DOCUMENT RESUME

ED 229 285 SO 014 456

AUTHOR Oliver, Donald W.; Shaver, James P.
TITLE The Use of Content Analysis of Oral Discussion as a Method of Evaluating Political Education.
PUB DATE 15 Feb 63
NOTE 36p.; Paper presented at the Annual Meeting of the American Educational Research Association (February 15, 1963). For related documents, see ED 003 364-365. Paper excerpted from a more extensive report.
PUB TYPE Reports - Research/Technical (143) -- Guides - Classroom Use - Guides (For Teachers) (052) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Citizenship Education; *Content Analysis; Controversial Issues (Course Content); Critical Thinking; Discussion; Interaction Process Analysis; Problem Solving; Secondary Education; *Social Studies; *Speech Communication; Speech Skills; *Student Evaluation
IDENTIFIERS Oral Examinations

ABSTRACT

Reliability data suggest that, although there are many problems, it is feasible to systematically evaluate a student's analytic and persuasive competence in free oral argumentation. The first part of the paper describes the contexts within which the evaluation project took place. Specifically discussed are the five areas of analysis on which the project concentrated: (1) problem identification and differentiation, (2) making explicit cross problem assumptions, (3) identifying and using appropriate strategies for dealing with different types of problems, (4) identifying common dialectical operations, and (5) identifying relevance problems. The second part of the paper describes the content analysis system used to quantify student behavior. The categories or units used to describe the interactions are examined (many examples are provided) and the importance of the frame of reference of the person who does the categorization is emphasized. Two evaluation studies were conducted. In the first, four trained scorers scored from 10 to 18 discussions between a student and adult interviewer, in which a student was challenged to defend a position on a controversial case. In the second study, two trained scorers scored 32 pupil-led discussions. On the average, there was a relatively high level of agreement among scorers in both studies. (RM)

THE USE OF CONTENT ANALYSIS OF ORAL DISCUSSION AS A METHOD OF EVALUATING POLITICAL EDUCATION

Donald W. Oliver and James P. Shaver

(Session 55)

THE USE OF CONTENT ANALYSIS OF ORAL DISCUSSION AS A METHOD OF EVALUATING POLITICAL EDUCATION*

Donald W. Oliver and James P. Shaver

When one reviews tests of political sophistication, he usually turns to measures which presumably get at some kind of "critical thinking." Although these tests are, in some cases, ingenious, one cannot help feeling somewhat skeptical about their usefulness or validity. Their major weakness, probably, is the failure of their authors to state the criterion behavior with which the test scores presumably correlate as well as the criterion situations in which this behavior might occur.

In our own efforts to evaluate competence in analyzing political controversy, we have tried to be more specific and systematic in identifying criterion behavior, and to assess this behavior in more realistic settings. Although our efforts have been slow and halting, we do feel progress has been made. We therefore suggest below some of the steps necessary for such progress.

1. The nature of the problem to be analyzed by the student has to be more narrowly construed than has thus far ordinarily been the case. To develop a general model as an analytical basis for a wide diversity of problems is too ambitious. Surely the important distinctions between problems which are essentially ethical, scientific, or aesthetic, for example, should be taken into account. Our work has

---

*This paper is taken directly from a more extensive report to the U.S. Office of Education entitled: The Analysis of Political Controversy: An Approach to Citizenship Education based on Cooperative Research Project No. 551. This explains some stylistic anomalies, e.g., footnote numbers.

focused on the analysis of the ethical component (as opposed to the political process component) in political controversy.

2. The situational context for measuring analytical skills should be described more carefully and the testing situation broadened beyond the simple multiple choice pencil and paper test. A number of elements should be considered in describing the test situation: (1) the form of the message with which the student is required to deal; (2) the content of the message which is the subject of analysis; (3) the extent to which the student is required only to analyze an existing message in terms of analytic concepts taught as against using these analytic concepts to evaluate or construct counterarguments; (4) the extent to which choices are prestructured for the student as against requiring the student to structure his own answer; and (5) the extent to which oral interaction is an element in the evaluative situation. These five considerations suggest the following kinds of choices. (These choices are obviously not exhaustive.)

a. form of message:

one sided persuasive communication (e.g., newspaper editorial or advertisement)	v.	two-sided dialogue

b. content of message:

scientific decisions v. public policy decisions v. aesthetic decisions

c. level of analytic abstraction

responses required in test might consist of abstract analysis of controversial dialogue	v.	responses might consist of appropriate rebuttal statements

d. level of structure

responses required in test situation might be pre-structured and simply chosen by student	v.	test might require that student create the response

e. interpersonal context

student might be operating within a live interaction situation (e.g., an oral examination)	v. student might be dealing with a controversy only as a non-participant or observer

In our work in test development we made the following decisions:

(1) to use the dialogue as the basic form of message;

(2) to deal with public policy decisions as the content of the message;

(3) to develop measures concerned both with abstract analyses of the controversial dialogue as well as measures requiring the student to develop appropriate rebuttal statements constructed on his own initiative;

(4) to develop both prestructured tests and open-ended unstructured tests;

(5) to develop measures in which the student both analyzes the arguments of others and participates in the give-and-take of an argument itself.

These decisions resulted in four measures which are discussed in some detail below.

