DOCUMENT RESUME

ED 103 493	95	TM 004 514

AUTHOR	Donlon, Thomas F.
TITLE	Testing in the Affective Domain. ERIC/TM Report No. 41.
INSTITUTION	ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C. Office of Dissemination and Resources.
REPORT NO	TM-41
PUB DATE	Dec 74
CONTRACT	OEC-0-70-3797(519)
NOTE	32p.
EDRS PRICE	MF-$0.76 HC-$1.95 PLUS POSTAGE
DESCRIPTORS	*Affective Behavior; *Affective Tests; Attitude Tests; Elementary Secondary Education; Emotional Development; Evaluation Needs; Forced Choice Technique; Interviews; Measurement Goals; *Measurement Techniques; Q Sort; Semantic Differential; Student Attitudes; Student Interests; Student Opinion; Test Construction; Testing; *Testing Problems; *Test Validity; Values
IDENTIFIERS	*Affective Domain; Guttman Scales; Likert Scales; Thurstone Scales

ABSTRACT

Recognizing that the emotional state of the student is integral to his ability to learn, educators now place emphasis on testing in the affective domain. With this increasing demand for test data, ethical considerations must be taken into account as measurement instruments are designed, administered, and interpreted. Difficulties in instrument design arise because of the complex and multidimensional nature of the affective domain. To date, the most useful method of categorizing the emotional state is through an assessment of student attitudes, interests, values, and appreciations. The most commonly used assessment technique is the self-report stimulus response selection approach which may involve a format that is forced choice or true-false. Scales include Guttman, Likert, Thurstone, and the semantic differential. Of the numerous types of item formats and scales, all have complex problems ranging from serious validity problems to high costs. Other methods for assessing the affective domain are the Q-Sort, interviews, and unobtrusive measures. (BJG)

Testing in the Affective Domain[1]

by

Thomas F. Donlon

ERIC/TM Report 41

December 1974

[1] The report upon which this article is based was funded by the ERIC Clearinghouse on Tests, Measurement and Evaluation. The Clearinghouse operates under contract with the National Institute of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Points of view or opinions do not therefore represent official National Institute of Education position or policy.

TM Report 41 December 1974

TESTING IN THE AFFECTIVE DOMAIN

Thomas F. Donlon

In 1951, E. F. Lindquist, writing in the first edition of *Educational Measurement*, focused on the need for tests of hitherto unmeasured educational objectives. "If the descriptions of educational development of individual students provided by tests are to be truly comprehensive," he wrote, "tests and measuring devices must be developed for many more educational objectives than are now being measured at all. In general, satisfactory tests have thus far been developed only for objectives concerned with the student's *intellectual* development, or with his purely *rational* behavior. Objectives concerned with . . . moral values, attitudes toward social institutions and practices, . . . have been seriously neglected in educational measurement" (Lindquist, 1951).

By 1971, when the second edition of *Educational Measurement* was prepared, Krathwohl and Payne (1971) could describe the work on the *taxonomy of educational objects*: *II The affective domain*, as evidence of progress and of the increased importance of affective goals in education. The Taxonomy, an ambitious attempt to structure levels of affective response, indicates the validity of Carmen Finley's observation: "In recent years there has been a growing awareness of the need for schools to include the affective domain in the development of objectives for learning" (Finley, 1973). Or, as Robert Strom and E. Paul Torrance have observed, "A decade ago there was less discussion among educators about the affective domain than there is today . . . [there is] an emerging priority for emotional achievement" (Strom & Torrance, 1973).

The reasons for this expansion of interest are diverse and numerous. Attempts to evaluate schools and their functioning, increased efforts at accountability, shifts in the responsibilities of schools and families are all factors in the process. A number of major social problems, such as drug abuse or the assimilation of minorities, are seen as challenges to affective education. Further, affective characteristics are seen not only as the end-products of education, but as process characteristics: Too many learning problems are traceable to problems of motivation and the self-concept, and the schools must confront these dimensions of their pupils.

The expanded emphasis on the affective domain inevitably brings renewed interest in the techniques for instruction and measurement in this area. These techniques, however, are not nearly as well developed as they are for the cognitive achievement areas. As Lindquist (1951) observed, " . . . attainment of these objectives is . . . difficult to measure, . . . so little is known about how to measure them, just as so little is known about how to teach them effectively." The problems Lindquist perceived are far from solved today. Nonetheless, there are a number of techniques available, useful in the assessment of characteristics such as interests, attitudes, and values. While all of these approaches are somewhat crude, and while all are vulnerable to distortions of inference, they constitute a valuable resource for educators who establish affective objectives and who seek to measure the attainment of them.

This paper is a brief statement of the major approaches to testing in the affective domain. The emphasis throughout is on paper-and-pencil approaches, and on objective strategies. An effort is made to characterize observational techniques and projective tests, but the major share of the discussion and information is devoted to paper-and-pencil approaches, in the belief that these have proven over the years to be the most practical methods for educational assessment in the affective domain.

