DOCUMENT RESUME

ED 279 690 TM 870 077

AUTHOR Nickerson, Raymond S.
TITLE Reasoning in Argument Evaluation.
PUB DATE 86
NOTE 72p.; One of 46 papers commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report "The Nation's Report Card" (TM 870 049). For other papers in this group, see TM 870 050-094.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Cognitive Tests; *Critical Thinking; Difficulty Level; *Discourse Analysis; Educational Assessment; Educational Testing; Elementary Secondary Education; *Logic; Measurement Objectives; *Multiple Choice Tests; *Persuasive Discourse
IDENTIFIERS National Assessment of Educational Progress; Teaching to the Test

ABSTRACT

A number of higher order cognitive skills are used in the task of evaluating arguments. Such skills should be assessed because the ability to evaluate arguments is an important one in all subject areas. In addition, it seems reasonable to assume that these evaluative skills will be representative of those required by other cognitively demanding tasks. Specific theories and curricula for evaluating persuasive arguments are not available to guide test construction; in lieu of a theory, reasoning in argumentation is defined and some of the processes involved are explained. Understanding of the difference between logical validity and empirical truth is one component of effectively evaluating formal arguments. Evaluation of informal arguments involves many skills: analysis, judging relevance and weight, synthesis, use of prior knowledge, information seeking and selection, and estimation. Based on the assumption that teachers do teach to the test, it is important to find ways to test the ability to evaluate arguments and to diagnose students' skills. The 31-page appendix describes ways to make multiple choice tests more informative, both to examiner and examinee. (GDC)

Reasoning in Argument Evaluation

Raymond S. Nickerson

BBN Laboratories

Cambridge, Mass.

Paper commissioned by

THE STUDY GROUP ON THE NATIONAL ASSESSMENT OF STUDENT ACHIEVEMENT

1986

REASONING IN ARGUMENT EVALUATION

Raymond S. Nickerson
BBN Laboratories
Cambridge, Mass.

The recent surge of interest in the teaching of "higher-order cognitive skills" stems in part from an increasing awareness that it currently is possible to emerge from twelve or thirteen years of schooling -- in the words of one widely cited report -- "ready neither for college nor for work" (National Commission on Excellence in Education, 1983, p. 12). Evidence has been provided by numerous investigators that even high school students who plan to go on to college often arrive there deficient in the kinds of cognitive skills that effective learning in college courses -- especially in science and math -- requires (Carpenter, 1980; Gray, 1979; Karplus, 1974; Kolodiy, 1975; Lawson and Renner, 1974; Renner and Lawson, 1973).

What is meant by higher-order cognitive skills is not always clear. The following are representative, however, of examples of behavior that is sometimes taken as evidence of the operation of such skills:

- analysis of tasks or situations into meaningful components
- judgment of relevance; distinction between those aspects of a problem that deserve attention and those that do not
- recognition of commonalities between similar structures in different contexts or domains; analogical reasoning
- effective planning of approaches to cognitively demanding tasks; revision of plans as necessary
- generation of novel, but workable, approaches to problems; creative thinking
- generation of useful qualitative representations of problems
- detection of flaws, oversights, inconsistencies in plans or proposed approaches to problems; critical thinking
- selection and use of intellectual tools appropriate to the task
- inferential application of knowledge
- application of principles or procedures learned in one domain in other domains, as appropriate.
- understanding of concepts, processes, relationships, principles at a more-than-superficial level (the kind of understanding that permits one to recognize whether a particular algorithmically-derived solution to a problem could possibly be correct)
- thinking at the level of principles and abstract relationships; classification on the basis of abstract properties
- sensitivity, in reading, to authors' assumptions, purposes, devices
- sensitivity to the strengths and weaknesses of own knowledge base vis-a-vis specific tasks

- effective management of own cognitive resources and monitoring of own performance

Such behaviors have broad applicability across wide-ranging situations and domains and development of the ability to engage them is a worthy educational goal.

This paper focuses on skills of the type represented in the above list as they pertain to the task of evaluating arguments. Some limitation of scope is necessary in an exercise of this sort and this focus is justified on two grounds: first, the ability to evaluate arguments effectively is itself an immensely important one, not only in science and math, but in everyday life as well, and, second, it seems reasonable to assume, or at least hypothesize, that skills that prove to be useful in argument evaluation will be broadly representative of those required by many other cognitively demanding tasks. In other words, the ability to evaluate arguments effectively is sufficiently important in its own right to warrant making its development a primary educational objective, whether or what is learned can be expected to generalize; and it is not unreasonable to expect some generalization to occur.

Our concern here is on testing rather than on teaching; however, inasmuch as a major reason for testing is to determine whether students have acquired the knowledge and skills that the educational process is intended to develop or impart, the question we have to ask straightaway is what knowledge and skills is education intended to cultivate with respect to argumentation. What are the major objectives of education in this area, or what should they be?

Here it would be helpful to have a theory of reasoning as it pertains to argumentation to guide thinking and test construction. Such a theory would provide not only a general conceptualization of argument, but also a framework for classifying the various types of arguments that exist. It would provide conceptual tools to facilitate analysis and a basis for characterizing arguments with respect to such properties as structural complexity and difficulty of evaluation. It would indicate how competency in argumentation would be expected to increase with age. And so on. To my knowledge there is no such theory, nor is there likely to be one that is widely accepted in the near future. To be sure, there do exist taxonomies of thinking skills or intelligence components that might be helpful

(Ekstrom, French, Harman & Derman, 1976; Ennis, 1985; Guilford, 1967; Sternberg, 1981). One or more of these taxonomies - perhaps an amalgam of them - might provide the basis for a conceptual framework, but none of them focuses on argumentation primarily and their application in this area would require some further development.

In the absence of a widely accepted theory, one might look to curriculum for guidance on test construction. When an established curriculum exists, as it does in basic mathematics and in the core sciences, developing a competency test may be a (conceptually) simple matter of consulting the curriculum to identify the educational objectives and then constructing test items designed to provide information regarding how well those objectives have been attained. Unfortunately there does not exist (to my knowledge) an established curriculum in reasoning and argumentation.

Given neither a theory nor an existing curriculum to guide test construction, one might turn the process on its head and take the position that a test may serve to define the domain of interest, at least for the present. After all, test items are

intended to reflect the knowledge and skills that one who is competent in the domain is expected to have. So the test battery as a whole might be taken as a reflection of the domain.

In my view, this approach has some merit. Eventually, one hopes to have a theory that not only will provide a rationale for test content, but will support a more adequate understanding of what competency with respect to reasoning in argumentation involves. In the meantime, the compilation of a collection of test items that are considered to tap various aspects of such competency can perhaps play a useful role in the development of the desired theory. The history of intelligence testing has shades of this approach. Tests intended to measure intellectual potential were developed on the basis of intuitive notions about what tasks such tests should contain long before there were any widely accepted theories of intelligence. Both the construction of these tests and the results obtained from their administration contributed significantly to the evolution of the idea of intelligence and to subsequent theorizing regarding its structure and function.

The approach taken here is closer to the last one mentioned

than to the other two. In lieu of a theory, I shall describe informally what I mean by reasoning in argumentation and indicate what appear to me to be some of the processes involved. And I shall try to be specific regarding at last some of the knowledge and skills that determine competency, by identifying specific things we should expect an individual who is competent with respect to argumentation to know, understand, or be able to do.

What is an argument

Broadly conceived, an argument is an effort to influence one's beliefs or behavior. Explicit verbal efforts to persuade are readily recognized as arguments. There are many more subtle ways to attempt to influence beliefs or behavior, however, and while some of these may often go unrecognized as arguments, the ability to see them for what they are and to react to them in a rational way is an immensely important one, especially in a media-rich society. A thorough assessment of reasoning ability as it pertains to argumentation would have to pay some attention to the evaluation of both direct, and indirect arguments of various types. Attention in this paper will be focused primarily however, on arguments that are explicit and relatively direct, which seems like the appropriate place to start.

