ED 395 971	TM 025 109
AUTHOR	Emmerich, Walter
TITLE	Appraising the Cognitive Features of Subject Tests.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-89-53
PUB DATE	Nov 89
NOTE	52p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Achievement Tests; Classification; Coding; Cognitive Processes; *Cognitive Tests; *Evaluation Methods; Higher Education; Knowledge Level; Literature; *Profiles; Psychology; *Test Construction
IDENTIFIERS	Graduate Record Examinations; *Subject Content Knowledge

ABSTRACT

The aim of this study was to develop a procedure that could be used to appraise the cognitive features of subject (achievement) tests. Cognitive taxonomies and an accompanying coding scheme were developed and applied to the Graduate Record Examinations subject tests in Psychology and Literature in English. The taxonomies were based on the manifest cognitive content of the test items. Cognitive Demand and Aspects of Knowledge taxonomies were prepared, and two coders were trained to evaluate the tests using the developed taxonomies. The cognitive profiles of these two tests were found to be strikingly different, indicating that the taxonomies were operating as intended. Ways of improving and extending the procedure are discussed. The taxonomies and the coding procedure developed provide new instruments for appraising the cognitive demands and aspects of knowledge contained in the items of contemporary subject tests. (Contains 6 tables and 13 references.) (Author/SLD)

RR-89-53

RESEARCH REPORT

APPRAISING THE COGNITIVE FEATURES OF SUBJECT TESTS

Walter Emmerich

Educational Testing Service
Princeton, New Jersey
November 1989

APPRAISING THE COGNITIVE FEATURES OF SUBJECT TESTS

Walter Emmerich

Educational Testing Service

Princeton, New Jersey

APPRAISING THE COGNITIVE FEATURES OF SUBJECT TESTS

Walter Emmerich

Educational Testing Service

Princeton, New Jersey

Abstract

The study's aim was to develop a procedure that could be used to appraise the cognitive features of subject (achievement) tests. Cognitive taxonomies and an accompanying coding scheme were developed and applied to the GRE Subject Tests in Psychology and Literature in English. The cognitive profiles of these two tests were found to be strikingly different, indicating that the taxonomies were operating as intended. Ways of improving and extending the procedure are discussed.

Acknowledgments

This study was supported by the Program Research Planning Council of the Educational Testing Service.

Special thanks go to Peter L. Cooper, Mari A. Pearlman, Yvonne Sherrell, and Lynn Silverstein, who gave valuable feedback on the coding schemes and served so ably as coders. I am also grateful to Drew H. Gitomer, who served as reviewer and consultant at crucial points in the study. I also wish to thank LaVerne Shelton and Carol B. Tucker for their reviews of earlier versions of the taxonomies. I am especially grateful to Robert H. Ennis for his thoughtful comments on the 1987 version of this report. The administrative assistance of Annie M. Picard and Janet Zahn was much appreciated.

Introduction

Background

Achievement or subject tests are designed to measure specialized knowledge and skills. When a test committee is asked to formulate the content specifications for a subject test, the committee may consider the test's cognitive features, but primary emphasis usually is given to substantive coverage of the field. This ordering of priorities is appropriate, but, as a result, we have little systematic information on the cognitive features of subject tests. The major purpose of this study was to develop a procedure for identifying the cognitive features of subject tests.

Initially we explored the possibility of constructing a cognitive taxonomy based on the cognitive skills measured in aptitude testing, such as analogical and deductive reasoning. Although it was possible to classify subject-test items into categories defined by the aptitude item types of the GRE General Test, most of these categories were largely or totally inapplicable to subject-test items. And, although we realized that it might prove necessary to develop cognitive taxonomies on a test-by-test basis (e.g., Bowman & Peng, 1972; Teitelbaum, 1981), we were intrigued by the possibility of developing a general taxonomic scheme that could be used to compare a variety of subject tests.

Such a scheme should provide a rich set of descriptors of the cognitive features of subject tests. Also, the taxonomy would need to be linked to a system for coding test items having closed-option formats, since many subject tests are of this type. The coding system would need to be highly reliable in the sense that it yields high intercoder agreement levels, a rarely attained but desirable feature of cognitive taxonomies (Fleishman & Quaintance, 1984).

Test Selection

The intent was to investigate a small number of subject tests intensively at the item level. The tests were to be chosen from a test program that covers a variety of subject areas as well as a broad range of cognitive skills. Subject tests from the Graduate Record Examination Program (GRE) fit hese requirements. We selected one GRE subject test within each of two broad domains, Literature in English (Humanities) and Psychology (Social Sciences). These tests have relatively high candidate volumes. They also appeared likely to offer contrasts in the cognitive skills measured, important for an initial effort to develop a broadly applicable cognitive taxonomy. Although these particular subject tests emphasize verbal skills more than quantitative skills, we also kept in mind the goal of developing a procedure that might be applicable to quantitative subject areas as well. Two operational forms of each of the two subject tests (here called Forms G and H) were selected in order to determine whether the cognitive features of alternative forms of a test are reasonably stable.

Analysis of the Test Format

Constraints

A subject test is designed to incorporate representative portions of a subject area's corpus. However, closed-response item formats place constraints on the measurement of subject-matter expertise. Since our intent was to focus on the cognitive features of multiple choice items, we recognized that the resulting taxonomy or taxonomies could turn out to be incomplete.

The type of test item under consideration consists of a stem, a single correct option (key), typically four incorrect options (distracters), and perhaps additional stimulus material, such as a narrative (passage), a graph, and/or a qualifying phrase. Considered as a whole, these parts of an item comprise its manifest content. We adopted the working assumption that a coder would be able to identify at least some of an item's cognitive features from the item's manifest content alone, without having to rely on expert knowledge of cognitive science. It did not seem feasible to ask coders to identify the range of cognitive skills involved in processing an item to the point of selecting one of the options. Accomplishing this task would have called for coders having expert knowledge of cognitive science.

Even then there would have been little ssurance that many of the items in a subject test could be analyzed using the same cognitive model, or that the coders could have agreed on the most appropriate cognitive models. Our decision to limit the cognitive taxonomies and accompanying coding scheme to the manifest contents of test items was crucial for the study, and the ramifications of this decision are noted throughout this report.

Units of Analysis

Having decided to consider only an item's manifest content, it was still necessary to demarcate an appropriate unit (and/or subunits) at the item level. This requirement called for a method of analysis that could be applied generally to subject-test items. In a properly written subject-test item, the item's essential meaning and intent can be captured by joining the item's stem and key to form a declarative statement. Such statements were considered to be the basic units to which the developed taxonomy would apply. This decision regarding the basic unit of analysis did not rule out the possibility of also coding an item's (additional) stimulus material, if any, and so provision could be made for such codings as well.