GENERAL CONSIDERATIONS

Terminology

There is a large variety of concepts and labels in the affective domain; drawing precise verbal distinctions among them is simply not possible. Interests, attitudes, values, and appreciations have been suggested by Tyler (1973) as the main areas of the affective domain which are of interest to educators. "Personality test" is another term that is widely used and troublesomely ambiguous. In general, measurement specialists distinguish it from a test of attitudes or interests and reserve it for tests designed to measure persistent and emotional characteristics of mental functioning, such as introversion-extroversion, or aggressivity-docility. In this use, the scores on personality tests describe general and emotional qualities of the mind. Tests of interests, attitudes, and values, then, are often not called personality tests because they have a specific content component external to the person: The person is interested in something outside the self, a sport or a book, or the person has a negative attitude toward Indonesia. The distinction is logically not very clear, however, and in the Seventh Mental Measurements Yearbook (Buros, 1972), the category called "Character and Personality" includes the well-known Study of Values, which measures broadly general interests or values. Similarly, in Anastasi (1968), the discussions of interest and attitude measures are included in a section devoted to personality tests. At best, we can simply offer some crude definitions of terms and recognize that there is a great deal of overlap among them. This paper does this for Tyler's four categories. The definitions, however, are those of the present author.

Attitudes or opinions are personal judgments of the nature or value of something. As such, they are not facts, and they may be either broadly emotional ("The United States is the best country in the world.") or quasi-intellectual ("The United States should have the largest navy in the world."). Similarly, in education, student attitudes may be emotional ("I hate school.") or have a strong and specific intellectual component ("I feel seniors should have a place where they can go and smoke."). Because of their emotional and intellectual nature, attitudes are very difficult to define. Shaw and Wright (1967) provide an extensive discussion of these definitional problems, discussing such words as opinion, belief, and trait. Further, attitudes may be conscious and easily stated by the holder or virtually unrecognized and unverbalizable.

Interests are areas of experience about which a person wishes to undertake further learning or performing. In this sense, an interest in something is a positive attitude toward it; an interest, then, is a kind of attitude. "Tennis is fun" may be the attitude which underlies an interest in tennis. To an extent, interests are more intimately connected to the self-image than attitudes. That is, attitudes, particularly quasi-intellectual ones such as whether the United States should have the world's largest navy, may change over time as the person learns new facts. Interests shift also but they are probably more stable components of the person than most attitudes are. Interests in some ways arise from deeper psychological processes involving the establishment of the self and its fulfillment.

Values are very broad attitudes or interests. The Allport-Vernon-Lindzey Study of Values, for example, describes a person in terms of six broad areas as originally proposed by Spranger (1928). These are: theoretical, economic, aesthetic, social, political, and religious. They are perhaps best thought of as broad classes of attitudes or interests; they are considered to be dominant aspects of the personality, motivating drives, and fundamental governors of behavior.

An appreciation is an achieved perception of the value or nature of something. It is an attitude, a judgment, but it connotes a learned set of perceptions which precede the affective reaction. Like an interest, it is almost always conceived of as positive, although one can speak of an appreciation of the dangers of drugs or of reckless driving.

Attitudes, interests, values, and appreciation are probably the four main aspects of educational measurement in the affective domain. Opinions, as suggested above, can be thought of as a subclass of attitudes. While other words offer potential clarity in some contexts, these four labels are a workable and comprehensive base.

In some ways, the self-concept and self-related evaluations do not fit neatly into the framework. The self is a very central concept, close to the core of the person. It is, in a sense, a learned appreciation. Although there are definite attitudes toward the self, it is probably wise to recognize this area as distinct from other appreciations and attitudes. The techniques for gathering information about the self are not essentially different from the techniques for learning about other, internal characteristics, but the degree of revelation is different, and the development of instruments in this area poses special challenges.

Ethical Aspects

Measurement and instruction in the affective domain face some problems which the traditional cognitive and achievement areas do not confront as directly.

These relate to the rights of persons to develop in a manner determined by their natural characteristics, by the kinds of persons they are. The extent to which schools attempt to influence aggressivity, for example, has ethical aspects. If a given student seeks counseling and is supported in it by the parents, then a counselor might test for aggressivity, identify it as the troubling area of the person, and work to modify it. But schools cannot enter the affective domain as "engineers" seeking to create specific kinds of people who are valued by educational authorities.

This ethical conflict between the need to give students self-benefiting attitudes and the danger of unnecessarily imposing values on them has surfaced in a number of contexts. The role of schools in the acquisition or rejection of religious values is a good example. Do the schools have the right to foster positive attitudes toward religion by permitting basically respectful pageantry during religious holidays? Even if the problems of recognizing religious minorities are surmounted, the rights of others are a sensitive issue in an egalitarian society.

A less difficult area but one that is not without its problems has to do with the attempts to influence students' attitudes toward drugs. There are deeply held emotional values running through all areas of the affective domain, and the recognition of the diversity of these values is an essential element of successful programs. Operations in the affective domain, be they measurement or instruction, must be constantly reviewed for ethical considerations.

Relating to this is the question of cooperation in measurement in the affective domain. As difficult as it is for the measurement worker to surrender potential information, or to deal with self-selected subsets of his original, total group (because some elect not to respond to certain material), the best principle is one which clearly indicates to the respondent that the cooperation is optional, that there is a principle of privacy, and that no response is required if it will produce discomfort or conflict within the person. This is perhaps particularly needed in tests of self-concept.