It is important to note that reasoning in argumentation as the term is used here is not intended to equate to ability to win verbal disputes. "Argumentation" is intended to connote broadly the processes involved in constructing and, especially, evaluating arguments. The winning of verbal disputes requires ability in case building, which means marshalling evidence favoring a particular position while ignoring or discounting evidence that opposes it. To be able to reason well in this context means to be able to judge evidence on its merits and to reach conclusions that the unbiased inferential use of evidence supports; it does not mean skill in compiling evidence selectively for the purpose of bolstering conclusions already drawn.

Traditionally, a distinction has been made between formal deductive arguments and informal inductive arguments. The former are those that adhere to one or another cannonical logical form and typically involve proceeding from relatively general premises to more particular conclusions. Informal inductive arguments do not adhere strictly to specific forms and typically involve arguing from particular to more general assertions. Sometimes the term <u>inductive</u> has served to connote any type of argument

that is not wholly deductive. Extended arguments are likely to contain both deductive and inductive components. The distinction between deduction and induction has not lacked its critics, but it has been a widely accepted one for a very long time and is convenient for present purposes.

The teaching of formal logic as a normative or prescriptive model of deductive thought is, of course, a tradition that goes back as far as institutionalized education and there are numerous text books on the subject, aimed primarily at college and, to a much lessor degree, high school level students. The explicit teaching of informal reasoning and, more precisly, reasoning in the context of informal argumentation is not an old established tradition. There are however, a number of fairly recent books that deal with this topic (Damer, 1980; Kahane, 1984; Nickerson, 1986; Ruggerio, 1981; Scriven, 1976). These typically focus on the various ways in which informal reasoning is or appears to be ineffective or wrong. Resnick (1986) points out that the work of philosophers in this area, while featuring a new emphasis on informal logic, still reflects a normative or prescriptive stance, whereas psychologists have tended to be more inclined to study how people who are judged to be good thinkers think and

then to try to teach these techniques to others. There is however, no well developed and clearly articulated theory, either prescriptive or descriptive, of informal reasoning that is espoused by any major subset of either the philosophers or the psychologists whose research focuses on human thought.

Formal deductive arguments

The following is an example of a formal deductive argument taken from a logic textbook (Searles, 1968):

Nothing intelligible ever puzzles me
Logic puzzles me
Therefore, logic is unintelligible

This is the modus tollens form of the conditional syllogism

If A then B
Not B
Therefore not A

Written more true to this form, Searles' example, which came originally from Lewis Carroll, could be stated

If it is intelligible, it does not puzzle me

It (logic) puzzles me

Therefore, it (logic) is not intelligible

The following argument has the same form:

If he were really sympathetic to the demands of the coal miners, he would have voted for the XYZ bill

He voted against the bill

So he is not sympathetic to the miners' demands

The structure of these arguments is obvious and elegantly simple: two premises and a conclusion. Rules for evaluating the validity of deductive arguments of this and similar types are well known. Not all deductive arguments are this simple and short, but longer ones usually can be reorganized as sequences of short ones, with the conclusion from each mini argument becoming a premise in one of those that follows it.

The top-level rules for evaluating a formal deductive

argument are conceptually simple: if the form of the argument is valid and the premises are true, the conclusion must be true. If either the form is invalid or one or more premises is untrue, the truth value of the conclusion is undetermined. In evaluating such arguments, therefore one has to consider questions of both validity and truth.

It is possible to test for the first of these capabilities separately, within limits. The ability to distinguish between valid and invalid deductive forms, for example, can be tested by using arguments that are devoid of semantic content. Consider for example the following arguments:

All bletes are crogs
All crogs are trons
Therefore all bletes are trons

No barps are clints
No clints are frumps
Therefore no barps are frumps

We need not know anything about bletes, or crogs, or the other entities in these arguments to recognize that the first one

is logically valid, and therefore if its premises are true then its conclusion must be true also, whereas the second is not, so being assured that its premises are true would not let us conclude that its conclusion is true.

We know from a considerable body of research that when deductive arguments do have semantic content, people often find it difficult to ignore that content in judging their formal validity (Staudenmayer, 1975; Wason and Johnson-Laird, 1972). Consider, for example, the two following arguments:

No insects are reptiles
No reptiles are mammals
Therefore no insects are mammals

No Scandinavians are Asians
No Asians are Swedes
Therefore no Scandinavians are Swedes

The two are identical in form, and both are invalid because in neither case does the conclusion follow from the premises. In evaluating the first argument with respect to the question of validity one can easily be misled, however, by the fact that the

conclusions and both premises are true. In the second case, the invalidity is more apparent because, while the premises are true, the conclusion is false.

A clear understanding of the difference between logical validity and empirical truth is critical to effective evaluation of formal arguments. While such an understanding is not enough to guarantee that one will always avoid reasoning errors that stem from an insensitivity to this distinction, it should make such errors less likely and increase one's ability to modify one's thinking appropriately when they are pointed out.

The degree to which formal -- and in particular Aristolilean -- logic is also descriptive of normal untutored human thought is a matter of long-standing controversy within psychology. Investigators have compiled long lists of how thinking often appears to be illogical. Some theorists have argued that thought is not greatly constrained by principles of logic (e.g. Harman, 1986), or that even when it produces results that are consistent with logic those results may be based on operations other than logical ones (Johnson-Laird, 1983). Others have taken the position that thinking is basically logical and that what appear

to be evidences of illogic can usually be traced to linguistic confusions or misinterpretations (e.g. Cohen, 1981; Henle, 1962). It is not helpful for present purposes to explore these controversies. The expedient assumption is made here that some knowledge of logic, while perhaps not a necessary, and certainly not a sufficient, cause of competence in everyday reasoning is probably some -- and possibly considerable -- help; no one to my knowledge claims that it hurts.

The insufficiency of a knowledge of logic to guarentee competence in argument evaluation is not a controversial issue, however, because most of the arguments one encounters in everyday life are neither exclusively deductive nor expressed in a cannonical logical form. Evaluation of these arguments requires more than the ability to distinguish between valid and invalid syllogisms.

Informal arguments

The following are two informal arguments taken from a daily newspaper (Boston Globe, 1982, p.8). The first is from a proponent of a Massachusetts law requiring that a refundable

deposit be paid for certain beverage containers sold in the state. The second is from an opponent of the same bill. A "yes" vote on an upcoming referendum would keep the law in place; a "no" vote would do away with it. The arguments were restricted in length.

Argument For:

The bottle bill was passed by two-thirds vote of the Legislature last year in order to clean up the litter cluttering our lawns, streets, and parks. This sensible legislation will put a stop to an enormous waste of money and resources.

Similar laws have proven popular and successful in achieving those aims in states much as Maine, Vermont, Connecticut, and Michigan. New York has passed a bottle bill which will take effect next July.

Among the benefits of the Massachusetts bottle bill:

- Reduced litter: An 80 percent decrease in beverage container litter, the most prevalent and dangerous type;

- Lower-prices: A 5 percent reduction of beverage prices;

- Tax savings: Millions of tax dollars saved due to reduced garbage collection and disposal costs;

- More jobs: A net increase of more than 2000 skilled and unskilled jobs for Massachusetts residents.

- Less waste: 33 percent less energy consumed by the beverage industry.

Argument Against:

In this age of limited resources, it is understandable that Massachusetts voters said "no" to forced deposit containers the only time they had an opportunity to vote on the issue. They knew that requiring return of cans, plastic bottles and other containers will create many problems, and were concerned about the needs to:

- Save town and private recycling operations.

- Save the state's water supply during a declared emergency.

- o Save local separation programs because landfill space is rapidly disappearing.

- o But most importantly, save money for already over-burdened consumers.

Further, our position is that the will of the electorate is better served by people voting in free elections than legislators acting in a highly political environment.

A "no" vote is a vote for an industry-funded litter control and recycling law - which will give money to communities for cleanup and recycling projects and anti-litter education programs.