For a variety of reasons, however, we decided to ignore an item's distracters. When the examinee selects among the options, a form of cognitive processing is likely to occur that is largely inaccessible to coders who are attending to an item's manifest content. Even if these processes were fully accessible to coders, an odd situation would arise if the distracters were to be included in the taxonomic codings. Joining an item's stem and key not only forms a declarative statement, but that statement can be taken to be true (in a properly constructed item). And such truth statements either can be found in the subject area's corpus or can be rigorously deduced from that corpus. The same cannot be said of a declarative statement that is formed by joining an item's stem with any of its distracters. Indeed, such statements can be taken to misrepresent the subject area's corpus. It seemed odd to characterize a subject test's cognitive features by including those parts of test items that do not belong to the corpus of the subject area. And, although all of the items in the present study were cast in a five-option multiple choice format, a closed-option test item need not include distracters at all, as in the case of a true-false format.

Manifest Cognitive Content

What does it mean, in more precise terms, to speak of an item's manifest cognitive content? Subject-test items appear to contain two relevant classes of content. The first class bears on the item's intent--what the examinee is asked to do. Obvious examples are forming a correct factual statement or drawing a correct inference from contextual material. We shall be using the term "cognitive demand" to refer to this type of accessible content.

The second class refers to the substantive information (knowledge) presented in the declarative statement formed by joining the item's stem and key. At least two pieces of substantive information are presented in an item, one in the stem and one in the key. These two pieces appear to be classifiable into one or another knowledge category, such as "theory," "relationship," or "entity," which, collectively, we are calling "aspects of knowledge." When the item contains additional stimulus material, such as a passage or a qualifying phrase embedded within the stem, the stimulus material also presents an aspect of knowledge.

In summary, an item's manifest cognitive content seems to be divisible into two broad classes, each calling for a taxonomy. One taxonomy would provide categories for identifying an item's cognitive demand, and the other would provide categories for identifying an item's aspects of knowledge. The latter would be applied twice to the item (stem and key), and to the item's stimulus material, if any, as well.

Variable and Invariant Classifications

Variations in item-writing standards, conventions, or styles can result in different versions of essentially the same item. Would the present approach to categorizing an item's manifest cog .tive content yield essentially the same taxonomic results despite variations in the way that the item is written? We sensed that it would in most cases, although the issue was not investigated systematically. Nevertheless, it is informative to consider hypothetical examples of how variations in an item's wording do and do not result in altered categorizations, especially with regard to aspects of knowledge.

Consider the following declarative statement, formed by joining an item's stem and key: "Freudian theory (stem) relates deprivation to fixation (key)." In this example, the aspect of knowledge contained in the stem can be designated as "theoretical", whereas the aspect of knowledge contained in the key can be designated as "relationship". Note that the item could have been written so that the same two aspects of knowledge are reversed: "A relationship between deprivation and fixation (stem) is postulated in Freudian theory (key)." This situation is reassuring because essentially the same taxonomic classifications would occur despite a variation in how the item is written. But consider what can happen when the item is written to include a qualifying phrase: "According to Freudian theory (qualifying phrase), deprivation is related to (stem) fixation (key)". In this version, the initial qualifying phrase can be designated as "theory" and the stem as "relationship", and so again there would be these two codings. However, the key of this third version contains the word "fixation", an additional aspect of knowledge, one that can be designated as "entity". Since this particular category was absent in the other versions, we see how a transformation in wording can alter the categorizations.

Keeping this matter in perspective, however, such indeterminancies probably would introduce negligible random errors into the item classifications, a small price to pay given our wish to capture the multiple aspects of knowledge that are to be found in an item's stem and key (and stimulus material or qualifying phrase, if any). Preserving these structural featu es was considered to be important because item variations in psychometric properties, such as difficulty level, might be partially determined by where the aspect of knowledge is located in the item structure (e.g., stem vs. key), and perhaps also by the interaction of this factor with the particular aspect of knowledge involved (e.g., "entity" vs. "theory").

Development of the Taxonomies

Background

The taxonomy for the cognitive domain presented in the classic work, Taxonomy of Educational Objectives (Bloom, Englehard, Furst, Hill, & Krathwohl, 1956), turned out to play an important role in the present study. Even though the efforts of Bloom et al. antedated the modern era of cognitive science, their taxonomy was sufficiently broad to provide a bench marl for determining whether we were overlooking any important cognitive categories. We also were impressed with how often the taxonomic categories developed by Bloom and his colleagues seemed to be applicable to the manifest contents of today's subject-test items. In this regard, the distinction we have drawn between an item's cognitive demand and its aspects of knowledge is similar to that drawn by Bloom et al. between "Intellectual Abilities and Skills" and "Knowledge", respectively. As will become apparent, however, our approach also differs in important ways from that of Bloom et al.

Our procedures in developing the taxonomies are described below. These procedures were systematic, for the most part, and they were informed by reviews of the literature as well as by the helpful comments of internal consultants and a coder in the course of developing the coding procedure. Nevertheless, we do not view the developed taxonomies as final products. Different intuitions would have yielded other versions, and our versions probably will require modification in the light of findings based on their use, such as the findings presented later in this report.

Level of Abstraction (Horizontal Structure)

It was important to identify terms (or phrases) that refer to specific cognitive demands and aspects of knowledge in subject-test items. These terms would be at different levels of abstraction, however, and so there was the need to determine how many levels to represent when structuring the selected categories in terms of level of abstraction. A previous study had suggested that three levels would suffice (Metfessel, Michael, & Kirsner, 1969). The levels that we eventually adopted were: (1) A term signifying the major category, accompanied by a brief descriptive or clarifying statement; (2) A defining term and/or phrase for each of the subcategories of a major category. (For some major categories this level was omitted.); (3) A set of terms and/or phrases at a more concrete level. The latter sets of terms were to be nested within subcategories.

Inspection of Table 1 reveals the horizontal structuring of each of the five categories selected for the Cognitive Demand Taxonomy, and inspection of Table 2 reveals the horizontal structuring of each of the six categories selected for the Aspects of Knowledge Taxonomy.

Selection of the Categories (Vertical Structure)

Selection of the categories themselves was guided by examining several sources in the literature, including the Bloom et al. taxonomy, a subsequent effort to further specify the educational objectives implied by that taxonomy (Metfessel, Michael, & Kirsner, 1969), taxonomically-oriented works from the cognitive sciences (e.g., Fleishman & Quaintance, 1984; Guilford, 1967; Messick, 1984; Mosenthal, 1985; Sokal, 1977), and a report on the reasoning processes that might suitably be measured by the GRE General Test (Tucker, 1985).

