 150 references on the subject. The results of these very considerable efforts to unlock the secrets of the survey interview suggest that there are not yet many hard and fast answers to questions about how to organize and conduct the kind of interviews that will guarantee a minimum of misinformation in the measures they generate. But this outcome is hardly surprising, since interviewing of any kind (diagnostic as well as survey interviewing), regardless of the degree to which procedures are standardized, is, in essence, an *art* involving a multitude of human transactions that can vary from one situation to another in ways that are imperfectly predictable.

One presumes, however, that, like any other art, interviewing is one that can be acquired through training and experience by persons who have a knack for talking comfortably with a wide variety of individuals in a wide variety of circumstances. The major agencies involved in survey research (such as the National Opinion Research Center at the University of Chicago, the Institute for Social Research at the University of Michigan, and the Bureau of Applied Social Research at Columbia University) have given much attention to the practical problems of devising administrable interview schedules for many purposes and to the even harder problems having to do with the selection, training, and supervision of interviewers—all with a view to developing procedures that, on their face, should serve to make the interviewing operation minimally vulnerable to threats of invalidity. Much of this how-to-do-it material can be found in the publications by Collins (4), Gorden (8), Merton (11), and Weinburg (19).

There is insufficient space in this monograph to give more than the barest hint of the state of the art as it comes

through in the materials cited, but a few common sense rules of thumb may be mentioned:

1. Before an interview schedule goes into the field, it should be tried out informally on a few respondents similar to those in the sample to be interviewed in order to check on the intelligibility of the questions and to revise the language as necessary.

2. Non-English-speaking respondents should be interviewed in the language with which they are familiar, and children should be interviewed in a language which is well within their vocabulary range.

3. Before an interviewer is sent into the field on his own, he or she should be observed in action and criticized by those who have had wide experience in the art. This exercise is usually handled in two ways: by simulation techniques and by having the neophyte accompanied into the field by a trained observer.

4. Before going into the field, any interviewer should become so thoroughly familiar with the interview schedule (format, wording, types of allowable probes, method of recording responses, and so on) that he or she will be in a position to carry on each interview in an easy, conversational manner.

5. When the interview contains sensitive questions of any sort (about personal income, for example, family relations, religious beliefs, or moral values), the interviewee must be assured that his answers will be held in confidence, and *the interviewer's promise of anonymity in reporting results must be scrupulously kept.* To the extent that this sort of trust is broken in any survey, the validity of all survey data is threatened.

6. During the course of the interview, the interviewer should never suggest answers to the interviewee, should maintain a neutral stance on all questions, should avoid interjecting his or her own opinions either by tone of voice or by explicit comments, and should at all costs avoid getting into an argument with the respondent.

7. If a predesignated interviewee is not reached on an initial call, the interviewer should make every effort to call back at a time when the interviewee will be available. Only a few missed cases can so bias the results from a preselected probability sample that their validity as a measure of any population attribute can be reduced to zero.

8. A spot check by supervisory personnel should be made from time to time to provide assurance that each interviewer is actually conducting the interviews assigned to him and is not turning in fictitious records.

OPEN INTERVIEWS

As we have seen, the survey type of standardized interview usually includes some open-ended questions that encourage the respondent to enlarge on his answers. In reporting this free-response material, one has three choices. One may simply smooth it into readable prose and let it speak for itself. One may take a phenomenological approach and try to figure out what typically goes on inside the respondent's head and present one's inferences regarding the same. Or one may take an additional step and code the responses in accordance with the logical categories one finds in them. It is this coding operation that provides the basis for turning free-response material into measurable attributes.

Since the wholly open interview produces free-response material only, it puts a heavy burden on the investigator to read his or her way into the material, and to come up with a coding scheme that can be used to create, at a minimum, ordinal categories for containing and comparing the output of any respondent with that of other respondents—in a word, to devise some sort of measure.

In the following pages, we shall consider three examples of studies which have used the open interview in an educational setting and which demonstrate more or less successful attempts to tease out measurable attributes from large quantities of free-response data. The first study deals with the forms of student development in a liberal arts college (13); the second, known as the Pathways Project, deals with the educational experiences of black youth growing up in the ghetto (14, 15) and the third study deals with teachers' understandings of the curriculum in open education (2). In each of these cases, it is to be noted that we are focusing on "studies"—research efforts in which the open interview was not only a tool for data gathering but was itself an object of inquiry.