Here is another example of an informal argument, this one in support of the further development and use of nuclear power in the U.S.:

That there is nothing inherently unworkable about nuclear power is borne out by the success of nuclear projects overseas and indeed of many projects in the U.S. Moreover, although a variety of alternative sources of

energy are available to utilities, and several promising new non-nuclear technologies are under development, abandoning nuclear power might well render the nation's electricity supply substantially less efficient and environmentally benign. Oil and gas, albeit currently plentiful, will eventually grow scarcer and costlier. Coal is difficult to burn cleanly: acid rain and other by-products of coal combustion are causing serious damage to the environment and will be expensive to control, and in the future, the carbon dioxide released by the combustion of fossil fuels may have a severe effect on global climate. The potential of solar electric technology to compete economically outside a fairly narrow range of favorable sites or specialized uses has not yet been demonstrated (Lester, 1986, p.31).

These examples of informal arguments were composed by their originators with an adult readership in mind, and therefore probably would not be appropriate to use in assessing the reasoning abilities of young schoolchildren. They or similar arguments would be appropriate for use with high school students, however, and comparable arguments could easily be found or composed for use with younger age groups.

Evaluation of informal arguments

While informal, all these arguments are still much neater than many of those one encounters in daily life. The people who constructed them were motivated to make them clear, compelling and concise. Even so, they are not nearly as easy to evaluate as are formal deductive arguments; and there are no widely agreed-upon rules for doing the evaluation. To evaluate such arguments one must, at least, recognize the claims that are being made (each of the first two of these arguments contains several claims in addition to those that are highlighted by being set off as "bulleted" items), decide how much support each of these claims -- if true -- gives to the conclusion or position the argument is intended to substantiate, determine -- at least for each of the more weighty claims -- how much credence to give it, and somehow aggregate the results of these considerations into an overall assessment of the compellingness of the argument as a whole. In other words, evaluation of informal arguments of any substance and complexity involves at least the following components:

- Analysis. Figuring out what the essence of the argument is. This is made the more difficult when the argument is expressed poorly or its originator's intentions are camouflaged by a superfluity of words, but in order to evaluate an argument one must know, or make an assumption about, what the argument is. This means identifying the conclusion(s) one is intended to draw and what is being asserted in its (their) support.

- Judgements of relevance and weight. To determine how much credence to give to the conclusion of an argument, one must be able to judge the assertions that are made in its support as to their relevance and, given tnat an assertion is considered relevant, how much weight to attach to it. In both cases the judgement is one of degree, relevance can vary continuously as can weight. The same types of considerations must be given also to counterarguments that one may construct, inasmuch as the assertions comprising them can also be more or less relevant and can vary in the degree to which they increase the credibility of the counter conclusion.

- Aggregation or synthesis. Somehow, having considered the argument's parts, one must arrive at an assessment of its compellingness as a whole. One must decide how persuasive it is relative to the most compelling counterarguments one has been able to construct, and on this basis accept or reject its conclusion(s), perhaps with qualifications or provisos, and probably at some level of surety less than absolute.

I do not mean to suggest that one always does each of these things in a conscious and deliberate fashion, but it is clear that the effective evaluation of informal arguments involves them, at least implicitly. Further, these activities draw upon other capabilities in turn. The following few illustrate the point.

- Assessment of own knowledge. If one knows a lot about the topic of the argument, one is in a better position to evaluate it than if one knows only a little. Equally as important as one's knowledge of a topic, however, is one's awareness of the extent and limitations of that knowledge. Especially vulnerable is the person who believes his knowledge to be extensive when, in fact, it

is very limited. This consideration becomes especially important in certain types of plausibility judgements. Suppose, for example, that I judge a particular assertion to be highly improbable on the grounds that it declares to be a fact something of which I am unaware. My tacit assumption must be that if that which is declared to be a fact really were a fact, I would be aware of it, and since I am not aware of it, it must not be a fact. Such a basis for judging plausibility is justified only to the degree that I am highly knowledgeable with respect to the subject. If my knowledge of the subject is very limited and I am aware that it is, then the fact that I do not know some assertion pertaining to that subject to be true should give me very little reason to conclude that it is false.

- Information seeking and selection. It may be desirable to obtain additional information to supplement what one has in one's head, especially if one's knowledge of the domain is limited. Skill in the finding of information is a useful one not only in the context of argument evaluation, but much more generally. (At least some

aspects of this skill should be relatively easy to address through training, namely those pertaining to the use of formal information-finding resources such as encyclopedias, atlases, almanacs, and indexes. Instruction in the effective use of such resources should, I believe, receive considerable emphasis throughout the educational system. This is not to suggest that students should be led to adopt the idea that the answers to all questions can be found or should be sought in books, but it is to their advantage to know how to find those that are to be found there.)

- Estimation and approximation. Many arguments have quantitative components: assertions about costs, populations, incidence, probabilities, distances, rates of change. In order to evaluate such arguments it is necessary to judge the plausibility of such assertions, and inasmuch as the quantities involved are often -- probably typically -- unknown to the evaluator, such judgements must be based on the ability to estimate or approximate them.

- Detection of inconsistencies. Inconsistencies can occur within an argument or between an argument's assertions and known or assumed facts. Such inconsistencies, when recognized, weaken an argument, so the ability to detect them is an important one. This is a knowledge-based ability; one cannot detect inconsistencies between assertions and facts unless one knows the facts.

To assess an individual's competence with respect to argumentation probably requires explicit attention to each of these aspects and several more. It is not safe to assume that an individual who is competent at detecting fallacies in deductive arguments posed in syllogistic form will be equally competent at judging the amount of credence that should be given to assertions offered in support of informal inductions.

Inasmuch as most of the arguments encountered in daily life are neither complete and well-formed nor strictly deductive, the practical question that must be decided is not whether the conclusion follows logically from the premises, but how much credence one should give to the conclusion in view of the claims that have been made in its support. Even in the case of strictly

deductive (say syllogistic) arguments that are encountered in daily life, evaluation involves more than judging formal validity. If one is trying to determine whether to accept the conclusion as true, one must be satisfied not only that the form of the argument is valid, but that all of its premises are true. The ability to judge the truth or falsity of assertions, or their degree of plausibility, requires knowledge of the domain to which those assertions pertain. It requires also the ability to judge the adequacy of one's own knowledge as it relates to that domain: one must be able to judge how one's knowledge of a domain compares to what there is to know about the domain and, in particular, whether one knows enough to have confidence in one's assessment or needs to seek further information. And, it requires the ability to use one's knowledge inferentially.

An important aspect of the evaluation of informal arguments involves going beyond the information contained in the argument itself. The most compelling reasons for rejecting an argument are often found in what could have been said but was not. A general strategy that is useful in evaluating arguments is to attempt to make explicit those relevant facts that were left unstated. It may be especially useful to attempt to explicate

any facts that are inconsistent with the conclusion drawn. That is to say, it may be especially useful to attempt to construct a counterargument. The ability to do so will depend heavily, of course on one's knowledge of the domain. It is perhaps here, in the realm of counterargument construction, that reasoning in argumentation is most heavily knowledge dependent. It is here too that one's metaknowledge - one's awareness of the extent and limitations of own's own knowledge of a domain - is especially important. If I know a lot about a domain, and realize it, and am unable to construct a compelling counter to a given argument, I am likely to give more credence to the argument's conclusion than if I am aware that I know too little about the topic to be able to construct a counterargument even if there were a simple but compelling one to be constructed.

The ability to reason effectively about arguments is an immensely important one in daily life, simply because arguments -- attempts to persuade -- confront us all more or less continuously. Without the ability to evaluate arguments rationally we would be at a loss to know which of the numerous claims that we encounter daily to accept and which to reject.