3. The analytic operations one wishes to evaluate must be described systematically and the student must be required to identify not only when the operation has been carried out successfully, but also when the operation is appropriate or relevant in a problem context. Systematic efforts have been made along these lines by several investigators including Ennis, Dressel and Mayhew, and in the PEA studies[47] discussed earlier in this chapter. The major problem

[47] Robert H. Ennis, "A Concept of Critical Thinking," pp. 81-111; L. Dressel and B. Mayhew, General Education: Explorations in Evaluation; Eugene B. Smith and Ralph W. Tyler, Appraising and Recording Student Progress.

with these efforts, we feel, is the isolation of the analytic operation from the total problem context. Although there is a general concern with problem analysis, all choose to measure fragments of the analytic process isolated from other components of the process. This often results in building systems of thought to think about thinking which are so unwieldly and complex or, at times, useless, that one scarcely has time to think about the problems they are meant to clarify.

Major intellectual operations assessed in Project measures. Our own approach has been much simpler than those used or suggested by Ennis or Smith and more "wholistic" and interrelated than the work of Dressel and Mayhew. We concentrated on five areas of analysis:

a. problem identification and differentiation;

b. making explicit important cross problem assumptions;

c. identifying and using appropriate strategies for dealing with different types of problems;

d. identifying common dialectical operations; and,

e. identifying relevance problems.

Each of these areas of analysis is related to the others, as should be clear from the subsequent discussion.

a. Problem identification and differentiation. We were interested in the extent to which the student could differentiate among various classes of problems within a controversial setting. For our purposes we differentiated between empirical controversy, value controversy, and definitional or analytic controversy. Evidence of the student's ability to make these distinctions can be obtained in at least two ways: (1) he correctly labels or describes the nature of the problem, or (2) he uses a correct strategy for dealing with the

controversy. For example, the student may say that a particular statement presents a problem over the meaning of a word, or it presents a problem concerning which of two conflicting factual claims is true. Or the student may say that we deal with this kind of problem by finding out how the word is commonly used in this context, while we deal with that kind of problem by looking for consistencies and inconsistencies in the reports of the event, investigating the credentials of the observers, etc. The use of an incorrect strategy can be illustrated by the following example:

> Two men are arguing over whether the absence of fire escapes from two-story buildings constitutes a menace to public health. One says it is a menace; the other says it is not. The statistics regarding deaths caused by the absence of such fire escapes are available, and are not brought into the debate.

In this kind of situation students commonly state that the way to settle the argument is to go out and observe buildings without fire escapes or observe the statistics and see whether or not such buildings constitute a menace. From our point of view, one must observe not only the statistics and the buildings, but how the word "menace" is commonly used.

b. Making explicit important cross-problem assumptions. The categorization of a particular problem as definitional, factual, or ethical is only the initial step in a complex process of justification. For every problem that is identified, one must make countless assumptions about agreement (or question whether or not there is agreement) regarding related problems inherent in the same issue. Suppose one says, for example, "The Federal Trade Commission is needlessly censoring television advertising." The antagonist says "no!" If an argument ensues, it can take the form of a definitional

disagreement (perhaps over the word "censorship"), a factual disagreement (over whether or not such censorship actually takes place), or a value disagreement (over whether or not the government should interfere with the free flow of information over the airways.) Each of these disagreements tends, however, to make assumptions about agreement in other areas, which may, in fact, be questioned. The value disagreement, in this case, for example, may assume that censorship is a real possibility or that it has actually taken place--it assumes a fact. The factual disagreement generally assumes that there are common values both with respect to the free flow of information and to government regulation or control in the interests of the welfare of the community; otherwise the fact would not be worth arguing about. The definitional argument may also assume that censorship is imminent or is taking place, and that censorship can be bad. Analyzing important empirical and definitional cross-problem assumptions is especially important in the process of argumentation, since one of the subtlest ways to press one's value commitments with respect to a given controversial situation is to distort or exaggerate the information describing the situation, or to choose vague but loaded words to describe it.

c. The process of identifying and using appropriate strategies for dealing with different types of problems. Both analytic processes discussed above involve essentially problem identification and differentiation. From observing a great many discussions, we noted that discussants tend to avoid dealing with the various aspects of an argument systematically and prefer rather to "go around in circles." The following example illustrates a common pattern of reasoning:

Southern schools shouldn't be segregated because

segregation of the races is bad, moreover,

the Negroes in the South are given poor treatment under

segregated conditions;

Analytically, the line of reasoning can be described as follows:

A policy decision is supported by a specific value judgment (which really adds little new information).

Both of these statements are followed by an unsupported and controversial generalization.

The case is then terminated with a restatement of the initial policy decision.