The Forms of Development Study

The Forms of Development Study exhibits the use of the open interview at its most open. At the same time, it demonstrates a highly self-conscious effort to organize masses of free-reponse material in such a way as to conform to certain of the canons of measurement as I briefly sketched them at the beginning of this monograph. The study started out to be no more than a few case histories descriptive of the changes that take place in students during four years in a liberal arts college. The study wound up with a nine-point ordinal scale purportedly capable of measuring how far any liberal arts student has moved through successive stages of development encompassing the student's intellectual, emotional, and moral outlook.

The measure so derived defines both a complex of student attributes and also the elusive concept which we call a "liberal arts education." From beginning to end, the study took approximately 10 years and produced on tape 464 unstructured interviews with 140 students deemed to be representative of the student population in the college they were attending. At the outset, the plan was to interview each student in the sample at the end of each of his or her four years in college. This resulted in 84 sets of tapes covering all four years. These provided the material for formulating a nine-point scale on which each year's output for each student could be positioned. The author of the study is careful to point out that, in accordance with the general principle that "the act of observation always influences the events observed" (p. 27), the interviewing itself must have affected to some unkown extent some part of the changes in scale position that occurred among the respondents from year to year. But this predicament,

though often overlooked, is one shared by every kind of measurement.

What, then, is the nature of the open interview, as exemplified in this study? It can best be understood by examining the interviewer's role as *stimulator*. In a word, the task is to help the respondent to reinstate, in effect to relive, salient events of the preceding year as they happen to come to mind, and to introspect out loud about the thoughts and feelings accompanying each experience. One well-known manual on interviewing in another connection refers to this process as "retrospection" (11, pp. 28-39). In the Forms of Development Study, the interviewer, after observing a few polite amenities and turning on the tape-recorder, typically began the initial interview in the freshman year with the question: "Why don't you start with whatever stands out for you about the year?" (p. 19). From then on, the interviewer left the respondent to grope his or her way back, while interjecting only innocuous expressions of interest to fill in the most awkward pauses and to keep the respondent going on the self-search. Two excerpts from the transcript of the opening exchanges of one of the initial interviews give a rough idea of how the thing goes:

I. You let me know if you mind if I record, OK? Sit down?/Thanks./Well, as you gathered, I guess, from the letter, we thought maybe you'd be willing to come in and sort of look back/Yeah/...and tell us how the year went, and how you feel about it. (*Long pause*)

S. Uhhm. Well. it's a subject I'd like to talk on, actually. I suppose every freshman wants to shoot off about their freshman year. (*Pause*) Good things, bad things, I guess./Yes/I don't know, I (*Pause*) I really don't know where to start. (*Pause*)

I. Well. wherever, I think, sort of - ah - looking back over what sort of things *stood out*, in one way or another as you - (*Pause*) ah -

S. Well, I know that it was sort - ah -sort of unwise for me to make any decisions about classes or courses to take before I came here. Actually - ah - I had a tentative list of courses. and the second day I was here, everything was completely changed. I, my ideas. values, everything was *completely* changed the minute I started talking to roommates and other people in the dorm, and so forth.

I. Sort of right away. some sort of change? [pp. 20-21]

Then later on. the student having brought up the bad experience with a course for which he hasn't found "the key," the interviewer proceeds:

I. The others are up and this, in this course, you. you don't feel as though you've found the handle. more or less, that you spotted in the others?

S. Well - ah - I think this. this course is really a good course, that's the bad part of it. (*Pause*) I, I think the reading list is probably the best I've had/Uhuh/ the lectures are good, and so forth. But I, I...if you don't get a good mark every now and then, it sort of sours the course for you (*Chuckle*)/Yeah/I think that's the prime thing - ah - (*Pause*) Oh, I don't know, the other thing I can remember is that - ah - (*suddenly raises voice*) - *I think that pre-meds are the - ah - the greatest group of cut-throats I've ever met in my life.*

I. You sort of found that here, too? [pp. 22-23]

From these excerpts, it may be noted that, once started, the interviewer does not lead the respondent into speaking on any particular topic, but simply tries, in a quite informal way, to help the respondent get out his or her thoughts and feelings about any experience that happens to come to mind. That is, the interviewer's interjected remarks are not in any sense in the nature of "probes," but more in the nature of helps to the respondent to do his or her own probing. It may be noted further that, as the respondent becomes accustomed to the nature of the dialogue, his or her "retrospections" come more readily to the surface—(the outburst about the "greatest group of cut-throats"). At no time during the interview does the interviewer play the *evaluator* role or give any hint of doing so. The role of *recorder* is delegated to the tape machine.