Competency in argument evaluation

Exactly what should we expect an individual who is competent with respect to argumentation to know? To understand? To be able to do? The following are among the items that would be on my list:

- Understand the difference between logical validity and empirical truth
- Know what is required to disprove a universal statement
- Understand the difference between a cause-effect relationship and a correlational relationship
- Given the true assertion "If A, then B," know what can be concluded

  (1) if A is known to be true
  (2) if A is known to be false
  (3) if B is known to be true
  (4) if B is known to be false

- Know how to evaluate an informal argument
- Understand the difference between proof and corroborating evidence
- Understand the difference between consistency and implication
- Understand what constitutes a contradiction
- Know how the truth or falsity of compound statements depends on the truth or falsity of their components
- Understand what counterarguments are and the role they can play in argument evaluation

- Know how to explicate tacit assumptions underlying an argument
- Be able to identify the key assertion(s) in an informal argument
- Be able to distinguish between highly relevant and relatively irrelevant claims in an argument
- Know how to design tests of hypotheses
- Recognize some of the more common errors of reasoning
- Understand the limitations of argument by analogy

This list could be greatly extended.

An attempt to develop a test of reasoning ability as it pertains to argumentation could have the salutary effect of focusing attention on the question of what the objectives of education should be in this regard. What reasoning skills pertaining to argumentation should educational programs be designed to develop? Given the lack of a well-articulated theory of reasoning and argumentation and also of an established curriculum, the development of such a test will probably require a considerable amount of experimentation. There can be no doubt, however, of the importance of reasoning competency both to the individual in daily life and to some modicum of rationality in the behavior of geopolitical entities as well. Whether we start

with a theory, a curriculum, or a test is less important perhaps than that we start. An effort to develop an appropriate test could do much to stimulate progress in theory construction and curriculum development as well.

Testing

So far, I have focused on the question of test content: what are the capabilities for which a test of reasoning in argumentation should test. There remains the pragmatic question of how to test for these capabilities.

Given unlimited time and resources, testing would probably include the open-ended evaluation of a variety of informal arguments of varying degrees of complexity. The difficulty of administering and scoring such items, however, makes their extensive use improbable. The practical question is whether items can be constructed that get at the various aspects of skill in reasoning in argumentation, and are sufficiently easy to administer and score as to ensure their practicability.

I believe that they probably can be, but am not prepared to

support the belief with convincing existence proofs. I believe further that the development of such items is very important objective. The reasoning behind that belief goes something like this.

> Assumption 1: Teachers will teach to the test; they will attempt to maximize the performance of their students on the tests that are being used to assess academic achievement.

This strikes me as neither surprising nor objectionable. A major purposes of testing, after all, is to evaluate the effectiveness of instruction. Can we really expect teachers to ignore the tests, this being true? Would we want them to do so?

> Assumption 2: Tests can be structured in such a way that it is possible for students to learn to do well on them without acquiring the intellectual competence that high test scores were intended to reflect.

Much of the concern about the current status of standardized testing derives from the belief that reliance on multiple choice testing techniques tends to encourage undue emphasis on the rote

learning of "facts", not because such an emphasis is known to be in the best interest of the students' long-term intellectual development, but because it is what is needed to help them do well on their tests. The cost of an undue emphasis on fact learning, in the view of the critics, is the neglect of those aspects of intellectual competence that are not easily assessed with multiple-choice or other equally convenient objective techniques.

Note that Assumption 2 is flatly contradictory to another one that might be made, namely that ability to do well on a test is compelling evidence that the test taker has the intellectual competence the test is intended to assess. This is the issue of test validity.

> Assumption 3: Recognition of the role that tests play in guiding teaching behavior, as well as their assessment function, should strongly influence test design; in particular, tests should be designed intentionally so that when teachers teach to them, they will be working toward, and not in opposition to, the fundamental educational objectives.

Because I view skill in the evaluation of arguments as extremely important, both to substantive academic achievement and to rational behavior in everyday life, I believe its development should be a major educational objective, and this is much more likely to be the case if it is a focus of educational testing than if it is not.

The down side of the teaching-to-the-test phenomenon relates not so much to what gets taught, but to what gets left out. Given limited time and resources, overemphasis of A means underemphasis of B. When teachers concentrate on teaching students whatever they need to know to do well on achievement tests, they are likely to neglect to teach what is not addressed by those tests. It follows that if the content and structure of tests are determined more by considerations of test administration convenience than by the goal of addressing all of the most important aspects of educational achievement in a representative way, the testing program may well subvert the original educational goals.

If the assumption that teachers will teach to the test is sound, and if one of the recognized uses of testing is to shape

teaching objectives, one might justify including in a test items that accurately represent the target skills even if those items are impractical to score. Inclusion of such items could diminish the effectiveness of a test as a measurement instrument however, so that is probably not an option.

Perhaps the most serious risk in test construction is that considerations of feasibility and convenience will overwhelm considerations of validity and representativeness, leading to tests that are easy to administer and score but that measure only limited aspects of what should be measured and divert teaching objectives in the process. On the other hand, it is difficult to challenge the assumption that any test that is to be widely used to assess the effectiveness of education or specific aspects thereof throughout the country will have to be relatively easy to administer and to admit of objective, unambiguous scoring. The problem of devising a test that has these characteristics and that really gets at reasoning ability in the context of argumentation is clearly a challenging one.

When a test of reasoning in argumentation is developed, it would be very desirable if, in addition to yielding a scalar

indication of reasoning ability, such a test provided considerable diagnostic information as well. With respect to this issue, I will end these comments with the simple observation that there are ways of making multiple choice tests much more informative than they typically are. Implementation of these techniques may be impractical in the context of a universal testing program. Although if such techniques were widely used in the school systems, so one could assume that students above some grade level were familiar with them, they might become feasible. If they were feasible, they would have distinct advantages from the point of view of diagnosticity, inasmuch they would provide considerably more information to the evaluator than do such tests as currently administered and scored. One such method is described in the attached appendix.

It must be acknowledged that even with the modification described, multiple choice tests still are designed to test primarily knowledge and not process. Moreover, the procedure does not address the issue of test construction or item selection. There is a need for some inventive thinking about the construction and administration of tests to assess ability in argument evaluation; it is an ability well worth having -- it is

difficult to imagine many that are more important -- and, consequently, given the role of testing as a forcing function in education, the development of adequate techniques for assessing this ability is worthy also of considerable effort.

References

Boston Globe (1982). Bottle refund law? October 28, 1982, pp. 1,8

Carpenter, E.T. (1980). Piagetian interviews of college students. In R.G. Fuller et al, (Eds.), Piagetian programs in higher education, (pp. 15-21). Lincoln: University of Nebraska-Lincoln, ADAPT Program.

Cohen, L.J. (1981). Can human irrationality be experimentally demonstrated. The Behavioral and Brain Sciences, 4 317-370.

Damer, T.E. (1980). Attacking faculty reasoning Belmont, CA: Wadsworth.

Ekstrom, R. B., French, J.W., Harman, H.H., & Dermen, D. (1976). Manual for kit of factor-referenced cognitive tests. Princeton, NJ: Educational Testing Service.

Ennis, R.H. (1985). Critical thinking and the curriculum. National Forum, 65, 28-31.

Erickson, J.R. (1978). Research on syllogistic reasoning. In R. Revlin and R.E. Mayer (Eds.), Human Reasoning. New York: Holt.

Gray, R.L. (1979). Toward observing that which is not directly observable. In J. Lochhead & J. Clement (Eds.), Cognitive process instruction. Philadelphia: The Franklin Institute Press.

Guilford, J.P. (1967). The nature of human intelligence. New York: McGraw-Hill.

Harman, G. (1986). Change in View: Principles of Reasoning, Cambridge, MA: MIT Press (Bradford).

Henle, M. (1962). On the relation between logic and thinking. Psychological Review. 69, 366-378.

Johnson-Laird, P.N. (1983). Mental Models. Cambridge, MA: Harvard University Press.

Kahane, H. (1984). Logic and contemporary rhetoric: the use of reason in everyday life. 4th Edition, Belmont, CA: Wadsworth.

Karplus, R. (1974). Science curriculum improvement study: Teachers handbook. Berkeley: University of California, Berkeley.