The taxonomic categories were constructed by means of an iterative procedure in which lists and groupings of cognitive terms were formulated, reviewed, and reformulated. In the first

TABLE 1

COGNITIVE DEMAND TAXONOMY

MAJOR CATEGORY	CODE	SUBCATEGORY	RELEVANT TERMS
Correctly Synthesize Components into a Recognized Pattern	Syn-O	Organize:	Sequence, order, categorize, classify, subsume
	Syn-I	Integrate:	Coordinate, unite, connect, assemble, harmonize, combine
	Syn-R	Reorganize:	Rearrange, reorder, reclassify, recombine, replace (element), reverse (e.g., figure-ground, cause-effect)
Correctly Support or Weaken a Claim, Procedure, Outcome	SW-S	Substantiate:	Demonstrate, prove, confirm, verify, document, counter (alternative)
	SW-C	Constrain:	Limit, qualify, delimit, contain
	SW-N	Negate (or nullify)	Cast doubt on, critique, undermine, contradict, counter, exclude, disprove, falsify, note flaw
Correctly Analyze Information	A-D	Distinguish:	Differentiate, contrast
	A-I	Infer (tight):	Conclude, induce, deduce, diagnose, extrapolate, interpolate
	A-G	Generalize:	Plausibly universalize, find common element or ground
	A-S	Simplify:	Extract, purify, weed out, dissect, decompose, abstract
	A-P	Problem-Solve:	Calculate, measure, test, observe, inquire, experiment, unravel, trouble-shoot, investigate
	A-E	Evaluate:	Judge relative merits, decide, appraise, weigh, compare
	A-R	Resolve:	Equilibrate, balance, counterbalance, satisfice, optimize
	A-T	Transfer:	Analogize, apply, carry over
Identify a Correct Piece of Relevant Information NOT Given	I-I	Recall:	Recognize, name, discern, locate, match
	I-D	Define:	Operationalize, concretize, spell out
	I-E	Exemplify:	Illustrate
	I-C	Clarify:	Elucidate, explain, explicate
Accurately and faithfully Restate GIVEN Information	R-D	Depict:	Pinpoint, characterize, portray
	R-S	Summarize:	Paraphrase
	R-T	Translate:	Literally code or decode

TABLE 2

ASPECTS OF KNOWLEDGE TAXONOMY

MAJOR CATEGORY	CODE	SUBCATEGORY	RELEVANT TERMS
Language: Content of communication or expression	L-T	Term(s):	Sign, symbol; single word or short phrase; single figure; simple equation
	L-M	Meaning: (Single)	Standard meaning or definition; explication of meaning from context (e.g., from discourse)
	L-D	Discourse:	Narrative or exposition in a paragraph, sentence, or long phrase; set of figures, graphs, or equations; combination (e.g., exposition plus graph)
Entity(ies): Real, or to be taken as real (e.g., fictional account)	E-T	Tangible(s): (Concrete, in past, present, or future; Recurrence over space and/or time NOT implied)	Occurrence, outcome, product, event, object, name, stimulus, response, act, date, time, person, place, source, author; special condition, circumstance, datum (e.g., statistical), observation, score, measure, magnitude, quantity
	E-C	Category(ies): (A class of tangibles; Recurrence over space and/or time IS implied)	Type, kind, manner, tone, style, species, topic, domain, classification, subclass, diagnosis, division, subset, prototype, theme, state, condition, phase, stage, era, tradition, genre, body of work
Relationship(s): Between entities, whether tangible, categorical, or linguistic	R-I	Individual Relationship:	Principle or generalization, syntactical rule, if-then statement, association, contingency, correspondence, connection, correlation, influence, resultant, simple sequence or process, cause-effect, absence of association (or of any of the above), independence, null effect
	R-S	System of Relationships:	Pattern, form, syntax, order, organization, taxonomy, hierarchy, network, subsystem, genealogy, complex series or process, proof, dialectic, chronology, syndrome
Procedure(s): Step or steps toward a goal	P		Method, means, usage, format, way, technique, design (e.g., experimental), procedural control,, treatment, routine, plan, heuristic, procedural rule, method of analysis (e.g., data, chemical)
Criterion: Evaluative standard	C		Qualitative or quantitative standard of acceptability, merit; or of unacceptability: Presence of absence of relevance, reliability, veridicality, plausibility, appropriateness, logicalness, reasonableness, coherence, consistency, validity, completeness, comprehensiveness, generative power
Theory: Unproven, or to be taken as unproven	T		Recognized, but not fully accepted belief or organized set of beliefs: Hypothesis, model, paradigm, formulation, approach, conceptualization, viewpoint, claim, perspective, conjecture, school, speculation, attitude, opinion

cycle, attention was given to the terms suggested by the above literature as well as by disclosed sample items from a variety of GRE Subject Tests. A cognitive term was considered regardless of whether it appeared to be applicable to current subject-test items. Also, the initial focus was on compiling reasonably complete lists of terms for each of the taxonomies rather than on assigning the terms to their proper levels in the hierarchy.

With regard to the Cognitive Demand Taxonomy, the first cycle yielded over a hundred terms, many of which seemed suitable for the third (concrete) level of the horizontal structure discussed earlier. These terms were arranged into nineteen groups, and these groups were given the following tentative headings: Affirm, Augment, Constrain, Distinguish, Evaluate, Extend, Fit, Identify, Infer, Integrate, Negate, Organize, Reduce, Reorganize, Represent, Resolve, Restate, Solve, and Transfer. At this point it became apparent that the tentative headings could themselves be grouped into the five major categories presented in Table 1. Having thus established the major categories of the Cognitive Demand Taxonomy, the tentative headings and terms they subsumed were then modified and/or rearranged, when required, to ensure that the various levels of the hierarchy within each category would be nested coherently. Subsequently, the taxonomy was fine-tuned in the course of developing the coding procedure.

In developing the Aspects of Knowledge Taxonomy, examination of sample items from a variety of GRE subject tests often suggested terms that seemed to fall at the higher levels of the horizontal hierarchy. In the first cycle, eleven tentative headings were used to subsume over a hundred terms, only some of which were generated from the sample items. The number of headings was reduced and the six major categories presented in Table 2 were established. In general, these major categories resemble those of Bloom, et al., although in some instances these categories as well as the subcategories and concrete terms were updated by incorporating those features of Mosenthal's (1985) analysis of expository discourse that seemed applicable to a multiple choice test format.

Having established the major categories for the Aspects of Knowledge Taxonomy, the subcategories were determined by reconsidering the residual headings and full list of terms. The next step was to ensure that the concrete terms (third level) were sufficiently varied to be broadly applicable to a variety of subject areas. This taxonomy also was fine-tuned in the course of developing the coding procedure.