This pattern of thinking is usually an inefficient way to approach an issue. Disagreements over decisions, for example, may be clarified by construing the decision in terms of the general values of the culture, discovering analogous situations in which the same values are in conflict so that the individual can see the problem in a broader context, identifying differences among the analogous situations which cause one to change one's decision from one sitution to another, and testing whether or not these presumed differences do, in fact, hold up under careful scrutiny. If the differences do not hold up, the individual can change his decision and simply accept his own inconsistency. Disagreements over fact can be approached by constructing testable hypotheses and looking for evidence. Definitional disagreements can be approached by a process of categorical reasoning (Do the essential criteria that define a category apply to an object, event, or person we wish to include within the category), by a process of empirically testing how a word or label is commonly used, by pragmatically testing whether the particular use of a word will avoid ambiguity and confusion, or simply

by seeking stipulative agreement on the use of the word in this particular situation. The strategies suggested here are obviously not exhaustive. The point is that some strategies are appropriate for some kinds of problems and inappropriate for others. And more important, we assume that making strategies of problem resolution explicit tends to lead the discussant to greater focus and more systematic analysis than simply making casual and perhaps unrelated statements in an attempt to justify a decision.

d. Identifying and using common dialectical operations. In the process of analyzing political controversy we have noticed a pattern of thought which seems common to the three major types of disagreement. The essential operations in the pattern may be described as follows:

generalizing: assuming that what is good, true, or useful in specific instances is good, true, or useful in general.

specifying: supporting, contradicting, or simply elaborating a general statement by pointing to specific instances in which the general statement holds or does not hold. The concept "specifying" includes, for us, then, the operation pointing out the consistency or inconsistency of a specific statement with a more general statement.

qualifying: generalizing in a qualified way so as to take into account exceptional instances in which the general statement appears to be inconsistent with related facts, values, or definitions.

The following example will illustrate:

Statement (A) "Desegregation will improve education for the Negroes in the South."

(The individual making the statement is assumed to know of specific instances which make him think this generalization will hold, or to have heard the general statement from some authoritative source, and to believe, therefore, that it is generally true. This is an example of a general statement, although only implicitly does it illustrate the process of generalizing.)

Statement (B) "Sure, just like education for the Negro improved the first year when Little Rock's Central High School was integrated."

(This statement provides a specific example of a situation in which the initial generalization is presumed not to hold. It illustrates the process: specification-inconsistency.)

Statement (C) "When a careful plan of community education has been carried out, and the community is ready for it, integrated schools provide a better education for the Negro and white alike.

(This is a reworking of statement (A), with a qualification added which presumably takes into account the inconsistent example specified in statement (B).) The process of going from specific to general (generalizing), then returning again to specific test examples of the general statement (specifying), and then restating the general statement to take into account inconsistencies and exceptional cases applies to all three types of problems--value problems, factual problems, and definitional problems. These concepts give us, then, an abstract conceptualization of dialectical strategy which cuts across problem types. The process of

arriving at a qualified generalization, we think, is a very important aspect of the definition of reflective thinking.

e. The process of identifying relevance problems. Relevance can be viewed from two points of view: What statements are relevant to the argument as a whole; and what statements are relevant at a particular point in the discussion. Each discussion has a context set mainly by the topic. Some statements are clearly not relevant because, while they may be within the political-ethical frame, they are on a different topic, e.g., a statement about labor unions in a discussion of desegregation. Within the discussion itself some statements may be relevant to the discussion in general, but inappropriate to the immediate context. For example, the central problem of a discussion, at some point, may turn on what are in fact behavioral differences between whites and Negroes. Before this issue is in any sense settled, someone may move to the question of whether or not differences are culturally conditioned or genetically conditioned. While this issue may be relevant to the total discussion, it may be inappropriate at this particular point in the discussion, since it shortcircuits the immediate issue under analysis.

The Development of an Experimental Measure

We shall now describe the results of our own efforts to evaluate competence in the analysis of political controversy. We developed four instruments, which actually constitute four strategies for getting at the same type of competence, but emphasizing different elements in the testing situation. All four tests are labeled the Social Issues Analysis Test (SIAT), with numbers designating different measurement procedures. Below we shall describe SIAT No. 4, the last measure developed.

SIAT No. 4, A System for Analyzing and Evaluating Free Discussion

From our point of view, the most natural situation within which to place the student in order to evaluate his analytic and persuasive competence is one involving free oral argumentation. This kind of argumentation must, however, be evaluated by general subjective ratings or by systematic content analysis of the interaction process. We have chosen to explore the latter approach in our own work. Following a description of the types of discussion setting in which we obtained student behavior, we shall present in some detail a description of the content analytical system developed to assess that behavior.

Types of Discussion Situations Subjected to Content Analysis

The system for content analysis discussed below is used to score tape recorded rather than live discussions, although with some simplification it could be used for analysis of live situations. It has been used to score three different types of discussion, all based on a controversial case: interviewer-student discussions; discussions composed only of students; and teacher-led instructional discussions. Both the interview situation and student-led discussions were used for evaluation. Teacher-led instructional discussions were taped to check the long term consistency with which teachers can play different teaching roles, a problem which will be discussed in Chapter Twelve. Here we will be concerned with the system as it was used for evaluation.

Student-led discussions. Our initial thoughts about what would be the most appropriate test situation in which to evaluate the student's ability to analyze a controversial case brought us to the student-led

discussion. This decision was based primarily on the assumption that the intellectual leadership of the adult teacher and the effects of approving or disapproving teacher cues would be minimized in a discussion group composed entirely of students. In such a situation approximately twelve students are seated in a circle or semi-circle. Each student is given a copy of the controversial case; the case is read, and the students are then asked to arrive at a consensus regarding what is to be done regarding the problem in the case within a stated period of time. Consensus is an important requirement of the task to prevent the majority rule. (This procedure was adopted from Bales' work with five-man groups.) In general, the discussion that ensues is witnessed by no one except members of the group. The teacher or experimenter reads the initial directions and leaves the room. He enters again only at the end of the period allowed for the discussion to hear the decision at which the group has arrived. (The discussion is, of course, recorded.)