Kolodiy, G. (1975). The cognitive development of high school and college science students. Journal of College Science Teaching, 13, 20.

Lawson, A. E., & Renner, J.W. (1974). A quantitative analysis of responses to Piagetian tasks and its implications for education. Science Education, 581111,(4), 454-559.

Lester, R.K. (1986). Rethinking Nuclear Power. Scientific American, 254(3), 31-39.

National Commission on Excellence in Education (1983). A nation at risk: The imperative for educational reform. U.S. Department of Education.

Nickeron, R.S. (1986). Reflections on reasoning. Hillsdale, NJ: Lawrence Erlbaum Associates.

Renner, J.W., & Lawson, A.E. (1973). Promoting intellectual development through science teaching. The physics teacher, II

Resnick, L.B. (1986). Education and learning to think. PA: University of Pittsburgh, Learning Research and Development Center.

Ruggiero, V.R. (1984). The art of thinking: a guide to critical and creative thought. New York: Harper and Row.

Scriven, M. (1976). Reasoning. New York: McGraw-Hill.

Searles, H.L. (1948). Logic and scientific methods. New York: The Ronald Press Company.

Staudenmayer, H. (1975). Understanding conditional reasoning with meaningful propositions. In R.J. Falmague (Ed.) Reasoning: representation and process Hillsdale, NJ: Lawrence Erlbaum Associates.

Sternberg, R.J. (1981). Intelligence as thinking and learning skills. Educational Leadership, 39, 18-20..

Wason, P.C. and Johnson-Laird, P.N. (1972). Psychology of reasoning: Structure and content. Cambridge, MA: Harvard University Press.

Appendix A: Getting more information from multiple-choice tests

The problem of knowledge assessment can be viewed from two sides. From the examiner's point of view, the problem is to discover what an individual knows about some domain of interest. From the point of view of the examinee, the problem is to reveal what he knows, and sometimes not so much by means of the testing procedure as in spite of it. No testing procedure is adequate unless it is adequate from both of these points of view.

If we limit out attention to highly structured objective testing techniques, and more specifically, to multiple-choice testing, there are at least two major problems that have to be addressed, which will be referred to here as the sampling problem and the measurement problem. The sampling problem stems from the fact that we cannot hope to discover all one knows about any reasonably complex subject (unless one knows very little about it), by means of any practicable objective examination. Therefore, the test designer must decide where to probe, which questions to ask, in order to elicit a representative sample of the knowledge the examinee has. In terms of test construction, the sampling problem is a problem of choosing an adequate set of test items. I will not consider this problem further here.

The problem of measurement is that of deducing from an individual's answers to a set of questions what he knows with respect to material represented by that set of questions. It is this problem that is the primary concern in what follows. In particular, I will consider the question of how to use multiple-choice tests in such a way that they are less constraining to the examinee and more informative to the examiner than when used in the conventional manner.

Consider first how multiple-choice tests are typically administered. The following is illustrative of the type of question that one might find in a multiple-choice exam designed to assess one's general knowledge of literature.

The short story "An Occurence at Owl Creek Bridge" was written by:

1. Nathaniel Hawthorne
2. Anton Bruckner
3. Edgar Allen Poe
4. Ambrose Bierce
5. Anton Chekhov

Conventionally, one's task on such an exam is to indicate which of the several possible answers one thinks is correct. The examiner, on looking at the test taker's response, learns whether or not he marked the correct alternative. And that is all he learns. He has no inkling of the basis on which the selection was made. Perhaps the individual was certain that a particular alternative was the correct one, and chose it for that reason. On the other hand, maybe he knew nothing whatsoever about the question and his selection represented a pure guess. What is perhaps more likely than either of these extreme possibilities is that he was less than absolutely certain regarding the correctness of a particular alternative, but he knew enough about the material to make a reasonable (from his point of view) choice. For example, if our hypothetical examinee were very familiar with the works of Nathaniel Hawthorne, he might confidently rule him out on the assumption that if Hawthorne had written the piece, he (the examinee) would have been aware of the fact. He might eliminate another of the alternatives if he feels reasonably certain that the story is a part of American literature and he recognizes Chekhov as a Russian author. Bruckner, he might recognize as a composer, and therefore feel safe in eliminating him from consideration. Thus, by applying

his knowledge of literature and music, he might narrow the number of viable alternatives from five to two. Now suppose that he finds it impossible to decide between the two remaining alternatives -- he considers them to be equally likely to be correct. At this point he guesses. His chances of guessing correctly are, of course, much better than they would have been had he not been able to eliminate three of the original possibilities. If he guesses correctly, he will get credit for the question; if not, his answer will be scored as an error. In either case, the examiner has discovered very little about the examinee's knowledge with respect to the subject matter of this question. Clearly, a method that permits the examinee to convey the fact that he knows enough about the question to rule out some of the proposed answers would constitute an improvement over the conventional forced-choice technique.

But, we can go a step further. Consider the case of a second hypothetical examinee faced with the same question. Suppose that he, too, can rule out two of the alternatives, say Bruckner and Chekhov, on some rational basis or other. He cannot eliminate Hawthorne with certainty, but feels that the likelihood that he was the author is very small. With respect to the

remaining two -- and from his point of view the most likely -- alternatives, this person is less ambivalent than the former one. He recognizes both Poe and Bierce as American writers of tales of horror and the macabre; and from what he can recall of the story, having read it many years ago, he finds it highly plausible that either of them might have produced it. He has in fact read extensively both of these authors, and, although he knows that he has not always managed to keep their respective works clearly distinguished in his mind, he is fairly confident that the author of the story in question was Poe. If he were asked to make a wager, he would give 3 to 2 odds -- but no more -- that Poe is the man. Thus, to summarize this examinee's opinion concerning the correct answer to the question, we might say that he considers Poe to be the most likely alternative, Bierce to be only a little less likely, Hawthorne to be a remote possibility, and Cheknov and Bruckner to be definitely not in the running. Surely any answering technique that does not permit one to convey this sort of information is providing a less-than-complete assessment of what the examinee knows. In the case of our hypothetical example, the examinee would get the wrong answer in spite of the fact that he knows a considerable amount about the subject of the question.

A variety of answering procedures suggest themselves as being more sensitive to the nuances of knowledge states than is the forced-choice procedure as it has been used traditionally. The examinee might, for example, be given the option of responding only to those questions to which he feels he knows the answers, being penalized less for an unanswered question than for an incorrect answer. Or he might be asked to give confidence ratings that reflect his degree of assurance that his choices are correct. Or he might be told to order the alternative answers to each question in terms of their relative likelihoods of being correct. Or he might be asked to assign a number to each of the alternatives in such a way that the relative size of the number assigned to any particular alternative is correct; thus, assignments of the numbers 10 and 1, or .4 and .04, to alternatives 2 and 3, respectively, would indicate that the examinee believes alternative number 2 to be ten times more likely to be correct than number 3.

Any of these, or similar, approaches could yield more information about an individual's knowledge than the conventional forced-choice procedure. There are distinct advantages associated with the last one mentioned, however, when it is used

in conjunction with certain scoring procedures, one of which is described below.

The testing procedure that we are looking for should have at least the following two properties: (1) it should be sensitive to what the examinee knows -- the better an individual knows the material, the higher the score he should get; and (2) it should reward honesty -- there must be no way for the examinee to beat the system by doing something different from assigning numbers in accordance with what he actually thinks concerning the probable correctness of the alternatives.

If the first objective is to be met, the score that one receives on any given question should reflect not only whether the examinee selects the correct answer, but also how much confidence he has in his selection. In general, given the answering technique mentioned above, we would expect the score for any item to depend, at least in part, on the relative size of the number that is assigned to the correct alternative. Thus, assuming that the sum of the numbers that an examinee has used on the alternatives for a given question is 13, we would expect him to get a higher score if he has placed 10 on the correct

alternative than if he has put 4 on it. Why not then simply make one's score the ratio of the number placed on the correct alternative to the sum of the numbers placed on the alternatives associated with a given question so that, for example, if one had placed 10 of a total of 13 points on the correct alternative, his score for that question would be 10/13. Unfortunately, this simple rule does not satisfy our second desideratum. Given such a scoring procedure, the examinee should cheat. Specifically, he should always put zeros on all the alternatives except the one that he considers most likely, even if he is not very certain that that alternative is indeed the correct one.