Unit Codings

The decisions on the units of analysis had set the stage for applying the developed taxonomies to a subject-test item. The Cognitive Demand Taxonomy was to be applied to the declarative statement formed by joining the stem with its key, and the Aspect of Knowledge Taxonomy was to be applied separately to the stem and to the key, and also to the stimulus material or qualifying phrase, if any.

It was possible that more than one cognitive demand would be called for, as in the case of items demanding both factual knowledge (Identify) and inference (Analyze). Considering just these two categories, if we had allowed for double codings of equal weight, there would have been three possible codings of a given item: Identify only; Analyze only; Analyze and Identify. Although coders might choose reliably among these three alternatives, there was doubt that they could do so on the basis of the item's manifest content alone. Also, if provision had been made for both single and double codings, based on all five major categories, the coders would have faced a choice among fifteen alternatives even when coding an item's major category! Since this procedure probably would have jeopardized the reliability of the coding system, there was good reason to limit a coder to exactly one coding operation per item. The ordering rule devised for this purpose is discussed in the next section.

In the case of the Aspects of Knowledge Taxonomy, the item was to be divided into the two (or three) structural subunits, and the complexity of the information presented in a subunit could then be reflected by the applicable taxonomic category. Specifically, the information contained in the subunit could be as simple as a term or entity, it could consist of linked terms and/or entities, as in a relationship, or it could be as complex as a literary narrative or a description of a scientific procedure. It appeared, then, that a single coding of each subunit would not only be practicable, but that it could be used to characterize the complexity of the information contained in the subunit.

Ordering the Major Taxonomic Categories

In coding an item's major category, the coder would still be faced with a choice among five alternatives when applying the Cognitive Demand Taxonomy (Table 1), and with a choice among six alternatives when applying the Aspects of Knowledge Taxonomy (Table 2). Thus, there was still the need to establish procedures for making these choices. Moreover, the procedures should be based on suitable rationales for ordering the major categories within each of the taxonomies.

The search for ordering principles initially carried us back to the discussion of this matter by Bloom et al. (1956), who offered the following tentative hierarchical ordering of their six major classifications (from bottom to top): Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. One of their principles, which might be called the "cumulative principle", was that "the objectives in one class are likely to make use of and be built on the behaviors found in the preceding classes in this list" (p. 18). A related principle, which might be called the "complexity principle", suggested that the listed classes are orderable in terms of their cognitive complexity. Putting the two principles together, an item that calls only for Knowledge and Comprehension would require the cumulation of two classes at a relatively low level of complexity, whereas an item that calls for Evaluation would require the cumulation of six classes at the highest level of complexity. Bloom et al. also cited initial evidence suggesting that this hierarchy is associated with item difficulty (pp. 18-19).

These authors recognized that their formulation probably would be revised or replaced in the light of further inquiry. For example, the major categories we have suggested for each of the two taxonomies only partially overlap with their categories. More fundamentally, their hierarchical concept has not stood up well. For one thing, the difficulty of a subject-test item is probably determined in large part by the degree to which the item requires specialized knowledge or vocabulary, attributes that are at the lowest level in the Bloom et al. hierarchy. Nevertheless, their formulation remains instructive because it illustrates the sorts of strong ordering principles that one would like to be able to incorporate into cognitive taxonomies.

Although we were not in a position to offer strong ordering principles, some sort of ordering rule was required in the case of the Cognitive Demand Taxonomy because of the possibility, noted earlier, that an item might call for more than one cognitive demand. We sought an ordering rule that would enable the coder to select exactly one major category per item, without implying that the item made only one cognitive demand, and without requiring the coder to judge the relative weights of two or more applicable categories. In short, we had hoped to resolve a methodological problem without resurrecting a strong concept of hierarchy--in the Bloom et al. sense.

To this end, an ordering principle was adopted that bears on how far the examinee is asked to go beyond the information given in the item's stem (and stimulus material, if any) to arrive at the key (Bruner, Goodnow, & Austin, 1956; Sigel & McGillicuddy-Delisi, 1984). From this perspective, the following ordering of our major categories (from top to bottom) appeared to be reasonable: Synthesize, Support-Weaken, Analyze, Identify, and Restate. Coders are instructed to start at the top of the list of major categories (Synthesize), decide on whether that major category applies, and to go no further if it does apply. If that major category does not apply, the coder moves to the next major category in the list (Support-Weaken), and the coder goes down the list until the item's major category is chosen. Thus, the coding procedure directs the coders to choose, for a particular item, that major category representing the greatest possible distance from the information given. This procedure entailed the risk that coders would tend to overestimate the distance, but this risk seemed preferable to that of systematically underestimating the cognitive demands of subject tests.

In the case of the Aspects of Knowledge Taxonomy, the information contained in the subunit (its meaning and level of complexity) appeared to be sufficient for choosing among the six major categories. Consequently, this taxonomy's major categories did not need to be ordered in any particular fashion, at least for coding purposes. In performing a coding operation, the coders were instructed to consider all six major categories before selecting the one that applies best to the subunit.

Of course, deeper issues remain regarding how these major categories might be ordered. What we did was to separate the coder's task from the more fundamental question of how aspects of knowledge are in fact organized. A review of the Aspects of Knowledge Taxonomy suggests that its major categories may be orderable with regard to the following: (1) Complexity of information (as discussed on page 15); (2) A dimension extending from abstract representation to direct experience; and, (3) A dimension extending from that which is firmly established to that which is purely speculative. Such multidimensionality suggests that a proper ordering of the Aspects of Knowledge categories probably would entail a non-hierarchical scheme, perhaps varying among subject areas and/or among families of subject areas.

Selection of the Coders

Criteria

We had assumed that the coding task would require mastery of ordinary language and academic work skills at the baccalaureate level. The primary question that remained was the level of subject-matter expertise that would be required. Several levels were considered initially: (1) nationally recognized experts in the subject area, such as test committee members; (2) people who have attained an advanced degree (e.g., Ph.D.) in the subject area; (3) graduate students in the subject area; (4) people who have attained a bachelor's degree in the subject area, but who have not sought a more advanced degree in that area; (5) people who have received a bachelor's degree and who majored in a subject area other than that under investigation.

In deciding on an appropriate level, potential long-term costs as well as workability were considered. With regard to the cost factor, the intent was to establish a procedure that could be used periodically to monitor the cognitive features of varied and perhaps numerous subject tests. From this standpoint, levels 4 and 5 would be most desirable. The situation was much less clear with regard to workability, however. Initially we had assumed that considerable subject-matter expertise would be required, suggesting that level 2 would be appropriate. However, in the course of analyzing the item format and constructing the taxonomies, it became apparent that even levels 4 and 5 would be reasonable alternatives.