The Socratic interview. The student discussion group as an object of evaluation has certain obvious difficulties: (1) Different groups may have different degrees of interest in the task assigned and may experience procedural problems of different degrees of intensity; (2) if we choose to treat the groups as the unit of analysis, it is impossible to tell which and how any members of the group facilitate or inhibit the clarification or resolution of the problems under discussion; and (3) if we choose to evaluate individual responses in the group situation, we face the problem of equating the responses of high and low participants, or even ensuring that some students will respond at all.

These problems can be largely overcome by setting up a situation in which a single adult discusses a controversial case with an individual student. This is essentially a two-man group in which we attempt to control the level of sophistication of the person in the group who is not being evaluated. As we have used this situation each interviewer is given the freedom to pursue whatever issues the student chooses to raise. Each interviewer is, however, provided with a "brief" setting forth major issues and arguments as well as critical analogies with which to confront the student. On this and the next page is presented a case with which we have had considerable experience.

A LAW ON HOUSING RIGHTS

Many states have become concerned about the problem of protecting the rights of minority groups, especially the rights of Negroes. Below are excerpts from an imaginary law similar to laws recently proposed in some states. Following the description of the law are comments from two newspaper editorials, discussing the merits and shortcomings of the proposed law.

After we read this information, I want to discuss your opinion of the law, and how you might defend your opinion. I would encourage you to use any discussion skills or critical thinking skills which might help you think through this issue more intelligently.

* * *

A law has been proposed in the state legislature providing that "No person shall refuse to sell, rent, lease, or sublet a house or apartment to any person because of race or color."

The proposed law further states that it would be "enforced immediately, except when an owner or landlord can show that undue hardship will result from such immediate enforcement. In cases where undue hardship can be demonstrated, a reasonable delay may be granted."

An editorial in the "Southern Evening Gazette" commented on the proposed law as follows:

> This law is based on the false assumption that there is no difference between the two races. The fact is that police records show that, in relation to the population, more Negroes commit crimes of violence than do white people. We also know that broken homes are much more common among non-whites than among white people, and many more Negro children have parents who have neglected or deserted them. Especially important to the question of renting or selling houses is the fact that the yearly income of the average Negro family is much less than that of white families. Also, a larger percentage of Negroes is out of work each year. With these facts in mind, let us conclude by saying that no matter how much our "liberal" friends do not like to admit it, Negroes are different, and whites have the right to act on this basis. This proposed law takes away that right.

An editorial in the "Evening Sun" took another position:

> There are good reasons to support this law. A recent survey in a large city showed that many Negroes with good jobs and a good income want to move to the less crowded suburbs, but are refused housing by white

owners. We know of instances in which nationally known Negro artists and athletes have been denied housing in white areas. This is a national disgrace. Sociologists support the position that Negores often fail to live up to high moral standards because they accept the same opinion of themselves as whites show toward them. Thus, when whites continue to treat them as inferior human beings, Negroes accept the same image of themselves and fail to live up to their full potential. Furthermore, it is common knowledge that Negroes indulge in luxury items such as Cadillacs, fur coats, and expensive clothes, because they are unable to buy or rent decent housing which many want and can afford. We must conclude that this law is needed to right the wrongs committed against the colored race for the last 100 years.

* * *

DO YOU THINK THIS LAW SHOULD BE PASSED?

Having described the contexts within which evaluation has been attempted by the Project, we shall now go on to describe the content analysis system used to quantify the behavior elicited in these various situations.

Quantifying Selected Conceptual Operations Required to Clarify and Defend a Controvesial Position

The system we are about to describe is set up to identify a number of the major analytic operations or concepts described earlier in this

chapter. Those that are particularly appropriate for the content analytic system are summarized below:

a. identifying diffferent types of disagreements within a political controversy

b. using appropriate strategies for dealing with different types of disagreements

c. using appropriate dialectical operations to explore a disagreement

d. dealing with the problems of relevance

A general approach to content analysis.[48] Before actually getting into the substance of the system, we should make some general statements about our approach to content analysis as applied to oral discussion. While some of these statements may seem technical, they are presented for two reasons: They may be helpful to those who wish to experiment with this approach to the evaluation of the student-teacher dialogue, and they will reveal the many problems which must be considered if one wishes to develop such a system. Moreover, the handling of these problems forms the basis upon which one builds a set of assumptions underlying the evaluation of the instrument.

Systematic analysis of interaction, as we use the methodology, involves analyzing ongoing interaction into discrete units which are then categorized. There are three important considerations which must be taken into account in carrying out this process: Into what size units will the total train of interaction be broken? What is the

---

[48] Our orientation toward content analysis is clearly influenced by Robert F. Bales, Interaction Process Analysis (Cambridge, Mass.: Addison-Wesley Press, 1951). The instrument described here is similar to Bales' in the units of analysis and the observer's frame of reference.

frame of reference of the person who does the categorization? What is the specific nature of the categories used to describe the interaction?