This is easily seen by considering a two-alternative case. Suppose that the examinee really thinks that the chances are 7 in 10 in favor of A being the correct alternative. If he is honest, then he will assign 7/10, of whatever points he is going to use, on alternative A and 3/10 on B. Given our scoring rule, and assuming that our hypothetical examinee assigns numbers to the two alternatives in the ratio of 7 to 3, then the two values that his score may assume are 7/10 and 3/10. Moreover, from the examinee's point of view, the probability of getting a score of 7/10 is 7/10 (i.e., the probability that A is correct), and the

probability of getting a score of 3/10 is 3/10. Thus, the expected value of his score is $(7/10)^2 + (3/10)^2 = .58$, (the expected value of a variable being the sum of all possible values of that variable, each weighted by the probability of its occurrence). But suppose that our examinee is a gambler, and decides to put all his chances on the alternative he considers most likely to be correct. Now the two values that his score can assume are 1 and 0, and the expected value of his score (assuming that he really believes that A's chances are 7 in 10, rather than 10 in 10 as his answer would indicate) is 7/10 x 1 + 3/10 x 0 =.70. Thus, whereas we have instructed the examinee to assign numbers to alternatives in accordance with his judgement of the likelihood of their being correct, our scoring rule is such that he can expect to obtain a higher score by ignoring our instructions than by following them.

Fortunately, scoring rules exist that resolve this dilemma. One such rule -- the only one that will be considered here -- was described by Roby (1965) and is sometimes referred to as the "spherical-gain" scoring function. According to this rule, one's score on any question is the number assigned to the correct alternative, divided by the square root of the sum of the squares

of the numbers assigned to all the alternatives. The score for the j^{th} question is given by

$$s_j = x_{jc} \left(\sum_k x_{jk}^2 \right)^{-\frac{1}{2}} \qquad (1)$$

where s_j represents the score received for the j^{th} question on the test, x_{jk} the number that the student assigns to the k^{th} alternative for the j^{th} question, and x_{jc} the number assigned to the correct alternative for that question. To illustrate the method, consider a five-alternative question for which the second alternative is the correct one. Table 1 shows such a question along with several hypothetical answers to it and the score that each answer earns.

It should be obvious that $0 \leq s_j \leq 1$. The score will be zero if zero has been assigned to the correct alternative; it will be one if zero is assigned to every alternative <u>but</u> the correct one. (Note that a "pure guess" -- the assignment of the same number to each alternative -- will not result in a score of zero.)

Table 1. Question: Which one of the following presidents of the United States served two non-consecutive terms?

1. James Madison
2. Grover Cleveland
3. William Harding
4. Zachory Taylor
5. James Buchanan

(The correct answer is #2.)

Hypothetical answer:

	A.		B.		C.		D.	
	1	0	1	0	1	1	1	0
	2	7	2	10	2	1	2	0
	3	0	3	0	3	1	3	0
	4	6	4	0	4	1	4	0
	5	2	5	0	5	1	5	5
Score:	0.74		1.00		0.45		0.00	

It is clear that this scoring rule has the first of the two desirable properties mentioned above. That is to say, in general, the larger the number placed on the correct alternative (relative to the numbers assigned to the other alternatives), the larger the resulting score. Although the fact is less obvious, the rule also satisfies the second desideratum: the examinee maximizes his expected score only if he assigns numbers to the alternatives in accordance with his true belief regarding their relative chances of being correct. For a mathematical proof of this assertion, see Schuford, Albert and Massengill (1966). A reference to the example that was used earlier should be sufficient to make the assertion plausible. Consider again the two-alternative example for which the examinee thinks the chances are 7 in 10 in favor of alternative A. Recall that if his score is determined solely by the proportion of points assigned to the correct alternative, then his best strategy is to put zero on every alternative except the one he considers most likely to be correct; in which case, his expected score would be .70. To see that this is not true in the case of the scoring technique proposed by Roby, note that if the student puts all his stakes, say n points, on alternative A, his expected score will be:

$$7/10 \times (n/\sqrt{n^2+0^2}) + 3/10 \times (0/\sqrt{n^2+0^2}) = .70$$

If, however, he weights the alternatives in accordance with his judgement of what the chances really are, his expected score will be:

$$7/10 \times (7/\sqrt{7^2+3^2}) + 3/10 \times (3/\sqrt{7^2+3^2}) = .76$$

It also should be noted that the procedure permits the examinee to assign weights to the various alternatives in any way he sees fit. It may appear that there is some advantage in forcing the numbers assigned to the alternatives for a given question to add to one, inasmuch as we could then interpret them as probability estimates. We could have instructed the examinee to make his assignments so that they would indeed add to one; however, this is an unnecessary demand inasmuch as the score is unaffected by a change of scale. Moreover, if we wish to treat the assignments as probability estimates, we can easily normalize them by simply dividing each assigned number by the sum of the numbers associated with that question. If we did this, and replaced each of the original numbers with the resulting quotient (rounded off to two decimal places, say), then we could refer to each of the resulting numbers as a probability estimate (the

examinee's estimate of the probability that a particular alternative is correct), and to the collection of numbers associated with a given question as a probability vector. Table 2 shows the probability vectors that are obtained when this procedure is applied to the number assignments shown in Table 1. For the remainder of this note, it will be assumed that the examinee's number assignments are normalized in this way.

A further advantage of Roby's scoring technique is that it allows us to obtain not only an index of knowledge (what one knows about an item), but also one of confidence (what one thinks one knows). The index of knowledge has already been considered, and is given by Equation (1). (We have noted that this index gives the same result whether computed from the original number assignments or the normalized components of a probability vector. In both cases its range is from 0 to 1.)

The index of confidence (Roby refers to it as the "resolution index") is given by

$$c_j = \left(\sum_k^n p_{jk}^{2.} \right)^{-\frac{1}{2}}$$

where c_j represents the examinee's confidence in his answer to

Table 2. The "probability vectors" corresponding to the number assignments shown in Table 1.

Hypothetical answer:

A.	1	.00	B.	1	.00	C.	1	.20	D.	1	.00
	2	.47		2	1.00		2	.20		2	.00
	3	.00		3	.00		3	.20		3	.00
	4	.40		4	.00		4	.20		4	.00
	5	.13		5	.00		5	.20		5	1.00

the j^{th} question, and P_{jk} is the number assigned to the k^{th} alternative of the j^{th} item, after being normalized. Thus, Equation (2) is simply the denominator of Equation (1) after Equation (1) has been normalized.

As in the case of s_j, the maximum value of c_j is 1. It should be clear that $c_j = 1$ only if $x_{jk} = 1$ for one value of k and $\emptyset$ for all others. That is to say, in keeping with our intuitive notions about how an index of confidence should behave, it assumes its maximum value when the student has put all his bet on a single alternative. (Note that whether that alternative is correct or incorrect is irrelevant to this measure -- as it should be.) Unlike s_j, c_j cannot assume the value $\emptyset$. Its minimum value depends on the number of alternatives supplied with the question, and it is obtained when $x_{jk} = x_{jm}$ for all k and m; that is, the index gets its lowest value when the student assigns the same number to every alternative. Again, this is consistent with our intuitive ideas about confidence. The fact that the minimum value of the index depends on the number of alternatives is also in keeping with our intuitions about how a measure of confidence should behave: one should have less confidence in a guess among three equally likely alternatives than in a guess between two.