Yet the possibility remained that level of expertise would influence the codings. For example, a subject-matter expert might tend to see less in an item's cognitive demand than would a novice, perhaps especially when the distracters are ignored. Interestingly, Bloom et al. (1956) were aware of a related complication. They noted that a student's past experience in solving a class of problems can determine the level of processing required by the student, and they were particularly concerned about this possibility when applying their taxonomy to test exercises (p. 16).

Systematic differences between novices and experts in how they code subject-test items probably are interpretable in terms of different judgmental processes resulting from the coder's knowledge and its organization (e.g., Glaser, 1984). From this standpoint, observed differences between novices and experts would be interpreted in terms of available knowledge structures, not in terms of the "biased judgment" of either the novice or expert. With regard to an operational coding scheme, however, the question of bias would arise, and one would assume that the judgments of experts are less subject to bias than those of novices. Although this assumption seemed reasonable, we asked ourselves whether the selection of coders at levels 4 or 5 would be likely to result in serious bias.

In particular, we considered the possibility that subject-matter experts would tend to simplify an item's cognitive demand relative to subject-matter novices, in which case the codings of subject tests by novices would tend to overstate the cognitive demands placed on examinees by subject tests.

Our initial examination of sample items from subject tests suggested that such a bias would be most likely to occur in ambiguous instances, but, that for many items, the cognitive demand is not ambiguous. Also, many items seemed to place rather limited cognitive demands on examinees. Consequently, it seemed unlikely that novices who are trained to focus on an item's manifest content would stretch matters to the point where the cognitive demands of subject tests are seriously overstated.

On the basis of these considerations, we took a calculated risk and recruited the two coders at levels 4 and 5.

The Coders

The two recruited coders had attained an undergraduate degree but not a more advanced degree, and neither expressed intent to pursue graduate study in either Literature or Psychology, important for test security. One of the coders was a Research Assistant at ETS who had recently graduated with a major in Sociology. The other coder was hired as a temporary Research Assistant, had majored in Psychology in college, and had worked previously as a temporary employee at ETS.

Coding Procedure

Preparing an Item

Several preparatory steps are taken by the coder before the actual coding of an item begins. After being provided with the item's key, the coder casts the item into a declarative statement that joins the stem with the key. Sometimes the item already appears in this form in the test booklet, but often it must be recast, as in the case of a closed stem. Additional rewordings may also be required, as when there is reference to separate stimulus material (e.g., "In the passage above,...") or when a qualifying phrase is embedded in the stem (e.g., "According to Freudian theory..."). In such cases, the reference to the stimulus material or to the qualifying phrase is placed as the first clause in the declarative statement.

Whether the item's wording is recast or not, the coder writes the entire declarative statement on a sheet of paper. This written version of the item consists of, consecutively: a notation regarding the stimulus material (if any), designated as the "C" subunit; the stem, designated as the "X" subunit; the key, designated as the "Y" subunit. The coder compares the written declarative statement and its parsing (into the C, X, and Y subunits) with the original version in the test booklet in order to ensure that no changes in meaning have occurred in preparing the item for coding.

It became apparent that the key of an item is not always sufficient to define the Y subunit meaningfully. As an illustration, consider still another version of our earlier hypothetical item: "Freudian theory relates (stem) deprivation to fixation (key)." Here the coder would include the word "relates" as part of the Y subunit (key). On occasion, more than one term in the stem might be required to fill out the Y subunit of an item, and so we needed to decide how far to go in moving backwards when selecting words from the stem. When this problem arose, coders were to define the X subunit of the item as the first meaningful subunit in the stem (other than a C subunit) that is codable in applying the Aspects of Knowledge Taxonomy. The coder would then include the remainder of the stem, together with the key, in the Y subunit of the item.

Item Codings

The first coding of the item applies the Cognitive Demand Taxonomy to the declarative statement as a whole, although the coder is encouraged to use the subunits as an aid in identifying the item's cognitive demand. The cod^r is encouraged to ask: "What is the examinee being asked to do in going from the X subunit to the Y subunit?" In the remaining codings, either two or three in number, the coder applies the Aspects of Knowledge Taxonomy consecutively to the C subunit (if any), to the X subunit, and to the Y subunit.

The coder chooses the single applicable major category and the most applicable subcategory within the major category. The coder records these selections consecutively on prepared code sheets as well as next to the coder's written version of the item. However, in arriving at a code, the coder is allowed to move flexibly back and forth within the horizontal hierarchy. We believe that this feature strengthened the coding procedure because it encouraged the coder to ground even the major category coding in a concrete term provided at the third level of the hierarchy. Tc facilitate proper tracking by the coder, the subcategory code carries the designation of its major category as well (see Tables 1 and 2). Coding at the subcategory level is thus an integral part of the coding procedure, regardless of whether the investigator intends to proceed to the subcategory level. In the present study we carried out the analyses at the primary category level but not at the subcategory level.

Coder Training

After a brief orienting session with the trainer, the coder becomes familiar with three sets of materials: (1) The descriptive booklets for the relevant subject tests, containing sample disclosed items that are to be used in training; (2) Sets of instructions for coders; (3) The taxonomies themselves (Tables 1 and 2). As part of the familiarization process, the coder is asked to try out a few codings of the sample items, and the initial problems encountered are discussed with the trainer. In this study the trainer was the author.

In the second phase of training the coder continues to code sample items, followed by discussion with the trainer. This procedure is repeated until the trainer and coder feel reasonably confident that the coder has mastered the procedure. In the present study this phase of training also provided feedback to the trainer for purposes of refining the taxonomies, the coding procedures, and the code sheets.

In the third phase, pairs of coders work independently on the same items and then meet with the trainer to compare codings and to resolve any discrepancies in the procedures followed by the two coders. This process is repeated until intercoder agreement levels reach designated levels. In the present study, these levels were not actually specified because procedural refinements were still under consideration. However, there came a point when the investigator sensed that further training sessions and procedural refinements would not add appreciably to intercoder agreement levels. At that point the collection of study data began.

The time and cost requirements for training depend on a variety of factors, such as the number of coders and the particular subject tests involved. The coding task is intricate and intensive, mastery requires repeated feedback from the trainer, and the training period should not be overly compressed. Nevertheless, training does proceed steadily. As a rough estimate, the two coders were trained in about 40 hours including discussion time.

Codings for the Study

Each of the two forms of the Literature test consisted of 230 items and each of the two forms of the Psychology test consisted of 200 items. Thus, a total of 860 items were to be coded in order to provide data for comparing the cognitive features of the two subject tests. After receiving instructions on how to ensure test security, each of the two coders was assigned a total of 530 item in accordance with the design considerations discussed in the next section.