Most persons experience little difficulty in communicating, particularly if the measurements are retained with a reasonable degree of confidentiality. Schools, however, should be careful to institute adequate review procedures for all assessments in the affective domain, so that the rights of individuals are preserved. Holman and Docter (1972) have a succinct discussion of these issues and offer some bibliographic references, of which one, Ruebhausen and Brim (1965), is devoted to legal issues.

TECHNIQUES FOR ASSESSMENT

Paper-and-Pencil Approaches

By far the most common approach to measuring affective characteristics is to offer the person some way of providing a self-report by choosing alternatives or endorsing responses in a printed form. In a measure of self-concept, for example, the statement "I am much less organized than the average person" might be provided, and the person asked to respond with a choice or endorsement of some kind. Broadly, then, this approach is a stimulus-response technique, in which the stimulus is some verbal input, and the response is the individual's endorsement or rejection of it. There is considerable variation in the formats for such self-report surveys, both in the presenting of the stimulus and in the eliciting of the response. For example, in responding to the statement above about degree of organization, persons could simply indicate "true" or "false," or they might be given an opportunity to select from a somewhat broader scale of alternatives:

I STRONGLY AGREE with this statement

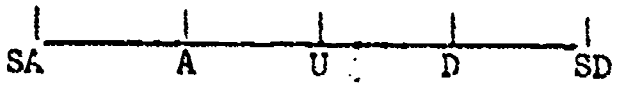
I AGREE with this statement

I am UNDECIDED about this statement

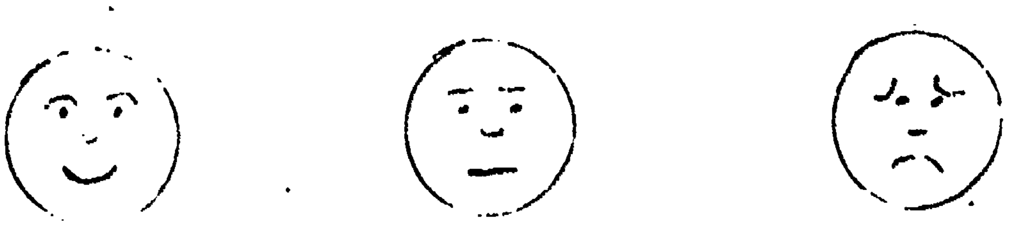
I DISAGREE with this statement

I STRONGLY DISAGREE with this statement

Frequently, after an introductory set of instructions, these possible responses are coded SA, A, U, D, and SD. The variety of formats is very great. A line can be drawn offering a kind of scale, and the individual can place a check mark along this line



Again, a variation on the offering of true-false endorsements of self-concept statements is simply to ask for a check mark on a check list of self-descriptive traits. On the other hand, not infrequently the response options are prepared on a separate answer sheet which can be scored by machine. In general, then, there are a large number of potentially workable formats and no overwhelming rationales for asserting the superiority of one to another: There has been fairly extensive empirical work on the relative merits of some of the different methods, as in the study by Jackson, Neill and Bevan (1973) comparing forced-choice and true-false formats; but in general an instrument developer can proceed to use practical judgment without fear that some technical rule will be violated. A practical and common sense adjustment of the general stimulus-response format to the needs and characteristics of the group being worked with is all that is needed. In adapting the methods to children, for example, such verbal categories as Agree-Undecided-Disagree can be replaced with the simple pictures



For children in the first four or five grades, responses through such a format may be more accurate and more highly motivated than responses through more abstract verbal endorsements. An approach such as this is part of the Minnesota School Affect Assessment (Ahlgren, Christensen, & Lun, 1973).

Where four or five choices are offered as optional responses, the method has similarities to the familiar multiple-choice tests which are widely used in cognitive tests. It differs, of course, because in the affective domain there is no "correct" answer, and because the optional responses tend to differ only in degree rather than in basic qualitative content as they do in cognitive tests. But there is in common a choosing among alternatives, a selection of options. In cognitive tests, giving each alternate wrong answer the proper qualities is demanding and skilled work. However, in most work in the affective domain there is little need for highly specialized skills in order to prepare an appropriate response. Nor is there a great need for special training in preparing stimuli. In order to assess attitude toward a school-expansion program, for example, simple statements along the following lines can be offered:

The proposed new school is too expensive	SA	A	U	D	SD
The proposed new school is too large	SA	A	U	D	SD
A swimming pool should be incorporated in the new school	SA	A	U	D	SD

The creation of such stimuli demands common sense, a knowledge of humanity, and an ability to create reasonably clear language, but no great technical expertise. This is not to say that there are not good and bad statements, or clear and unclear, and so on. But in much self-report work there is a straightforwardness of communication that places the creation of adequate stimuli well within the ability of a teacher or counselor.

A basic method, then, exists for affective measurement--a stimulus-response method which is relatively inexpensive to prepare, which requires no very formidable technical training, and which can be inexpensively scored in most cases. One might hope that Lindquist's pessimism cited earlier was premature. However, as they say in the jokes, that is the good news; now for the bad news.