To summarize what has been said to this point, the scoring technique that Roby described has the following advantages: (1) the score received on an item reflects not only whether the examinee would have selected the correct answer in a conventionally administered test,* but also how much confidence he has in the item he would have selected; (2) it permits the examinee to assign weights to the various alternatives in any way he sees fit; (3) it has the property that the examinee serves his own interests best (maximizes his expected score) by reflecting his true opinions through his assignments; and (4) it makes an explicit distinction between the validity of one's belief about an item and the degree of confidence that one has in that belief. (This distinction between validity of and confidence in an opinion is similar in principle to that that signal detection theory makes between an observer's sensitivity and his decision criterion.)

The procedure also has another interesting property: The fact that the right-hand side of Equation (1) is the formula for

---

* I assume here that he would have selected the alternative to which he assigned the highest number

calculating the distance between two points in an n-dimensional Euclidean space suggests an elegantly simple geometric interpretation to this way of representing knowledge and confidence. Suppose we interpret the n-components of our probability vector as the coordinates of a point in n-space. It will be convenient to confine our attention for the moment, to the case of n = 3, that is to question for which only three alternatives are provided. Because the vector components are all non-negative and add to one, we need be concerned with only a small portion of this space; namely, the corner of a cube defined by the points (0,0,0), (1,0,0), (0,1,0) and (0,0,1). The triangular area in the plane of the last three points and bordered by the lines connecting those points defines all points with non-negative coordinates that sum to 1, and, hence, it represents all admissable 3-component probability vectors. In other words, any probability vector that can represent an answer for a three-alternative question will define a point in the shaded area of Fig. 1. We might refer to this area as the "belief surface."

To give the model an even more geometrical cast, we may represent a probability vector by a geometric vector emanating

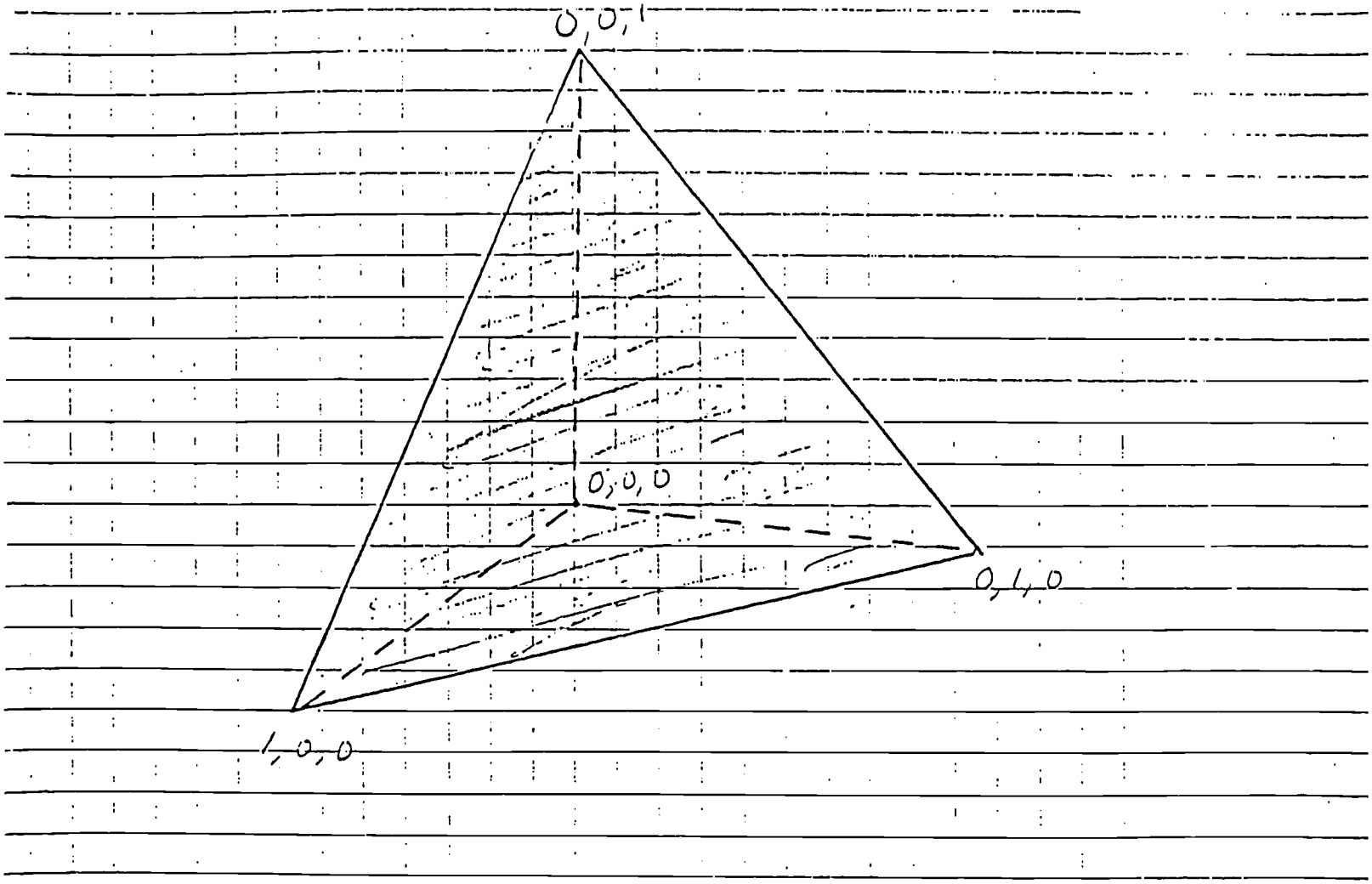

Fig. 1. The "belief surface" is the area of a plane defined by the points (1,0,0), (0,0,1) and (0,1,0) and bounded by lines joining these points.

from the origin and terminating at a point on the belief surface. Roby referred to such vectors as "B vectors." Each component of a B vector then is the projection of the vector on one of the axes of the space. This leads us to associate each axis with one of the alternative answers to the question, i.e., one of the "hypotheses." For convenience, we can let H_1, H_2, and H_3 represent the vectors originating at (0,0,0) and terminating at (1,0,0), and (0,1,0) and (0,0,1), respectively. The index of knowledge then, as given by Equation (1), is the cosine of the angle formed by the B vector and the axis representing the correct hypothesis (we might refer to the vector that represents the correct hypothesis as the T, or "truth," vector), the cosine being the ratio of the projection of the B vector on that axis to the length of the B vector. It is obvious that this angle must be between 0 and 90 degrees, and within that range the cosine of an angle varies inversely with the size of the angle, being maximum (1) when the angle is 0.

The index of confidence, given by Equation (2) has an equally simple geometric interpretation: the length of the B vector. Note that this length is minimum when the vector is equidistant from all three of the hypothesis axes; the point of

maximum uncertainty--or minimum confidence--is the centroid of the belief surface. The length of the vector increases as the point moves closer to any of the axes--not only the one representing the correct hypothesis. This representation of belief states provides a basis for hypothesizing a number of relationships between the validity of a belief and the confidence with which it is held, but that is irrelevant to the present context. What is relevant, is the possibility of exploiting this technique for the purpose of obtaining certain types of diagnostic information from aggregate test results.

To this end, it is convenient to think of the results of an examination in terms of a three-dimensional, say s x m x n, array of numbers, where s, m and n represent the number of individuals taking the test, the number of test items, and the number of alternatives per item, respectively. We will represent the number that the i^{th} examinee assigned to the k^{th} alternative of the j^{th} test item as p_{ijk}, and will assume that the assignments have been normalized so that

$$\sum_k p_{ijk} = 1$$

The test score for a student, say S_i, is given by

$$S_i = \sum_{j=1}^{m} \frac{p_{ijc_j}}{\left(\sum_{k=1}^{m} p_{ijk}^2\right)^{\frac{1}{2}}}$$

where p_{ijc_j} is the component of B_{ij} associated with the alternative that is defined as correct for item j.