Theoretically, the unit can range in size from an entire meeting or discussion to a particular segment of the discussion or meeting. This segment may be defined in terms of time, a completed verbal interchange, a bit of participation by an individual, or according to some linguistic convention. In general, the unit of our present system is defined by Bales:

> The unit to be scored is the smallest discriminable segment of verbal . . . behavior to which the observer, using the present set of categories after appropriate training, can assign a classification under conditions of continuous serial scoring. This unit may be called an act, or more properly, a single interaction, since all acts in the present scheme are regarded as interactions. The unit as defined here has also been called the single item of thought . . . . Often the unit will be a single sentence expressing or conveying a complete simple thought.[49]

We are interested, then, in classifying the "single item of thought." Examples of complete units would be:

> "I am sure that the Southerners would not accept immediate integration."
>
> "They are the ones who should do something about the situation."

In general, compound sentences are scored as two units; complex sentences as one. In some instances, several sentences may constitute one unit of thought, e.g., in presenting a single case situation or analogy.

---

[49] Ibid., p. 37.

Determining the observer's frame of reference poses a number of questions. One is deciding what the observer's point of view will be toward the group. The observer can:

> . . . . think of himself as a generalized group member, or, insofar as he can, as the specific other to whom the actor is talking, or toward whom the actor's behavior is directed, or by whom the actor's behavior is perceived. The observer then endeavors to classify the act of the actor according to its instrumental or expressive significance to the other group member.[50]

Another point of view is that described by Steinzor in which the purpose of the observer is to determine the intent of the actor.[51] A third point of view is that of the observer who is to be aloof from the process and not concerned with intent or the effects upon the group or its members. In the observational scheme of Heyns "the observer is outside the process and views each contribution in terms of its theoretical properties as a problem solving function."[52] Carter and his associates use a scheme in which the observer is not to be concerned with intent or effect, but with the functional significance of an act for the discussion situation through the eyes of an outsider.

In one of the schemes which we have used to quantify teacher style, socio-emotional categories requiring inferences about affective

---

[50] Ibid., p. 39.

[51] B. Steinzor, "The Development and Evaluation of a Measure of Social Interaction," Human Relations, II (1949).

[52] R. W. Heyns and R. Lippit, "Systematic Observation Techniques," Vol. 1 of Handbook of Social Psychology, ed. by Lindsay Gardner (Cambridge, Mass.: Addison-Wesley Press, 1954).

states of mind are included. In using these categories, the observer is to adopt Bales' position in regard to point of view. That is, he is to put himself in the position of the individual toward whom the act is directed and ask himself, "How would I perceive the actor's intentions if I were the recipient of that act?" or "What would that segment of behavior tell me about the actor's state of mind if I were the one toward whom it was directed?" However, in the present scheme the orientation is different. The purpose of observation is not to make inferences about affective states, but to look for cues which will indicate whether or not the actor is using desired categories of thought. The observer's point of view, therefore, is much like that of Heyns in that the observer serves as an expert in applying criteria of thought categories to the actor's statements. He is "outside the process," except as the discussion context is necessary to apply the criteria. For example, in deciding whether or not a student has stated a qualified value judgment or raised a question of relevance, the observer refers to the content of the statement, not to the manner in which other students might interpret it. As a matter of fact, our experiences confirm Bales' report that for scoring most acts the point of view of the observer is of small importance in obtaining interobserver reliability.

A second aspect of frame of reference is the extent to which the observer should take into account any prior knowledge he has of the group or of the individuals within the group. Our position is that it would be ideal if the observer scored each discussion with no prior knowledge of individuals or groups involved. We instruct the observer to try and forget all prior experiences with participants. If an interact arouses a prejudgment the observer is to control it and the

question becomes, "How would I score that act if I had never before heard this person?"

A third consideration in defining a frame of reference for the observer concerns the context of the discussion. How much of the context of a particular discussion should the observer take into account in classifying a particular act? Should the act be scored in isolation? Or should the act be scored in its relationship only to the previous act? For example, the statement, "I would agree with John" might be scored disagreement, if John's antagonist in the discussion had just spoken, but scored as "agreement" if the context of our scheme is the total discussion. Since our system uses two scoring systems superimposed on each other, it uses two contexts: one for what we call static categories; the other for dynamic categories. The dynamic system (see Table 10.4) consists of categories which explicitly require the scorer to deal with a context beyond the statement being categorized. (These are essentially dialectical operations.) This context may include one or several other sentences. Scoring in these categories is determined by relationships within or among statements. The static categories (see Table 10.5) theoretically can be scored without taking into account any context beyond the scorable unit. Every unit of behavior is scored in a static category. Dynamic operations are scored only when they are identified. Thus, when a dynamic operation is scored, a double categorization of the same unit occurs.

There are some exceptions, however, to the distinction between static and dynamic categories. The category "relevance," for example, is a dynamic category, but is scored as if it were static because the assertion or questioning of relevance usually contains an obvious cue

TABLE 10.4

DYNAMIC CATEGORIES

1. CONSISTENCY-INCONSISTENCY: Statements that indicate explicitly or implicitly that the speaker is aware of a real or possible consistency or inconsistency within his own or another speaker's position. The inconsistency may be between two values, two facts, or two definitions.

2. SPECIFICATION and GENERALIZATION: Specification occurs when the speaker gives a specific statement to illustrate or support a more general statement. Generalization occurs when the speaker draws a more general conclusion from one or more specific statements already given.