A particular coding assignment consisted of a block of 50 to 115 items from one form, enabling the coders to exchange test booklets between assignments, when appropriate. After coding the items within a block, a coder reviewed the codings for that block and made any final revisions. A coder was to complete all of the Psychology test codings before proceeding to the Literature test codings. This sequence was arbitrary, but the procedure of working within a given subject test until all of the codings for that test were completed facilitated the coder's task. However, the switch from one subject test to the other did result in some initial problems. Further discussion with the trainer occurred at the time of the transition, although the trainer attempted to keep these discussions at a general level so as not to inflate the estimates of intercoder agreement. It would be desirable to hold a "refresher" training session at those times when a coder is about to switch to a different subject test.

Because intensive cognitive work is required, the optimal coding period appeared to be about three consecutive hours per day (including rest periods), during which about 25 items were coded.

Profile Analyses

Intercoder Reliability

In determining intercoder reliability, the two trained coders independently coded the first fifty items in Forms G and H of both the Psychology and Literature tests. The reliability sample included 200 items, corresponding roughly to the number of items that appear in a single form of a GRE subject test.

The percentages of agreement for the major categories were 77% for the Cognitive Demand Taxonomy, 69% for the Aspect of Knowledge Taxonomy applied to the item stems (X), and 75% for the Aspect of Knowledge Taxonomy applied to the item keys (Y). Later in this report we discuss how these agreement levels might be raised. For the initial profile analyses, however, even moderate amounts of measurement error (coder disagreement) could be tolerated because the unit of analysis was the test form, an aggregate in this case of 200 or more items, not a single item or a small set of items that is expected to be taxonomically homogeneous.

Nevertheless, it was important to check on whether the amount of coder disagreement might be great enough to obscure a test form's profile. The relevant data are presented in Table 3. Regarding random coding errors, a randomly determined taxonomic profile would be rectangular in shape. Inspection of Table 3 reveals that the category distributions were not rectangular.

Table 3

Coder Profiles for the Set of 200 Reliability Items

Taxonomy	Category	Percentage Coder A	 Coder B
Cognitive Demand	Synthesize	0.0	0.0
	Support-Weaken	0.5	0.5
	Analyze	51.0	42.0
	Identify	46.0	56.5
	Restate	2.5	1.0
df=1, X^2 =3.48, *p* >.05		100.0	100.0
Aspects of Knowledge Applied to Item Stems	Language	22.0	20.5
	Entity	42.5	53.0
	Relationship	24.5	18.0
	Procedure	2.5	1.5
	Criterion	1.5	0.0
	Theory	7.0	7.0
df=3, X^2 =4.33, *p* >.20		100.0	100.0
Aspects of Knowledge Applied to Item Keys	Language	17.5	15.0
	Entity	60.0	56.5
	Relationship	16.0	23.0
	Procedure	2.5	1.5
	Criterion	1.0	1.0
	Theory	3.0	3.0
df=3, X^2 =3.10, *p* >.30		100.0	100.0

The possibility that systematic bias by a coder would obscure one or another profile was tested by comparing the two coders' profiles on the 200 items. As seen in Table 3, the two coders' profiles were quite similar within a given taxonomic application, and none of the three comparisons reached the 5% level of significance. (Chi Square tests are reported here and in the tables that follow. A category was excluded from an analysis if its expected frequency was too low.)

It was also important to ensure that systematic coder differences, even apparently minor ones, not be confounded with our estimates of the profiles for a particular test form. Consequently, the two coders' item codings were counterbalanced within forms when aggregating the taxonomic profiles for a given form. Specifically, each of the coders judged half of the items within Form G and within Form H of both the Psychology test (200 items per form) and the Literature test (230 items per form).

Stability of Test Forms

The first substantive question was whether two current forms of a subject test yield similar profiles on the developed cognitive taxonomies, suggesting within-test stability.

The profiles for the two forms of the Psychology test are presented in Table 4. Although there were minor differences between the profiles, none of the three profile comparisons reached the 5% level of significance. The profiles for the two forms of the Literature Test are presented in Table 5.

Table 4

Taxoncmic Profiles for Two Forms of the Psychology Test
(200 Items Per Form)

		Percentage	
Taxonomy	Category	Form G	Form H
Cognitive Demand	Synthesize	0.0	0.0
	Support-Weaken	0.5	1.5
	Analyze	36.0	35.5
	Identify	63.5	63.0
	Restate	0.0	0.0
		-----	-----
df=1, X^2 =0.01, p >.90		100.0	100.0
Aspects of Knowledge Applied to Item Stems	Language	3.0	5.5
	Entity	40.0	39.5
	Relationship	35.5	31.0
	Procedure	5.5	10.0
	Criterion	1.0	1.0
	Theory	15.0	13.0
		-----	-----
df=4, X^2 =4.98, p >.20		100.0	100.0
Aspects of Knowledge Applied to Item Keys	Language	1.0	5.0
	Entity	50.5	49.5
	Relationship	35.0	37.0
	Procedure	3.5	3.0
	Criterion	2.0	1.0
	Theory	8.0	4.5
		-----	-----
df=4, X^2 =7.49, p >.10		100.0	100.0

Table 5

Taxonomic Profiles for Two Forms of the Literature Test
(230 Items Per Form)

		Percentage	
Taxonomy	Category	Form G	Form H
Cognitive Demand	Synthesize ------	0.0	0.0
	Support-Weaken --	0.0	0.0
	Analyze ---------	70.4	70.0
	Identify --------	27.0	27.0
	Restate ---------	2.6	3.0
		100.0	100.0
df=2, X^2 =0.08, p >.95			
Aspects of Knowledge Applied to Item Stems	Language --------	36.5	37.0
	Entity ----------	59.6	59.1
	Relationship ----	3.5	3.0
	Procedure -------	0.0	0.9
	Criterion -------	0.0	0.0
	Theory ----------	0.4	0.0
		100.0	100.0
df=2, X^2 =0.08, p >.95			
Aspects of Knowledge Applied to Item Keys	Language --------	33.9	30.9
	Entity ----------	55.7	60.9
	Relationship ----	10.0	6.1
	Procedure -------	0.4	1.7
	Criterion -------	0.0	0.0
	Theory ----------	0.0	0.4
		100.0	100.0
df=2, X^2 =3.01, p >.20			

Again, although there were some minor differences in the profiles, none of these comparisons were statistically significant. It would appear that the two subject tests are quite stable in their taxonomic profiles, at least with regard to Forms G and H.