The full description of the nine positions of the developmental scale, together with defining examples of responses excerpted from the transcripts, occupies 118 pages of text. The elaborateness of this "scoring scheme" is due to the heavy reliance upon large chunks of actual response materials to flesh out the high-level abstractions by which each of the scale positions is labeled. Though greatly expanded, the technique recalls that which we saw in connection with the evaluation of responses to questions in the wise where the "scoring criteria" run to a mere 32 pages.

The reduction of the scoring scheme to a set of more quickly comprehended categories entailed the writing of a *Judge's Manual* containing a less formidable description of the scale positions (pp. 29-40) together with a chart and a 21-item coding scheme. The validation of the scale consisted of having a group of judges, with manual in hand, independently rate random samples of the students' transcribed interviews to see how closely they agreed on the positioning of each one. Despite the enormous complexity of the task, the results suggest that, with adequately instructed judges, the level of agreement can be quite high: Inter-judge agreement. in terms of "mean estimated reliability of the mean rating for individual interviews for each of the four years was found to be respectively, +0.966, +0.875, +0.872, and +0.916" (p. 12). These results suggest that if one is willing to take the time and trouble, the marriage of the wide-open interview with measurement is, within limits, achievable. Whether the trouble is worth taking depends upon how much one cares about measuring the subtler processes of students' development as they are mediated by the more elusive processes of educational institutions.

In the case of interviews as open as those in the Forms of Development Study, one is always left wondering how different any student's retrospections would have been if the interviewer had been a different person. This, of course. is the same kind of interviewer-respondent interaction problem that turns up in the standardized interview as well, and it is probably no more nor less a source of error in the open interview than in the standardized interview. Nor is it a source of error the extent of which can be readily estimated in either case.

The Pathways Project

From the point of view of this monograph. our principal interest in the Pathways Project stems from two facts: 1) In its early stages, it represented a huge effort to shape

masses of free-response material into measurable categories; 2) in its later stages, it wisely settled for a series of case studies in narrative form as the only feasible way of conveying some idea of the innumerable dimensions contained in the data.

Like the Forms of Development Study, the Pathways Project was of the longitudinal type. It began with interviews of 61 black youngsters all of whom were boys from poor families and all of whom were attending the same nearly all-black junior high school. After three years, 55 of the same boys were located and reinterviewed, and three years after that, 15 of the 55 were interviewed a third time. The first set of interviews ran from 10 to 20 hours with each boy in sessions of two or three hours each spread over a two-month period. At the same time, interviews were conducted with members of the boy's family, his teachers, and certain of his schoolmates whom he said he knew best or who he thought knew him best. The idea was to come as close as possible to a fully rounded and credible picture of each boy, as seen by himself and others, while he was coping with his education and his world. As a point of strategy, throughout some 300 interviews with the boy himself and his "focal cluster," the race and sex of the interviewer were always matched with the race and sex of the respondent. This arrangement not only was presumed to yield a freer flow of dialogue in each case but also tended to provide a kind of validity check on the credibility of the data about each boy. That is, it was possible to observe the degree to which the several people in a given focal cluster (including the boy himself) converged in their perceptions of the boy's attributes.