   Example of specification: "Desegregation is not going well. Only 7% of the Negro children in the South are now going to integrated schools after seven years of illegal segregation." The second sentence would be scored as the static operation "specific claim" and the dynamic operation "specification."

   Example of a generalization: "After World War II, Russia 'captured' the countries of eastern Europe, helped China to become a Communist nation, and tried its best to take over Greece and Turkey. Russia is the greatest empirialist nation the world has ever known." Statement two would be scored as a static operation "general claim," and as a dynamic operation "generalization."

3. QUALIFYING. A statement which deals with an implicit or explicit inconsistency by pointing out under what general circumstances an exception to a general principle is allowable or possible we score as a qualifying act.

   Example:
   Mr. A: Our civil liberties are our most precious asset. To try and restrict them for any citizen is unAmerican.
   Mr. B: If you had been in Germany in the early 1930's, would you have restricted some of the civil liberties granted Hitler when he was conducting mass hate meetings?
   Mr. A: I very well might have. I would say that civil liberties should be restricted, however, only when the government which is pledged to protect them is in real danger from an undemocratic and brutal force, which would destroy all civil liberties.

   Mr. A's modified position would be scored as static operation "general value judgment," and dynamic operation "qualification."

TABLE 10.5

STATIC CATEGORIES

GENERAL VALUE JUDGMENTS: Statements in which the speaker expresses a preference for a person, object or position in the argument in terms of a general social or legal value, such as: personal privacy, property, contract, speech, religion, general welfare of the groups, equality, justice, brotherhood, due process, consent and representation, etc. "Mr. Kohler certainly should have the right to use his property as he sees fit and to make contracts with his workers without union interference."

SPECIFIC VALUE JUDGMENTS: Statements in which the speaker expresses a preference for a person, object or position in the argument in terms of the specific case under discussion. "I think Mr. Kohler should have met the demands of the United Auto Workers."

GENERAL LEGAL CLAIM: Statements in which the speaker asserts that someone has a legal right to do something, expressed in terms of a general legal principle, such as: rule of law, due process, equal protection under the law, constitutional restraints, etc. "He has a right to a fair trial under the United States Constitution."

SPECIFIC LEGAL CLAIM: Statements in which the speaker asserts that someone has a legal right to do something, but does not give a legal principle as a basis for the right. "Mr. Kohler has a right to fire any worker he wants.

GENERAL FACTUAL CLAIMS: Causal, descriptive, or predictive generalizations. "Negroes are just as intelligent as whites."

SPECIFIC FACTUAL CLAIMS: Statements describing specific events delineated in time and space. "The first attempt at integration in Little Rock was on September 4, 1957."

SOURCE: A statement or part of a statement describing the source on which a claim, definition or value judgment is based. "Emergency is defined this way in Webster's New International Dictionary."

DEFINITIONAL CLAIM: A statement about how a word or phrase is defined or should be defined. It is also a statement of analysis by which several meanings of a single word or statement might be distinguished. "An emergency occurs when one or more people are in danger of being injured or losing their lives and property."

REPETITION: A statement in which the speaker repeats himself or communicates something already stated in order to focus the discussion.

TABLE 10.5—continued

CASE: A set of statements which describes specific, real, or hypothetical situations analogous to the one under discussion. Its main purpose is to elaborate the range of situations to which one might apply a value judgment. "Suppose Negroes and whites were given schools of equal quality, teachers of equal quality, books and educational facilities of equal quality: would Negro schools still be inferior to white schools?"

RELEVANCE: Statement which explicitly deals with the way a statement or groups of statements is related to the total argument or to the specific point under discussion. "I don't see what that statement has to do with the discussion."

DEBATE STRATEGY: Ad hominem or other remarks which explicitly discuss the tactics being used by a discussant. "You're just trying to confuse me."

TASK—PROCEDURAL: A statement directed at controlling the immediate interpersonal situation, and which assumes that everyone in the discussion is trying to do a conscientious job. "Let's take a vote." "Let's give everyone a chance to talk."

DEVIANCE CONTROL—PROCEDURAL: A statement directed at controlling the immediate interpersonal situation, and assuming that one or more people are violating group norms. "Get back in your seat and sit down." "You don't have to shout."

within the statement itself, and because there is often no static category which can be appropriately scored with it.[53]

Posture of the speaker. "Posture" refers to the attitude of the speaker toward the statement he is making or the function which that statement is performing for the speaker. We have identified and used four postures: declarative statements; interrogative statements; statements which question or express doubt about a prior statement (often in either the declarative or interrogative form, but with an overtone of argumentative intent); and, statements which express self-doubt (as, for example, uncertainty as to the validity of a claim which has been or is going to be made by the speaker). The posture of the speaker is scored with a symbol within the space provided on a scoring sheet for the appropriate static category.