The mean validity and mean resolution for the j^{th} item are, respectively

$$\mu_{V_j} = \frac{1}{s} \sum_{i=1}^{s} \frac{p_{ijc_j}}{\left(\sum_{k=1}^{n} p_{ijk}^2\right)^{\frac{1}{2}}}$$

and

$$\mu_{P_j} = \frac{1}{s} \sum_{i=1}^{s} \left(\sum_{k=1}^{n} p_{ijk}^2\right)^{\frac{1}{2}}$$

One of the functions of course-content examinations could be that of providing an instructor with feedback concerning how well various parts of the subject matter are getting across. An obvious thing to do in this regard is to compute a vector that represents the "class opinion" on every item and then to examine the "corporate" vectors (C vectors) in terms of both validity and resolution. It should be of some interest to the instructor to distinguish, for example, among items with respect to which the class is uniformed, and those with respect to which it is misinformed.

We may represent the corporate opinion with respect to the j^{th} item as

$$\bar{C}_j = \frac{1}{s}\left\{\sum_{i=1}^{s} p_{ij1}, \sum_{i=1}^{s} p_{ij2}, \cdots, \sum_{i=1}^{s} p_{ijn}\right\}$$

The validity and resolution of the corporate vector for the j^{th} item are then

$$v_{\bar{C}_j} = \frac{p_{ijc_j}}{|\bar{C}_j|}$$

and

$$R_{\bar{C}_j} = |\bar{C}_j|$$

where $|\bar{C}_j|$ is the length of $\bar{C}_j$.

We should note that the mean validity for the j^{th} item, μ_{v_j}, will not, in general, be the same as the validity of the j^{th} corporate vector, $v_{\bar{C}_j}$. Nor will the μ_{R_j} equal $R_{\bar{C}_j}$. The differences between these pairs of aggregate scores provide a rough indication of the degree to which examinees in a group vary with respect to their responses to an item. This is most easily seen in the case of the resolution, or confidence, measures. Consider, for example, the case in which two examinees both put

all of their bet on one alternative, but they do not both pick the same one. Both will have a high confidence index, and thus the mean resolution (considering only these two examinees) will also be high. The corporate vector that is computed from their responses however will have a moderate bet on each of two alternatives and thus will have a smaller resolution index. As may be seen from this example, the resolution of the corporate vector will tend to be smaller than the mean resolution, the magnitude of the difference depending on the degree of correspondence among the examinee's responses to the item. Note, however, that the difference is greatest when the examinees are differentially misinformed. A small difference could be obtained if they were well-informed, uniformed, or consistently misinformed. These states could be distinguished by comparison of the various aggregate measures. For example, high validity and high resolution can only be obtained if the class is uniformly well-informed. (Moreover, the corporate vector can be maximally valid only if all three of the other measures are also high.)

There are several ways in which a test administrator or teacher might use measures such as these as indicants of how

material is getting across. The most obvious thing to do is simply to order the items according to a variety of criteria. For example, one would probably like to see the items ordered in terms of any, or perhaps all, of the aggregate measures discussed above. In addition, one might wish to partition items on the basis of more than one measure simultaneously. For example, items for which the resolution of the corporate vector was high while validity was low would be of particular interest inasmuch as that combination indicates that the examinees are uniformly misinformed.

Some of the points that have been made above may be illustrated by reference to Table 3. The table shows several corporate vectors selected from the results of a college-class multiple-choice examination in which Roby's scoring technique was used. The numbers in the "key" column indicate which of the alternatives was correct for each item. Item 29 was one that the class, as a whole, knew quite well. All of the aggregate scores are high, and the corporate vector indicates that everyone was at least able to eliminate three of the alternatives from consideration. The class also did well as a group on item 43. In this case, however, the little uncertainty that there was was

distributed fairly evenly over the four incorrect alternatives, rather than being concentrated on one. Item 3 represents a case in which the class as a whole confessed to being relatively uniformed. Mean resolution was fairly low (the minimum for a five-alternative item is .45). Items 17 and 41 are clear cases of being misinformed; the evidence is the relatively high mean resolution scores accompanied by very low spherical-gain scores. Apparently, the students did not concur in their opinions on these items, however, in spite of the fact that many of them must have expressed high confidence in their answers. This we can infer from the differences between the mean resolution scores and the resolution of corporate vectors.

There are other aggregate measures besides those mentioned that might be used to advantage. A measure of consensus or coherence, for example, could be useful. Such a measure should reflect the degree to which the members of a group, say a class, share the same opinion (irrespective of its validity and resolution) concerning an item. (Note that this is different from the corporate vector, which, by itself, tells us nothing about consistency across the test takers.) One thinks first of some measure of variance or correlation for this application, but

Table 3. Illustrative items from five-alternative multiple-choice exam administered to college class. See text for explanation.

Item	Key	Corporate Vector					Spher. Gain Score of C Vector	Resolution of C Vector	Mean Spher. Gain Score	Mean Resolution
3	2	.275	.358	.233	.067	.067	.693	.517	.458	.718
17	2	.258	.133	.383	.154	.071	.261	.510	.195	.830
29	3	.125	.000	.875	.000	.000	.990	.884	.892	.976
41	1	.100	.350	.267	.100	.183	.201	.497	.121	.856
43	5	.017	.058	.042	.033	.850	.996	.854	.914	.898

such a simple thing as the difference between the mean resolution and the resolution of the corporate vector might suffice.

Another measure that is of some interest allows us to relate the opinion of an individual to that of a group as a whole. We might refer to such a measure as an index of concurrence, and one way to define it would be as

$$K_{ij} = \cos \theta_{ij}$$

where θ_{ij} is the angle between $\bar{C}_j$ add B_{ij} where

$$B_{ij} = \left\{ p_{ij1}, p_{ij2}, \ldots, p_{ijn} \right\}$$

i.e., the B vector for the i^{th} student for the j^{th} item. Concurrence then would vary between 0 and 1, being 0 when the agreement was minimal and 1 when it was complete. The mean concurrence

$$\mu_{K_{ij}} = \frac{1}{s} \sum_{i=1}^{s} K_{ij}$$

would be a candidate for the measure of consensus mentioned above.

My intent in the foregoing discussion was to demonstrate that multiple-choice tests have the potential to be much richer sources of information about what an individual or group does or does not know than they provide when administered in the usual select-one-alternative fashion. The question that naturally arises at this point concerns the feasibility of administering them. The hard part of this question relates to whether or not the answering technique could be made sufficiently understandable by test takers to be usable. A subordinate question of some relevance is the following: to what extent must examinees understand the theoretical basis for such scoring rules as the spherical-gain function in order to perform optimally on tests for which such scoring rules are used?

One plausible answer to the last question is: probably very little, if it is possible to provide test takers with immediate feedback, that is to give them their score on each item as soon as they complete that item. They would then learn from experience that they hurt themselves if they consistently express greater confidence than they really have. How long it would take for the point to be fully appreciated would probably vary from person to person, but one suspects that it would not take long in most cases.

Instantaneous scoring with a rule like the spherical-gain is not easily accomplished, however, unless one can administer the test by computer in real time. If one can administer the test by computer then the scoring problem is a trivial one, and feedback can be provided to the examinee graphically and in such a way as to make the implications of his number assignments apparent.

But what about when administration of the test by computer is not a possibility? Is it still possible to make effective use of such scoring techniques as the spherical-gain function? Some preliminary data gathered by the writer suggest that the answer may be yes. However, the data are too sparse to justify more than a qualified guess. To answer the question it would be necessary to administer a variety of tests under carefully controlled conditions. It would be of interest to attempt to determine not only whether or under what conditions performance converged on the optimal, but also how fast. Such knowledge assessment procedures would be of practical use in non-computerized environments, of course, only if techniques can be developed for bringing novice users to the point of efficient utilization fairly quickly.

References

Roby, T.B. Belief states and the uses of evidence. Behavioral Science, 1965, 10, 255-270.

Shuford, E.H., Albert, A. and Massengill, H.E. Admissible probability measurement procedures. Psychometrika, 1966, 31, 125-147.