Another difference is that the Pathways Project settled for an elaborate set of interview schedules of the open-ended variety to ensure that the areas of the boy's life and relationships with others would not be left wholly to random retrospection. The result was a set of upwards of 200 questions, many accompanied by a string of subquestions and suggested probes. Categories covered included such matters as the boy's health, life in his family, in his school, in his work, relations with the white world, troubles, disappointments, aspirations. One particularly evocative question addressed to the boy himself suggests something of how the "openness" of the interview was maintained despite the heavy load of instructions with which the interviewer had to contend:

> Let's pretend you wanted to disappear from the scene for awhile, but you had to get someone to take your place so that no one would know you were gone. You have to teach him, like a spy, how to act like you so that no one would know the difference. How would you tell him to act around home? With your friends? At school? (etc.) [p. 10]

The enormous scope of the response data obtained from only the initial interview with the boy himself is suggested by the fact that the coding of the responses involved 843 items or variables to be identified by the reader of the transcript. Many of these were subsequently merged in a number of ways in an attempt to define fewer attributes along which the respondents could be measured. One of these was a categorical scale labeled "strategic style" which purported to summarize for each individual the manifold ways he typically confronted his world at school, at home, and elsewhere. The strategic style dimension consisted of five categories labeled and ordered as follows: *withdrawn, conformist, cool guy, smart guy, tough guy.* This type of effort had some success. The fact that many of the measures so constructed were found to be correlated to some extent with one another and with outside variables in expected directions tended to support the hypothesis that many of the measures were not without a degree of construct validity. The strategic style measure, for example, was expected to be related to the boy's tendency to drop out of school early. The data tended to support the expectation thus:

Table 2
Relation between Strategic Style and Tendency to Drop Out of School

Drop-out behavior	*Strategic style*	
	Withdrawn, conformist, or cool guy (N = 38)	Smart guy or tough guy (N = 22)
Dropped out	24%	73%
Stayed in	76%	27%

(Table adapted from Rosenthal et al., undated, Table 10.22, p. 217)

By the time the third set of interviews with the 15 locatable young men had been completed and transcribed, it had become quite clear to the investigators that any further effort to form measurable variables out of the response data from all sources (which now filled 19 filing cabinets) would yield diminishing returns as a basis for representing the innumerable dimensions descriptive of educational experience and growth in the ghetto. Accordingly, they decided to capitalize on their extensive work with the data and summarize them in case studies of six of the young men who they believed would demonstrate the complexity and diversity of the entire group. One might think of this decision as a retreat from measurement. Not so. The initial effort to define as rigorously as possible a set of variables which might encompass the many facets of the response data paid off in two ways. It helped the investigators to identify the cases to be written up—cases that would by comparison with one another be most likely to comprise the full diversity of the black educational experience and the profound differences among the individuals involved. Further, it sensitized the authors to the risks as well as the virtues of the case study approach. They say:

> We are reflecting an experience with some three or four hundred people whose lives touched because they were all involved with someone who attended the George F. Ryan Junior High School [a pseudonym] in the late sixties. How representative are the people whose words and experiences we have recorded? *We cannot answer this with confidence, and would rather err on the side of conservatism.* Black people for too long have been lumped together in facile and erroneous generalizations. But it would be ingenuous to claim that we do not believe that the facts of black life and death in Roxbury are similar to those of life and death in other northern ghettos. [p. 14, emphasis added]

The content of the interview in this study was entirely different from that used in the Pathways Project. Its form, however, though much shorter, was roughly the same. It consisted of 54 open-ended questions with suggested probes and organized into 10 general categories having to do with aspects of the teaching-learning situation as experienced and viewed by elementary school teachers engaged in some form of open education. One set of questions, for instance, asked about the physical setting and materials of the classroom; another set asked about the children's activities; still another asked about the impact of school policies on the teaching-learning process. The overall aim was to see how each teacher viewed her or his job by having him or her retrospect about specific events.

After two years of study of the response material generated by interviews with 60 teachers at a number of different locations, the authors were able to put together several coding schemes for ordering the data. One such scheme made up of 17 items called "curriculum priorities," was used to compare the teachers with respect to various concerns having to do with their job. The nature of these curriculum priorities is best given by excerpts. One item of the code, labeled *Reflectivity and Intention*, is described thus:

> A. Concern that children know "what they are about" and "why." Concern that children think through what they are doing, understand (in their own terms) what they are doing. . . interject their own purposes into an activity. [p. 191]

Another labeled *Personal/Social Responsibility:*

> B. Concern that children mature in direction of basic cultural expectations—take care of own needs and belongings, respect the property of others, learn to take turns, share, etc. This is a concern for basic socialization of the child. [p. 194]