Orientation of the speaker to the discussion: analysis versus persuasion. We also distinguish and score whether or not the speaker is trying to persuade other group members that his substantive position in the argument is correct, or whether he is attempting to stay "outside" the argument and simply analyze how the group might construe the issues in the case. For example, "That person in the case should not have been allowed to speak because avoiding a riot is more important than his right to speak," is scored as persuasive. The statement, "The problem here is that the principles of freedom of

[53] At this point it should be noted that both the interaction system being discussed here and the one we used to describe teaching styles differ markedly from Bales' in their use of double scoring. That is, each act is scored in at least two subsystems simultaneously. Multiple scoring is possible largely because we score from tapes which allow us to control the rate of scoring. It would be much more difficult to use a system requiring multiple scoring in a live situation. For purposes of analysis, it is important to use an observer scoring sheet which tells us what particular acts have been multiple scored, so that we can distinguish the total units of behavior from the number of categorizations made.

speech and peace and order are both involved in the situation, and we must decide which value should be given greater weight in this instance," is scored as analytical.

Validity and reliability. Initially, of course, using the system to categorize statements in a discussion results in an abstract cognitive description of the discussion. This description must be translated into a quantitative score by determining which categories seem valuable from the point of view of our objectives, and then counting the frequency with which units are scored in these categories. This selection of valued categories is essentially a question of validity. Thus far the system appears to have not only intuitive or face validity, but also reflects the effect of experimental training in reflective thinking. Data on this point will be presented in Chapter Eleven. We have undertaken procedures by which validity can be more firmly estblished and preliminary results seem to bear out our faith in the instrument. Presently, we feel that the following categories have value for a discussion involving political controversy.

Static Categories

General Value Judgments and General Legal Claims are valued because they allow the student to deal with the controversial case at a more abstract and general level.

Specific Factual Claims and Sources are valued because they are appropriate ways of supporting more general claims. They are an important part of the empirical proof process.

Definitional Claims are valued because they tend to demand or give greater precision to the various positions in the argument.

Repetition is not valued, since it involves mainly statements which repeat something already said. When the student clarifies by drawing finer distinctions between positions or terms in the argument, it is scored as a Definitional Claim.

Case is valued because, by definition, it is an attempt to expose the point at which an individual will reverse his position, given an array of similar situations to judge. It is essentially a defining operation.

Relevance is valued because it indictes that the student is attempting to deal with the relationship between a particular statement and some larger facet of the total argument.

Dynamic Categories. For obvious reasons, all three dynamic operations are valued. They hve been selected for scoring precisely because we think they are important.

Orientation to Discussion. The analytic orientation to the discussion is valued because it tends to indicate that the student is attempting to stand back from the immediate persuasive aspects of the argument and provide a more impartial framework by which to deal with the controversy. The questioning posture may be valued especially in unsophisticated groups when it tends to require discussants to clarify or support a position.

It should be noted that these valued acts are not simply the product of a priori guessing about what acts operate to produce the most intelligent discussion. In arriving at our present position, we have listened to many discussions and done a good deal of cutting and fitting to make our quantitative scoring procedures consistent with our intuitive judgments about what behavior is actually important for clarifying a controversial situation.

Although validity is based mainly on these subjective judgments, we have carried out more systematic work on reliability to establish whether or not the subtleties of language can be objectively scored with these gross categories. Reliability, of course, has two meanings in this context. It can refer to the consistency of behavior under observation or to the consistency with which behavior is observed or categorized. It is the latter with which we are concerned at this point. Initially, as part of the training procedure, agreement among observers in the frequency of units assigned to specific categories was checked by a graphic method.[54] Having reached an acceptable level of agreement as estimated by this method, we turned to the agreement between observers on the total number of valued acts which should be credited to each student. The degree of association was estimated using the product-moment correlation. Initially four persons were trained to use the system. Each scorer was paired with every other scorer, so that six scoring combinations resulted. The discussions scored were Socratic interviews between a student and adult interviewer, in which the student was challenged to defend a position on a controversial case. The number of discussions scored by each combination ranged from 10 to 18. The results are shown in Table 10.6. There is no widely accepted criterion for the acceptance of such coefficients as satisfactory; as Heyns and Zander[55] point out,

---

[54]Binomial probability paper as developed by Frederick Mosteller and J. W. Tuckey and reported in "The Uses and Usefulness of Binomial Probability Paper," American Statistical Association Journal, XXXXIV, 1949, pp. 174-212. For a statement of its application to systematic observation, see R.F. Bales, Interaction Process Analysis, pp. 111-112.

[55]R. W. Heyns and F. F. Zander, "Observation of Group Behavior," Research Methods in the Behvioral Sciences, ed. by L. Festenger and D. Katz (New York: Dryden Press, 1953), p. 411.

whether one demands a correlation of .70 or .90 is contingent upon the uses to which the observational scores are to be put. As we are now reporting our system within a specific research context, it seems sufficient to point out that with the exception of one coefficient all approach at least .70, with two greater than .80, and one greater than .90. On the average, there is a relatively high level of agreement.

A second reliability study was carried out on a larger sample of discussions, computed by individual valued acts. In this case the scoring was done by two men, one an undergraduate at Harvard College and the other a student at the Harvard Graduate School of Education. Neither scorer knew the purpose of the scoring system or the distinction between valued and non-valued acts. The situtation scored was pupil-led discussion, in groups of 10-14 students based on controversial cases. The dynamic operations are not included because of the low frequency in these categories. (Between one-third and two-thirds of the discussions contain a frequency of less than three on the three dynamic categories.) These data are presented below in Table 10.7. The reliability of the three low frequency dynamic (dialectical) categories was tested for individual discussion on binomial probability paper. In the 32 discussions scores fell outside acceptable limits three times for generalization-specification, twice for qualification, and three times for consistency-inconsistency.