The basic method of self-report by responses to statements is full of problems which complicate the interpretation of the results and which weaken the validity of the measures. For example, in considering interest assessment, Schwarz (1971) remarks:

> The problem in assessing interest has been that simply asking the individual about his interests in various curricula or occupations seldom results in the information desired. . . . The answers to direct questions necessarily are generalized responses based in part on erroneous or irrelevant impressions. . . .

That is, the individual's perceptions are influenced by her or his own personal experience. The stimulus, then, is always somewhat ambiguous. Do you like journalism? The meaning of a "Yes" or "No" response to such a question can seldom be clear, for there is an unwieldy breadth to the concept of "journa. .m." Similarly, attitude-assessment stimuli such as "The proposed new school is too expensive." will receive similar endorsements from quite different-minded people. One person may respond "Strongly Agree" because education *per se* is held in low regard, another simply because some component of the plans--a new gymnasium or a vocational shop--is considered a frill. Inferences about attitudes drawn on the basis of marks or responses are highly vulnerable.

The solutions to these problems are not simple. While stimuli should be as specific as possible, detailed breakdowns of stimuli can prove cumbersome. Analyzing journalism into free-lancing, sports reporting, editing, cartooning, opinion columns, and so on, can produce tedious decision-making that taxes the information base of the respondent. Inferences simply have to be made on practical grounds. There are, however, other problems with direct self-report approaches besides the inherent, logical problem of the verbal ambiguity of stimulus and response. The so-called response sets reflect the influence on the respondent of his or her awareness that the instrument is a communication about the self. A common response set growing out of this awareness is social desirability. First proposed by Edwards (1957), this is the tendency to "put up a good front," to distort personal choices in the direction of what is considered socially ideal. Thus, interests in higher paying or prestigious occupations may be expressed not because one is, in fact, attracted to medicine or the law but because one cannot admit in the context of the affective test that these really are not where the interests lie. Often, as Edwards pointed out, the individual is not conscious of her or his deception. We all like to perceive ourselves in the best way and we make the socially desirable response to please ourselves as much as others.

The techniques for combating some of these problems offer only a limited success. One of the major strategies for guarding against social desirability as a response set, for example, has been the forced-choice technique. In this technique, the stimuli are not presented alone but in groups, and the responses usually consist of identifying the extremes of the set. For example, in assessing interest in school subjects, one might create sets of three subjects and force the respondent to indicate a "most preferred" (M) and a "least preferred" (L) subject:

Physics	M	L
History	M	L
English	M	L
Physical Education	M	L
Woodworking	M	L
Home Economics	M	L

If the social desirability of the stimuli is determined beforehand, through judgments by raters, all of the stimuli in each set of three can have about the same social desirability. The choices, then, are believed to be more securely based on actual preferential feelings about the stimuli.

However, it has been demonstrated that sophisticated test takers can still distort responses even in the forced-choice approach, and, further, that the scores that are reached by adding up the results have a somewhat negative characteristic: The judgments were all relative rather than absolute; and so the results reflect more the rank order of the stimuli than their absolute level. Further, the scores commonly have a built-in influence on their intercorrelations, called "ipsativity," which makes them somewhat difficult to interpret in standard statistical analyses.

The forced-choice approach, then, is a logical response to certain problems of response sets, but it is so imperfect that, in balance, it would not commonly be recommended to nonprofessional test constructors. The work of assessing social desirability or other stimulus characteristics beforehand will not often seem to be worth the results.

Several types of stimulus-response scales are so well known as to require specific mention: The Likert, the Thurstone and the Guttman approaches. Both the Likert and the Thurstone approaches present stimuli singly rather than as forced choices. In the Thurstone approach, however, the stimulus is simply checked or endorsed as true of the respondent or not true. In the Likert approach, the response is given on a graded scale of (usually five) categories, such as Strongly Agree, Agree, Undecided, Disagree, and Strongly Disagree. These differences in response methods lead naturally to differences in scoring methods also. The Likert scale creates different weights for each possible response on its scale (say, 5 for Strongly Agree, 4 for Agree, and so on) and adds up a total score of all the weights for the responses selected. The Thurstone scale determines a unique weight for each stimulus statement by asking judges beforehand and then takes the median value of the weights of all the statements selected. Between them, the Thurstone and Likert scales account for the bulk of instrument development in education and psychology. Thurstone procedures require somewhat more elaborate preliminary development and statistical knowledge. In the long run, however, both are stimulus-response scales, differing more in the nature of the response and the numerical value attached to it than in anything else. A practical description of Thurstone procedures is offered in a paper by Murray (1971), which is available as an ERIC document.

A Guttman scale is another concept in this field. This approach assumes that an ideal scale will have the property that any individual who responds positively to a higher-ranking stimulus will also respond positively to a lower-ranking one. Let us suppose that 10 statements are prepared as stimuli. These are specifically selected to vary in the level or intensity of attitude they reflect, and the respondent is asked to indicate those which he or she can personally endorse. In theory, if one knows the highest level statement which is endorsed, one knows that a) no endorsements of higher level statements were given, and b) all lower-ranking statements were endorsed. For a number of reasons, people are seldom this consistent in responding to statements, and a Guttman scale is an ideal not often attained in practice. Like the Thurstone approach, it requires a considerable amount of rather complicated statistical work.