Another labeled *Grade-Level Facts and Skills:*

> C. Concern that children learn and be able to demonstrate knowledge of the required skills and basic facts expected of them at their particular grade level. [p. 193]

Having coded the response material in this fashion, the authors found that the curriculum priorities could be grouped in a hierarchical order to form an ordinal scale of sorts on which a teacher's understanding of curriculum could be located. On one end of the scale is a group of priorities that the authors regard as "narrow" (such as item C above), at the other end are priorities regarded as "comprehensive" (item A above), and in between are those regarded as "middle range" (item B above). Cutting across the narrow-to-comprehensive dimension was a grouping which distinguished between "cognitive priorities" and "personal/social priorities" (p. 42). Priority A above is an example of a cognitive priority which is also comprehensive. An example of a personal/social priority which is also narrow is labeled *Good School Behavior/Docility* and is described thus:

> Concern that children conform to a stereotypical pattern of school behavior. . . emphasis on politeness, working hard, settling down, not causing disruptions, etc. This is a concern for socialization into an adult stereotype, with little regard for the nature of the children's internal experience. [p. 194]

For all respondents, the code indicated that a given teacher might have several curriculum priorities, but that in each case, one or two concerns tended to be dominant. By combining this information with "evidence that a teacher was experimenting with the surface curriculum in ways intended to be responsive to the interests of individual children," (p. 56), it was found that the 60 teachers could be classified in four distinct groups as follows:

Group 1. (12%) "Grade-level facts and skills" is clearly the dominant priority, and there is little evidence of experimentation or change in the surface curriculum from what the teachers had been practicing previously.

Group 2. (22%) "Grade-level facts and skills" is clearly the dominant priority, but there is much evidence of change and experimentation with the surface curriculum.

Group 3. (39%) "Grade-level facts and skills" is an expressed priority, but not the dominant priority. Middle-range priorities tend to be dominant, and there is evidence of a potentially rich surface curriculum.

Group 4. (27%) A comprehensive or middle-range priority is dominant, and there is little evidence of preoccupation with "grade-level facts and skills"—i.e., it is not codable as such. There is also a potentially rich surface curriculum. [p. 56]

This kind of ordinal categorization suggests that the authors, who conceive of themselves as working in what they call the "neo-phenomenological tradition in psychology" (a tradition that is often seen as eschewing "measurement" in any form), are nevertheless prone to organize their highly complex response data along lines that do indeed conform to some of the basic notions of measurement. As is the case with the two preceding studies, the measures are not easily evolved, and the authors are careful to point out that "[t]his and other methodologies need to be refined for sustained and programmatic research on the origins, nature, and influences of teachers' thinking" (p. 171). It seems not unlikely that the needed refinements in research methodology will also include refinements in measurement methodology looking toward more explicit assessment of the validity and reliability of the multiplicity of attributes that any well-conducted open interview may bring to the surface.

A WORD ON PRACTICALITY

Of the Open Interview

After examining the three studies of the open interview that we have briefly sketched above, the reader is apt to have doubts about its practicality as a measuring device in education. From the standpoint of those whose work is with the day-to-day operations of the educational enterprise, such doubts would hardly be surprising. The open interview as here described has been strictly a research tool—one whose validity resides primarily in its power to *discover* the human and institutional attributes that may inhere in the schools and the people in them. It has involved the collection of huge amounts of data from small samples of respondents and has required months and years of work by research teams to code and shape the data into measurable attributes. All this effort has not been without a good deal of pay-off in uncovering dimensions of human functioning and educational process not capturable by the standardized interview or indeed by any other methods of measurement. But the administrator or research director of an educational system may well ask: "How can I conceivably make use of the open interview in sizing up our day-to-day operations or in making decisions about people and programs in the system?"

The question does not have an easy answer. It gets to the heart of the perennial problem of the tenuous connection between educational research and educational practice. Nevertheless, studies employing the open interview well and carefully can, if disseminated, have at least two consequences that one might call "practical": 1) They can remind educational practitioners of the many dimensions of the educational enterprise that are lurking below the surface of day-to-day operations; 2) they can open the way to the development of more readily applicable procedures for measuring those dimensions and assessing their validity and reliability.