Profile Contrasts

It seemed likely that the Psychology and Literature tests would exhibit quite different taxonomic profiles. Indeed, these comparative analyses provide a major basis for evaluating the developed cognitive taxonomies. If the profiles are highly similar, that would suggest that our taxonomies are insufficiently sensitive to differences among knowledge domains. On the other hand, the occurrence of plausible profile differences between Psychology and Literature would provide initial evidence that the developed taxonomies are operating as intended.

Because the analyses of the form differences were non-significant (Tables 4 and 5), the items from forms G and H were aggregated (within each subject test) before the between-test profile analyses were conducted.

Table 6 presents the taxonomic profiles for each of the two tests. In all three instances the profiles are strikingly different (ps <.001). With regard to Cognitive Demand, the balance between Analyze and Identify is virtually reversed between the two subject tests. This contrast reflects an obvious

Table 6

Taxonomic Profiles for Two Tests:
Psychology (400 Items) and Literature (460 Items)

		Percentage	
Taxonomy	Category	Psych.	Lit.
Cognitive Demand	Synthesize	0.00	0.00
	Support-Weaken	1.00	0.00
	Analyze	35.80	70.20
	Identify	63.20	27.00
	Restate	0.00	2.80
df=2, X^2 =122.6, *p* <.001		100.00	100.00
Aspects of Knowledge Applied to Item Stems	Language	4.25	36.74
	Entity	39.75	59.35
	Relationship	33.25	3.26
	Procedure	7.75	0.43
	Criterion	1.00	0.00
	Theory	14.00	0.22
df=4, X^2 =323.9, *p* <.001		100.00	100.00
Aspects of Knowledge Applied to Item Keys	Language	3.00	32.39
	Entity	50.00	58.26
	Relationship	36.00	8.04
	Procedure	3.25	1.09
	Criterion	1.50	0.00
	Theory	6.25	0.22
df=4, X^2 =211.4, *p* <.001		100.00	100.00

difference in test construction. For Literature, many of the items are accompanied by stimulus material, usually in the form of a narrative calling for interpretation. In sharp contrast, the Psychology test uses stimulus materials much less frequently. However, it is unlikely that this difference arose solely from different conventions in test construction. Rather, the contrast appears to reflect a deeper difference in subject matter, suggesting that the cognizant test committees might judge that their respective tests each contains an appropriate balance between Analyze and Identify.

The contrasts between Psychology and Literature on the Aspect of Knowledge Taxonomy are no less striking and involve more than two categories. For example, about a third of the Literature items refer to Language (in either the stem or key), whereas less than five percent of the Psychology items refer to Language. The remaining contrasts are of particular interest because they are less predictable. References to a Relationship are much more frequent in Psychology (about a third) than in Literature (less than a tenth), references to Theory are not uncommon in Psychology, but are negligible in Literature, and references to Procedure are somewhat more frequent in Psychology than in Literature.

These differences in test profiles probably did not arise because the taxonomy itself is biased in favor of social scientific categories. The Relationship category was carefully defined to include any type of connection (or lack of connection) between entities, whether tangible, categorical, or linguistic. The Theory category was defined broadly to include any reference to a school of thought, belief, or conjecture, such as might occur in literary history, interpretation, or criticism. And the Procedure category, defined broadly as a step or steps toward a goal, could include, for example, a reference to a literary technique.

The reason for making these profile comparisons is to provide a basis for evaluating the potential of the developed taxonomies, not to evaluate the relative merits of the GRE Psychology and Literature tests. However, it is of interest to consider the kinds of interpretations that a test committee might give to a subject test's profiles. For example, the profiles might be seen as faithfully reflecting the subject area. Or, they might be seen as reflecting only selected portions of a subject area. The selection factor might be the educational level of the examinees, say intermediate, resulting in a test committee's toning down or perhaps even excluding the field's most advanced cognitive demands and/or aspects of knowledge. Or, subject-matter experts may judge that only certain portions of a field can bc measured successfully by means of a multiple choice test. Also, in the judgment of experts, the taxonomic outcome could misrepresent a field, even after taking account of the constraints of educational level and of the test's item format. Sorting out these interpretations might be a useful exercise for a test committee. Such an exercise, repeated periodically, could result in constructive changes in the cognitive features of subject tests.

Much more useful, however, would be profile comparisons among subject tests within a broad domain. In the case of the GRE Program, for example, the Psychology Test Committee might want to compare its taxonomic profiles with those of Economics, Biology (Population Biology), and/or Sociology, and the Literature Test Committee might want to compare its profiles with those of History and/or Music.

Profile Similarities

Also of interest are profile similarities among subject tests, especially as more of them are subjected to taxonomic analyses. Analyses of profile similarities would be an important part of an attempt to group subject tests into families. An intriguing possibility is that the observed demarcations among families reflect conventional divisions among broad domains of knowledge, such as between the Humanities and the Social Sciences. Also, in cognitive terms, some subject areas may be prototypical of their families, others may be at the margins, and still others may represent interdisciplinary "hybrids".

Examination of profile similarities also could reveal that some categories are of importance across most or all subject tests. Based on the initial findings (Table 6), the categories

Analyze, *Identify*, and *Entity* are promising candidates in this regard. It is also possible that one or more of the categories would be found to be of negligible importance within most or all subject tests. Such an outcome might signify that a taxonomy is in need of revision, that the low frequency categories are better applied to less or more advanced levels of knowledge within a domain, and/or that additional item types or testing formats are needed in subject tests. Based on our findings (Table 6), the categories *Synthesize*, *Support-Weaken*, and *Criterion* appear most likely to be subject to these sorts of interpretations as more subject-test profiles are investigated.

Improving the Procedure

Needed Improvements

The coders rarely emphasized difficulties in interpreting the taxonomies. Nevertheless, attention should be given to strengthening them, especially the Cognitive Demand Taxonomy. A very useful first step would be to extend the reliability and profile analyses to the subcategory level.

The developed coding procedure is sufficiently reliable to characterize the cognitive features of moderate to large sets of test items, such as a single form of a subject test. The procedure also could be used to form taxonomically homogeneous groups of items for research purposes--by including only those items that are classified identically by independent coders. Yet it would be important to improve the levels of intercoder agreement. Subtle differences could then be detected, whether in tracking the evolution of test forms over time, or in detecting profile differences between closely related subject tests. Moreover, higher intercoder agreement levels would reduce item attrition in research applications.

How might the coding procedure be improved? Although the present study was not designed to evaluate alternatives, certain findings do provide some clues.

Cognitive Demand

As already noted, the estimated agreement level for these codings was 77%. However, this estimate differed by subject test. It was 83% in the case of Psychology and 71% in the case of Literature.