For most practical purposes, then, educators who need to develop affective scales can probably rely upon Likert scales for attitude assessment as the most convenient approach.

The measurement of interests is typically approached in a similar, stimulus-response way. The Strong Vocational Interest Blank (SVIB), for example, offers as stimuli the names of occupations. Responses are through a three-point scale of Like, Indifferent, or Dislike rather than a "standard" Likert five-point scale, but the approach has many basic similarities. The Strong instrument differs in that it derives its scores by a system of weights computed by preliminary sampling rather than assigning say, 3, 2, and 1 to its three points and then adding them all up. The Strong approach is sufficiently complex to require a special discussion, but in terms of the format for stimulus and response, the Strong Vocational Interest Blank resembles the use of

Likert or graded-response scales. The weights in the SVIB are derived by comparing the response of specific occupations with people in general. For example, an item might be evaluated as follows with respect to bakers:

Working with my hands

Responses	Like	Uninterested	Dislike
Responses of bakers	55 percent	35 percent	10 percent
Responses of people in general	30	35	55
Difference	+25	0	-25
Weight	+1	0	-1

This approach is consistently used throughout, with the difference in the percentages being used to assign weights. The total raw score is the sum of all the weights, positive or negative.

The Strong approach to weighting is interesting but demands large populations; the rationale essentially focuses on statistically significant differences among groups, and could not often be successfully used in developing measures for use in a given institution. Similarly, the Kuder scales for interest measurement, requiring the respondent to select the most attractive and the least attractive of a set of three activities, does not offer an easily reproducible technique for individual institutions.

A potentially useful technique for institutional researches is the Semantic Differential. It derives this high-sounding name from its origin as a research tool for psycholinguists (Osgood & Suci, 1955). Osgood and his associates were interested in problems of the meanings of words. They devised a format for securing judgments and feelings about words. For example, the word WOLF might be presented this way.

WOLF

kind : : : : : : : : cruel
big : : : : : : : : small
sweet: : : : : : : : sour

Each of the pairs of opposite words creates a kind of scale, ranging from one opposite pole to the other. Thus, there are intermediate points between "kind and cruel." The instrument designer offers a number of intervals between the poles as potential choices, and the respondent selects an interval on the scale. The use of seven intervals is a fairly common practice, although no specific number is mandatory. Each interval is assigned a weight which, for convenience, is a whole number; thus, if the respondent checks the interval nearest "kind," this might be scored as a "7" on that scale, with the interval nearest to "cruel" being scored 1. The individual's score is the sum of all these scale values.

The test constructor has to know certain things in order to develop a successful Semantic Differential and score it. The various scales have to be able to be added together if the total score is to have meaning. That is, they have to correlate, so that there is a tendency for those who think wolves are kind to think other positive thoughts about them. It is appropriate to find out which scales go together by doing a statistical analysis of the results, weeding out scales that don't contribute but deriving separate scores for those that offer independent information.

The use of the Semantic Differential in assessing attitudes in educational settings is exemplified by "Semantic Differential for Measuring Attitudes of Elementary School Children Toward Mathematics" (Scharf, 1971). Below is a sample of the stimuli and the response scales:

Taking a Math Test is:

	Very	Sort of	Neither	Sort of	Very	
BAD						GOOD
HAPPY						SAD

The same set of response scales may be used to assess additional dimensions of the subject. Thus, Scharf studied such other stimuli as "My Math Class is" and "Doing Math is."

A Semantic Differential is easy to construct, and most respondents find it intuitively easy to understand what is wanted. An interesting feature of this approach is that the respondents will often tolerate quite unusual scales, make meaningful responses, and the responses to these scales can offer useful information. This has to be checked by empirical methods, of course, but after a set to respond has been developed, one can ask where the concept FATHER stands on a scale from Valuable to Worthless and get a plausible response, even though it is rare to hear people say "My father is very valuable!" Similarly, in a Semantic Differential reflecting attitudes toward a home room, one could create the following scales:

My Home Room is

QUIET	________	________	________	________	________	NOISY
CROWDED	________	________	________	________	________	ROOMY
HOT	________	________	________	________	________	COLD
DUSTY	________	________	________	________	________	CLEAR
KIND	________	________	________	________	________	CRUEL

It is frequently possible, in the context of a number of judgments, to have scales such as KIND or CRUEL be meaningful to the respondents and to offer a sufficiently oblique avenue for response that some of the defensive response sets are avoided.

Because of its ease of construction and its acceptability to respondents, the Semantic Differential is a very useful technique. It is fairly widely used in educational research. Descriptions of techniques for constructing Semantic Differentials are found in Kerlinger (1967) and Maguire (1973).