Of the Standardized Interview

Concerns about the practicality of the standardized interview are of a different order but just as real. In this case, the educational practitioner may wonder about how the cost-benefit equation works out when one compares the interview with other techniques of measurement that look much like it. Clearly, assessing the attributes of individuals by means of the standardized diagnostic interview is a lot more expensive on a per-capita basis than assessing the same attributes by means of standardized paper-and-pencil tests administered to individuals in groups of 40 or more persons. Similarly, the survey interview with its team of paid interviewers going from house to house and spending upwards of half an hour with each respondent costs much more per capita than the mailed questionnaire, which purports to transmit the same type of information at the price of a few postage stamps.

The question, of course, is whether the extra overall expense can be justified on the ground that the data obtainable from standardized interviews are sufficiently superior in terms of measurement quality to the data obtainable from the competing standardized substitutes. A good case can be made on *a priori* grounds that the interview data can indeed be superior inasmuch as the transactions by which they are produced can be more closely observed and controlled.

This is to suggest that the various threats to validity we have noted in connection with standardized interviews may be just as severe in the standardized substitutes, possibly more so. In the latter case, however, they are more likely to go unnoticed and are, therefore, less likely to be guarded against. If one could somehow factor such threats into the cost side of the cost-benefit equation, one might be able to obtain a somewhat clearer idea of the relative practicality of the different modes of measurement or perhaps of some combination of them. The solution to this problem might perhaps be hastened if we were to conceive of *practicality* as an attribute of the various measuring devices in education—that is, as an attribute which, like any other, would be most usefully defined by the operations with which we agree to measure it.

REFERENCES

1. Bridgeman, P.W. *The logic of modern physics.* New York: Macmillan, 1948. P. 33.
   Views on measurement and other matters by the scientist who introduced the term *operational definition* to the English language.

2. Bussis, A. M. et al. *Beyond surface curriculum: An interview study of teachers' understandings.* Boulder, Colo.: Westview Press, 1976.
   A phenomenological approach via interviews to the ways teachers in open education perceive the concept of *curriculum* and the roles they and their students play in it.

3. Cohen, S. R. An exploratory study of student attitudes in the primary grades. In ETS, *A Plan for Evaluating the Quality of Educational Programs in Pennsylvania,* Vol. II. Harrisburg, Pa.: Pennsylvania State Board of Education, 1965 (mimeo). Pp. 61-130.
   Use of interview employing a projective technique with young children to infer their attitudes to the classroom experience.

4. Collins, A. *The interview: An educational research tool.* Palo Alto, Calif.: Institute for Communications Research, ERIC Clearinghouse on Educational Media and Technology, 1970.
   Overview of standardized and unstandardized interviews, the role of the interviewer, pitfalls in interviewing.

5. Dyer, H. S. The usability of the concept of *prejudice. Psychometrika,* 1945 *10,* 3. 219-224.
   Investigates how closely a mixed group of independent judges agree in recognizing prejudice in the responses of 9th-grade children.

6. Educational Testing Service et al. *State educational assessment programs.* Princeton, N.J.: ETS, 1971.
   Use of unstructured interviewing in nationwide exploration of the who-why-what-when-how of assessment programs.

7. Educational Testing Service et al. *State educational assessment programs: 1973 revision.* Princeton, N.J.: ETS, 1973.
   Use of structured interviews based on preceding survey.

8. Gorden, R.L. *Interviewing: strategy, techniques, and tactics* (Revised Edition). Homewood, Ill.: The Dorsey Press, 1975.
   Comprehensive textbook on interviewing, contains laboratory problems for the learner of the art.

9. Hyman H. H. et al. *Interviewing in social research.* Chicago: The Univer. of Chicago Press, 1954.
   Classic review of research on the problems of interviewing, particularly as found in the work of the National Opinion Research Center. An appendix describes how NORC organizes its surveys, trains and supervises interviewers, etc.

10. Jones, L. V. The nature of measurement. In Robert L. Thorndike (Ed.), *Educational measurement* (Second Edition). Washington, D.C.: American Council on Education, 1971. Pp. 335-355.
    Succinct account of the basic ideas and principles of measurement.