It should also be reported here that this scoring was done immediately after a training period. We found that scorers tended to become unreliable after a relatively short period of independent scoring, creating some very difficult problems which will be discussed in Chapter Eleven.

We have noted that in applying the concept of reliability to the quantification of oral or written narrative materials through

TABLE 10.6

RELIABILITY ESTIMATES FOR FOUR OBSERVERS

USING THE CATEGORY SYSTEM

	A	B	C	D
A				
B	.55			
C	.82	.93		
D	.87	.69	.68	

TABLE 10.7

RELIABILITY BETWEEN TWO SCORERS FOR SPECIFIC VALUED ACTS ON THE SIAT No. 4 FOR 32 STUDENT-LED DISCUSSIONS

Valued Acts r	Correlation (r)
General Value Judgment	.76
General Legal Claim	.79
Specific Claim	.71
Source	.77
General Definition	.86
Specific Definition	.86
Case or Analogy	.73
Relevance	.51
Questioning Posture	.60
Analysis	.42
Total Valued Acts	.89

systematic content analysis a distinction must be made between the reliability of the behavior being measured and the reliability with which observers can categorize that behavior. The latter type of reliability provides very little problem with most paper-and-pencil tests because there is little room for error or disagreement once scoring keys have been adopted. The first type of reliability also provides little trouble because it is clear that the behavior referred to is that involved in making responses to the test. That is, the question is, "Can we predict from the responses on one test how the individual, or group of individuals, will respond on the same test at a different time?" This form of reliability is, however, fraught with ambiguities in the systematic observation situation because the observer's interpretations intercede between the individual's behavior and the categorizations which are quantified to obtain a test score. It is often not clear whether in speaking of consistency of behavior we are referring to the actual behavior manifested by the individuals being tested or to the categorizations of that behavior which result from observer behavior. Of course, in either case the reliability of the observer in categorizing behavior will affect the quantifications which we must use in our estimation.

Above we have treated the reliability with which observers can apply the category system. The question we would like to deal with now is whether or not, within the limitations of observer reliability, the observational instrument produces an estimation of behavior which is consistent over time. A factor here, of course, is whether the sample of behavior categorized is large enough to serve as a basis for prediction about future categorizations which will be obtained with the individual or group. With paper and pencil measures this reliability is commonly estimated by correlating individuals' scores

derived either from odd versus even items on a test, or from the administration of two forms of a test at two different times. To make an estimate of internal reliability similar to the odd-even comparison, we obtained two scores for each student in the Socratic interview situation. Each scoring sheet, containing spaces for scoring fifteen acts, was numbered in serial order and then used in that order in scoring the interviews. We summed scores on the odd scoring sheets and correlated these with the sums obtained from the even sheets. The correlation obtained is .67, corrected by the Spearman-Brown formula for total valued acts.

Some Concluding Statements on Evaluation

We would conclude this chapter on evaluating competence in political analysis by stating some of the major principles which have developed out of our own work.

1. The criterion of competence in political analysis must be established in a less structured and more realistic setting than that allowed by the multiple choice pencil-and-paper test. Our own evidence indicates that there may be little or no relationship between competence to defend one's point of view in public, and competence required in any one of the common "critical thinking" tests.
2. The use of content analysis to assess competence in political analysis is complicated and fraught with reliability problems. The reliability problems are two-fold: (a) How much behavior does one need to measure before one can make reliable predictions about a person's analytic competence? and (b) To what extent can people agree on how this behavior should be described quantitatively? Nevertheless, we feel these

reliability problems can be overcome sufficiently to allow the use of such a system as a criterion index against which to compare less costly and less complex methods of measurement. To proliferate the measurement field with the simple measurement tools already available, which are so heavily saturated with general reasoning and verbal factors, makes little contribution.

3. The translation of the objectives of our pedagogical approach to political controversy into specific learning outcomes which can be measured with a set of categories such as described above presents, we believe, unusual possibilities for curricular evaluation. Because learning outcomes can be measured in a situation less structured than paper-and-pencil tests and approaching more closely the circumstances in which teh desired concepts will later be applied, the results take on greater meaning and validity. Our reliability data suggest the feasibility of this approach to assessment both in experimentation and classroom teaching. It should be noted, too, that just as a teacher might during any one period of time teach for only one or a few of the concepts included in the category set, so might the set be modified to include fewer categories in order to simplify scoring.

There is, however, no denying the impracticability of careful content analysis for the day to day needs of the average classroom. Teachers, in general, have neither the research competence nor the time to learn and use such a complex system. Ultimately, however, the more complex instrument might be used to establish the validity of simpler category systems, or even of pencil-and-paper tests. There is

little doubt in our own minds that present methods of measurement which attempt to assess the process of reflective thinking with a series of fragmented multiple choice items show insufficient respect for the subtlety and complexity of this competence. It is our conviction that measurement programs will become more significant to teachers and research people when evaluation begins with a recognition of the complexity of the phenomena they are attempting to describe and assess.

4. We should bear in mind that content analysis itself, while it may give us a more reliable picture of an interaction sequence, will not tell us what types of actions or sequences of actions should be valued. This can be done only by some kind of philosophical analysis into the question of what constitutes "rational conduct."