In spite of the formal differences between them, Likert Scales and Semantic Differentials have a broad commonality as stimulus-response scales. Each calls for a response to a stimulus by selecting from a graded series of options, and it ought to be possible to secure somewhat the same results by adapting one technique to the other. For example:

My homeroom is quiet.	SA	A	U	D	SD
My homeroom is crowded.	SA	A	U	D	SD
My homeroom is hot.	SA	A	U	D	SD
My homeroom is dusty.	SA	A	U	D	SD
My homeroom is kind.	SA	A	U	D	SD

This Likert-type equivalent to the Semantic Differential given above ought to provide much the same information. .

It has been suggested that Semantic Differential scales be provided with adverbial descriptors, as follows

Weak : extremely : quite : slightly : slightly : quite : extremely : Strong

Thus, Wells and Smith (1968) found that there was greater differentiation and an avoidance of end points when the adverbial modifiers were included. Such additions underscore the similarity to the Likert approach.

There is a verbal efficiency to the Semantic Differential, however, that probably gives it an edge when the sought-for attitude can be captured in words or brief phrases which satisfy the requirements for a scale of opposites. With more complex concepts and opinions, such as "My homeroom is an excellent place to learn what's going on in school," Likert-type stimuli probably have the edge.

Likert scales or the Semantic Differential are workable techniques, but the interpretation of the scores they yield has been essentially normative. That is, if one creates a ten-statement Likert scale of attitudes toward mathematics, the best basis for evaluating it would seem to be normative by giving the scale to some students and studying the responses, letting statistical rarity guide the assessment of what is or is not important. Similarly, responses to a Semantic Differential of 10 scales would be handled in this way. The increasing attention to criterion-referenced measurement in the areas of skills and knowledges, however, has implications for affective measurement as well. A careful review of the instruments, considered in the light of the context in which they are administered and the decisions to which they should contribute, may suggest a critical level or levels, and a knowledge of such levels may help in the design or redesign of the instrument. Self-concept measures, for example, may be evaluated by predetermined evaluative criteria established by teachers and counselors. It is not easy to reach or defend such criterion levels, but it is probaoly an important safeguard against the passively accepted nonrational standards which can result from an overly timid reliance upon norms.

Similarly, the individual stimuli or statements in an affective instrument are often worthy of a careful review. A total score, with its abstract label, is more reliable and probably more valid than the individual components, but the content of the individual stimuli can often give insight as to where to go from here. Almost certainly, the individual stimuli themselves can be analyzed further. People who oppose the new gym and students who don't like math often have more to say on the subject. It is possible to do backup sampling or checking to bore further into the nature of the situation.

Measures of attitudes, values, and interests are often of greatest interest as descriptions of groups rather than individuals. A convenient way to display such information is to show the proportion of the group that selects one of the responses. This approach is appropriate for either a Likert-type scale or a Semantic Differential. This way of formulating results is often most interesting because of the contrasts it affords between subgroups with different abilities. The following example contrasts high school juniors and seniors with respect to attitudes toward dress code:

The dress code in our high school is too strict.

	SA	A	U	D	SD
Juniors	40	32	18	10	0
Seniors	28	36	10	16	10

Such contrasts of subgroups are often powerful contributors to an understanding of the social context within which attitudes operate. Further, they are often of greatest interest to the respondents themselves. Assessment in the affective domain is usually intrinsically interesting to the members of an institution, for it functions as a sort of mirror of the social context. Announcing the results of questionnaires and surveys, analyzed by subgroups with which people can identify--for example, administrators, faculty, students--not infrequently leads to the pinpointing of areas of difference which may be obstacles to communication. In a sense, affective results are somewhat freer of the ego threat that often lies in achievement scores; people will talk about them more.

From the foregoing review of major paper-and-pencil strategies, it should be clear that affective assessments depend in large measure on statistical operations. Establishing Thurstone scale values, determining Likert scale internal-consistency, and defining the clusters of semantic-differential scales all require some basic statistical operations. There is a danger in this need for analysis, however. The affective domain is extraordinary in its complexity and multidimensional nature. We do not know any grand design for the affective domain, and interests, values, appreciations can be organized and subdivided in a variety of ways. It is possible to literally explode the interest domain by factor-analytic methods, subdividing it into a larger and larger number of increasingly specific interests.

The moral in this is to resist the temptation to overmanipulate the data. Psychological constructs are typically fragile things, often depending on scientific populations and circumstances in order to demonstrate them. The institutionally based worker should keep in mind the decisions or needs which confront the institution and the logic of the data as they relate to these requirements. It does no good to offer a fifteen-factor analysis of rather tenuously labeled qualities such as "Attitude toward Science" or "Attitude toward Punctuality" if what is needed is some general assessment of the degree to which the students feel positively about the school. Noncognitive and affective assessment contains this pitfall, and instrument users and developers should be aware of it.

Other Approaches

Pencil-and-paper affective self-report instruments are the basic techniques for assessment, but there are a number of others worth mentioning. This paper will focus on three: Q-sorts, interviews, and unobtrusive measures.