11. Merton, R. K. et al. *The focused interview: A manual of problems and procedures.* Glencoe, Ill.: The Free Press, 1956.
    Describes and illustrates strategies and tactics for inducing "retrospection."

12. Opinion Research Corporation. *A statewide view of our public schools.* Princeton, N.J.: Opinion Research Corporation, 1972.
    Brief summary of extensive unpublished survey of how various people in New Jersey feel about their schools.

13. Perry, W. G., Jr. *Forms of intellectual and ethical development in the college years: A scheme.* New York: Holt, Rinehart, and Winston, 1970.
    Shows how to convert longitudinal data from wide-open interviews with college students into a measuring device.

14. Rosenthal, R. A. et al. *Different strokes: Pathways to maturity in the Boston Ghetto.* Boulder, Colo.: Westover Press, 1976.
    Longitudinal study illustrating how free-response data can be converted into case studies that help to define the dimensions of the black educational experience.

15. Rosenthal, R. A. et al. *Pathways to identity: Aspects of the experience, self-concept, and racial identity of black youth from poor families: Selected findings from time I interviews.* Cambridge, Mass.: Pathways Project (undated, unpublished, mimeo).
    Predecessor of preceding reference. Illustrates herculean efforts to convert massive amounts of free-response material into measurable attributes.

16. Thomas, W. L. *The Thomas Self-Concept Values Test* (Manual). Chicago: Achievement Motivation Program, 1972.
    A standardized diagnostic interview that purports to measure how a child views himself and how he thinks others view him.

17. Wax, R. H. *Doing fieldwork: Warnings and advice.* Chicago: Univer. of Chicago Press, 1971.
    Shows how social anthropologists conduct their fieldwork and discusses the methods and problems of eliciting credible data from informants.

18. Wechsler, D. *Manual: Wechsler Intelligence Scale for Children-Revised* (WISC-R). New York: The Psychological Corporation, 1974. Pp. 53, 55.
    Paradigm of the standardized diagnostic interview as a measuring device.

19. Weinburg, E. *Community surveys with local talent: A handbook.* Chicago: National Opinion Research Center, The Univer. of Chicago, 1971.
Details of NORC's "experience in training the poor to interview the poor" in a series of health surveys. With minor modifications, useful training manual for interviewers in education.

20. Weiss, C. H. Interviewing in evaluation research. In Elmer L. Struening & Marcia Guttentag (Eds.), *Handbook of evaluation research,* Vol. I. Beverly Hills, Calif.: Sage Publications, 1975. Pp. 355-395.
Reviews recent research on the problems of interviewing.

ADDITIONAL REFERENCES

The additional items below are included for the reader who wishes to become more familiar with the uses, techniques, and problems associated with interviewing in general.

Biddle, B. J. et al. *Orientation, methods, and material studies in the role of the public school teacher,* Vol. 1. Columbia, Mo.: Univer. of Missouri, 1961.
Extensive use of interviews with teachers, parents, pupils, principals, and others to identify respondents' perceptions of the teacher's role.

Dunn, J. E. & Abrahams, N. M. *Use of biographical and interview information in predicting Naval Academy disenrollment.* San Diego, Calif.: Naval Personnel and Training Research Laboratory, 1970.
Compares predictive validity of the interview with that of other methods used in selecting Naval Academy candidates.

Educational Testing Service. The interview as an evaluation technique. In *Proceedings of the 1953 invitational conference on testing problems.* Princeton, N.J.: ETS, 1954. Pp. 116-136.
Three psychologists discuss the possibilities and shortcomings of the interview as a selection technique.

Miller, D. M. et al. *Elementary school teachers' viewpoints of classroom teaching and learning.* Madison, Wisc.: Univer. of Wisconsin Instructional Research Laboratory, 1967.
Chapters 9 and 10 illustrate ways of applying sophisticated statistical techniques to response data obtained from interviews with teachers.

Yarrow, L. J. Interviewing children. In P. H. Mussen (Ed.) *Handbook of research methods in child development.* New York: Wiley, 1960. Pp. 561-602.

Zeisel, H. *Say it with figures* (Fifth Edition, Revised). New York: Harper & Row, 1968.
How to put numbers on raw data en route to measurement that tells the story. Many examples.

70774 • Y4/P1.5 • 275328 • Printed in U.S.A.