Each of these is a somewhat more involved procedure than the paper-and-pencil approaches. Q-sorts and interviews tend to focus on one person at a time, and unobtrusive measures may demand rather elaborate recording devices. In the Q-sort technique, the experimenter asks the respondent to place a collection of stimuli in order from one end of a continuum to another. For example, the stimuli can be adjectives, and the respondent can rank them along a continuum from "most like me" to "least like me." The ranking is most often done by putting the stimuli into a distribution, and the distribution is prescribed in advance. Thus, if there are 10 adjectives, the respondent may be told to put them into 5 piles of 1, 2, 4, 2 and 1 adjectives each: Thus, respondents select the one adjective that is most like them, two more that are next most like them, four middling adjectives, and finally the next-to-least pile of two and the single "least like me" stimulus. There are disputes among Q-sorters about what kind of instructions to give the respondent concerning the number of piles and the number of stimuli in each, but whatever the approach, the method yields a sorting of the stimuli along a quantitative continuum, and hence its name.

The Q·sort method has its most interesting properties in the emphasis it places on individuals. In instructional evaluation, for example, a precourse Q-sort of attitude statement can be compared with a postcourse Q-sort, and the similarity between them assesses as a correlation. Similarly, it is common practice to analyze group Q-sorts so as to locate clusters of similar people rather than the more familiar clustering of stimuli into scales. People typically enjoy a Q-sort if there aren't too many statements; as more get added, or as the rules as to piles and the numbers in them get complex, it becomes a less attractive method.

Q-sorts are attractive in that they foster hypothesis testing. The experimenter can try to devise a set stimuli which he believes the subjects will sort in a predictable way, and then test this hypothesis. Attitudes toward mathematics, for example, may be hypothesized to be distributed in one way for successful students and in another for unsuccessful students.

Two contrasting Q-sorts of self-descriptive adjectives, one from a person who basically likes himself, one from a person who dislikes himself, might look like this:

	Sort 1	Sort 2
	Likes Himself	Dislikes Himself
Most true of me	Able	Unimportant
	Good Sorry	Passive Kind
	Energetic Kind Interesting Friendly	Unsociable Energetic Friendly Strong
	Passive Unsociable	Good Interesting
Least true of me	Unimportant	Able

It is possible to calculate correlations between such sorts, for they are basically elaborate rankings of the adjectives. The correlation between these two individuals would be highly negative. The study of similar correlations based on an individual over time is often useful as an index of personal stability or change.

The interview an expensive, time-consuming, and in some ways frustratingly unreliable, approach to affective assessment. It is extremely rich and full-dimensioned in the data it offers to the interviewer. It is superior to paper-and-pencil approaches in that it permits the correction of misunderstandings and the use of the interviewee's natural language. Its weaknesses lie in the pressure it puts on the respondents to put their best foot forward, to conceal the less "noble" aspects of self from the interviewer.

Interviews have to be planned, interviewers have to be trained. Matching interviewers to interviewees to reduce incompatability is a good practice. As an appraisal technique for determining attitudes and interests, the interview can produce a wealth of information of subjects, with questions devised on the spot by the interviewer after considering previous answers.

Cooperative subjects often volunteer a great deal of information which the experimenter failed to inquire about. To facilitate this effect, interviewers should provide the interviewee with as full an account of the purpose of the interview as can be given.

The greatest difficulties with the interview are the very size of the data it offers and the fact that you need to quantify this information in some way. It is, in a sense, a less-structured stimulus-response model in which there are obvious timuli (interviewer behaviors) and obvious responses (interviewee behaviors), and thus a chance to make inferences about affective characteristics, but the stimuli and the responses are so numerous and complex that it's difficult to know how to organize the information. Structuring, in the sense of predetermining most or all of the interviewer questions, helps this by controlling interviewer behavior, but there are still real difficulties in summarizing the information.

Nonetheless, interviewing is a sensible and valuable technique for affective assessment with many important by-products in terms of the human quality of the direct communication. Particularly where affective assessment is undertaken in order to determine the nature of an institutional environment, interviewing should be part of the overall assessment strategy.

Webb and his associates (1966) produced a book devoted to this topic, for which the best earlier statements were contained in Sellting et al. (1959). Essentially, this approach departs from the stimulus-response methods of the earlier techniques and seeks to make inferences about affective behavior by collecting data about everyday behavior. It is a challenge to the investigator and at the same time a corrective for some of the more indirect technqies. A common example of unobtrusive measures is how close people stand to each other when they talk, as a measure of their mutual acceptance of each other. Another example is the amount of audience coughing during theater performances, as a measure of interest in the play. Interest in museum displays is assessed by the wear and tear on the tile floors in front of the displays. Archives are searched for records of class attendance, book usage at a library, and so on, in an effort to draw inferences as to interest.

The method avoids some of the problems of alteration of response because of awareness of being assessed. But, because it is indirect and logical, it is open to errors of inference. Behaviors such as checking out library books are complexly determined and attributing a circulation increase to a poster campaign on reading may be entirely an error. Further, there are ethical aspects to the approach. Can you eavesdrop on student conversations? Can you check the wastebaskets after class? Watching behavior covertly may sound scientific but it can be dangerously close to snooping.

Nonetheless, the unobtrusive methods are to be recommended. They force investigators to think of the behavioral consequences of affective states, and in so doing may help them to devise real-world measures which are more intuitively satisfying than the results of paper and pencil surveys.

Projective techniques are widely used in the study of personality characteristics. The Thematic Apperception Test, in which the subject tells stories based on a series of drawings, is a good example. They may be extended to attitude measurement, however, although little formal work along these lines has been reported. Perhaps the most promising format for attitude assessment among the projective techniques is the sentence completion approach. Subjects are asked to supply completions for attitude-relevant sentences such as the following:

> The greatest social need of our time is . . .
>
> The greatest problem in dealing with minorities is . . .

The greatest difficulty with such approaches lies in their unstructured format. You can learn a great deal about attitudes, but you are at the mercy of the respondent, in some ways. Thus, responses to the "greatest social need of our time" will cover a gamut of concerns in which a given one, such as socialized medicine, may be very infrequently mentioned. The methods lend themselves more to exploring attitude domains, learning their likely boundaries, and are not really suitable for hypothesis testing.

Summary

The increasing interest in the affective domain in recent years has been met by a slow but steady expansion of technique and rationale in this area. Efforts have been made to formulate the definitions and to organize the logical structure that is essential for measurement. While much remains to be done, much has been accomplished. A number of specific strategies for assessment are available, ranging in complexity and rationale from paper-and-pencil scales to the unobtrusive methods and projective techniques.

These methods are seldom totally satisfying. While the development of instruments has a straightforward logic and requires little technical theory or body of knowledge, the gist of the methods is perceivable by the respondent, and the distortion of responses, either consciously or unconsciously, is the greatest single problem in working with them.

More so than in the area of cognitive achievements, there are ethical considerations to measurement in the affective domain. Attitudes, interests, values and appreciations are characterized by their affective component; the result is that communication about them is sometimes uncomfortable. Further, the establishing of objectives in this area is complicated by the problem of imposing values on others, of rewarding or recognizing certain types of persons at the expense of others. A maximum openness in the sharing of information helps to relieve this ethical tension, and often secures the kind of respondent cooperation which is desirable, considering the limitations of the techniques.

As important as the problems of measurement and assessment are, it is well to remember the cognate problems in the areas of instruction and curriculum. It is one thing to establish objectives in the affective domain; it is not so easy to institute sensible procedures for attaining them. Much progress in assessment will doubtless be made in the future, but it is likely that the greatest gains in the logical and ethical aspects of work in this domain will come through related gains in methods of instruction.

REFERENCES*

Ahlgren, A., Christerisen, D. J., & Lum, K. Manual for the Minnesota school affect assessment. Minneapolis, Minn.: Center for Educational Development, University of Minnesota, 1973.

Anastasi, A. Psychological testing. (3rd ed.) New York: The MacMillan Co., 1968.

Buros, O. K. (Ed.) The Seventh Mental Measurements Yearbook. Highland Park, New Jersey: The Gryphon Press, 1972.

Edwards, A. L. The social desirability variable in personality assessment and research. New York: Drydan, 1957.

Finley, C. About this report. Editorial comment in NCME Measurement in Education, 1973, Vol. 4, No. 3,

Holman, M. G., & Docter, R. F. Educational and psychological testing. New York: Russell Sage Foundation, 1972.

Jackson, D. N., Neill, J. A., & Bevan, A. R. An evaluation of forced-choice and true-false item formats in personality assessement. Journal of Research in Personality, 1973, 7, 21-30.

Kerlinger, F. Foundations of behavioral research. New York: Holt, Rinehart and Winston, 1967.

Krathwohl, D. R., & Payne, D. A. Defining and assessing educational objectives. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington, D. C.: American Council on Education, 1971.

*Items followed by an ED number (for example ED 06, 762) are available from the ERIC Document Reproducation Service (EDRS). Consult the most recent issue of Resources in Education for the address and ordering information.

Lindquist, E. F. Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), Educational measurement. (1st ed.) Washington, D. C.: American Council on Education, 1951.

Maguire, T. O. Semantic differential methodology for the structuring of attitudes. American Education Research Journal, 1973, 10, No. 4, 295-306.

Murray, N. B. Construction of a Thurstone attitude test. Paper presented at Annual Convention of the California Educational Research Association, 1971. ED 058 306.

Osgood, C., et al. The measurement of meaning. Urbana, Illinois: University of Illinois Press, 1957.

Ruebhausen, O. M., & Brim, O. G. Privacy and behavioral research. Columbia Law Review, 1965, 65, 1185-1211.

Scharf, E. S. The use of the semantic differential in measuring attitudes of elementary school children toward mathematics. School Science and Mathematics, 1971, 71, 641-649.

Schwarz, P. A. Prediction instruments for educational outcomes. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington, D. C.: American Council on Education, 1971.

Spranger, E. Type of men.

Strom, R. D., & Torrance, E. P. (Eds.) Education for affective achievement. New York: Rand McNally and Co., 1973.

Tyler, R. W. Assessing educational achievement in the affective domain. NCME Measurement in Education, 1973, Vol. 4, No. 3.

Webb, E. J., et al. Unobtrusive measures: nonreactive research in the social sciences. New York: Rand McNally and Co., 1966.

Wells, W. D., & Smith, G. Four semantic rating scales compared. Journal of Applied Psychology, 1960, 44, 393-397.