One of the coders had majored in Psychology and the other had majored in Sociology. Since Sociology is closer to Psychology (Social Sciences) than to Literature (Humanities), perhaps the discrepancy in coder agreement between the two tests was due to a greater degree of coder expertise in the case of Psychology. Also, for those Psychology test items included in the reliability analyses, the coder having more background in Psychology had base rates of 25% for *Analyze* and 74% for *Identify*, whereas the coder having less background in this field had base rates of 36% for *Analyze* and 63% for *Identify*. This difference was not statistically significant, but its direction was consistent with the hypothesis, noted earlier, that greater expertise is associated with a shift in the cognitive demand that the coder attributes to some subject-test items.

This situation suggests a procedural modification that could strengthen the internal validity of the profile analyses. Coders would be selected so that the level of coder expertise is roughly comparable across all of the subject tests to which the taxonomies are applied. This still leaves open the question of what the level of expertise should be, as discussed below.

Returning to the matter of the observed difference in agreement levels, perhaps subject-test differences in base rates (Table 6) also are relevant. For Psychology, approximately 36% of the items were classified as *Analyze* and approximately 63% were classified as *Identify*, whereas, for Literature, approximately 70% of the items were classified as *Analyze* and 27% were classified as *Identify*. In the case of Literature but not Psychology, a separate passage (narrative) often is included as stimulus material. When such complex information is part of the item's manifest content, the coder may experience uncertainty regarding when to move down the list from *Analyze* to *Identify*. More generally, perhaps the incidence and/or amount of coder uncertainty would be greater for those subject tests that require examinees to go relatively far beyond the information given. For Literature, the test's generally "distant" cognitive demands may require a relatively high level of subject-matter expertise in order that coding uncertainties be resolved reliably. For Psychology, however, the test generally makes less "distant" demands, perhaps resulting in less uncertainty and therefore requiring less subject-matter expertise to yield satisfactory coder agreement.

It appears reasonable to suppose that the coder agreement level for the Literature Test provides a lower-bound estimate of the reliability of our coding scheme for the Cognitive Demand Taxonomy. It would be strategic, therefore, to use this test in future efforts to improve the scheme's overall reliability, say to about 80% perfect agreement. It also appears that an "intermediate" level of subject-matter expertise would suffice, perhaps signified by attainment of an undergraduate degree in the Humanities. However, the extent of expertise required within the intermediate range still remains to be determined. From the practical standpoint of designing a cost-efficient coding system, it might be especially informative to compare the average coder-agreement level for a group of trained coders who majored in Literature with the average for a group of trained coders who majored in the Humanities but not in Literature.

Aspects of Knowledge

As already noted, the estimated agreement levels for these codings were 69% for the X subunits (stems) and 75% for the Y subunits (keys). These estimates also differed by subject test.

They were 85% and 86% respectively for Literature, and 53% and 64% respectively for Psychology. In contrast to the situation for the Cognitive Demand Taxonomy, it was Psychology rather than Literature that posed problems for the coders. It will be recalled, however, that when the Aspects of Knowledge Taxonomy was applied to the Psychology test items, more than two of the categories had non-trivial base rates for both the X and Y subunits (Table 6). Consequently, the upper limit of coder agreement is somewhat constrained for Psychology, and probably for other subject tests as well.

For this taxonomy, the coders attained satisfactory levels of agreement on the Literature test even though neither had majored in that subject. This outcome suggested that neither coder experienced special difficulties in applying the taxonomic categories to the subunits--once they had demarcated those subunits. This impression was largely sustained by the author's informal review of the reliability protocols for the Psychology test. However, there were numerous instances in which the Psychology items were parsed differently by the two coders, and such differences often resulted in coding disagreements. Indeed, differences in parsing typically affected not only one of the subunit codings, but the other(s) as well.

For the Psychology test, the highest category base rates occurred for *Entity* and for *Relationship* (Table 6). By definition, a relationship includes at least two entities.

As a result, a coder difference in parsing an item was likely to result in a reversal of the codings of the X and Y subunits by the two coders. Also, there was a (non-significant) coder difference for this test. One of the coders chose Entity more frequently than Relationship when coding both the X and Y subunits, whereas the other coder reversed this pattern when coding the X subunits and exaggerated this pattern when coding the Y subunits.

There appears to be a simple way to reduce much of the ambiguity in the parsing procedure. Instead of parsing the declarative statement primarily from left to right, the coder would first demarcate the Y subunit. In the case of a closed stem, the coder need perform no translation of the Y subunit because the key stated in the test booklet is already a codable unit. For open stems, the coder would perform, when necessary, a minimal linguistic translation that renders the key a meaningful unit. The X subunit in the stem would then be identified and separated from a qualifying phrase (C subunit) in the stem, if any. The single declarative statement would still be written down for the purpose of coding the item's Cognitive Demand. What would be added are the two (or three) additional subunit statements noted above. These additions would increase the writing time, but that increase should be offset by reductions in ambiguities faced by the coders.

Coding the C Subunits

We did not formally compare the C subunit codings of the Psychology and Literature tests, an omission that was deliberate. The X and Y subunit codings appeared to be sufficient in making these comparisons, and it was sensed that the addition of the C subunit might unduly magnify the Literature test's emphasis on Language, an emphasis that was fully apparent from the X and Y subunit codings alone. Nevertheless, the C codings probably should be retained as part of the procedure because they facilitate other aspects of the coder's task and they could provide additional information that is of interest to test committees.

Conclusions

The cognitive taxonomies and coding procedure developed in this study provide new instruments for appraising the cognitive demands and aspects of knowledge contained in the items of contemporary subject tests in the Humanities and the Social Sciences. These instruments were designed to appraise the manifest cognitive contents of subject-test items. The instruments could be used to supplement and perhaps even extend information-processing and psychometric analyses of subject-test items. For example, groups of test items selected to be homogeneous, from a taxonomic standpoint, might be shown to differ in their information-processing requirements and/or psychometric properties.

After some further development, these instruments also could be used to provide GRE test committees in Psychology and Literature with accurate portrayals of the cognitive features of their respective tests. The data base could be extended to include other GRE subject tests in the Humanities and in the Social Sciences. Then, within each of these broad domains, the cognitive profiles of the separate subject tests could be compared with one another.

Another important step would be to strengthen the taxonomies by extending the reliability and profile analyses to the subcategory level. Such analyses would help determine, among other things, how the major categories might be revised and/or reorganized to good advantage, especially in the case of the Cognitive Demand Taxonomy.

Although the coder agreement levels were acceptable for some purposes, it would be desirable to raise them. Our findings suggested how this might be done. Regarding cognitive demand, we would determine the level of subject-matter expertise that is sufficient to obtain accurate codings of test items requiring the interpretation of text, as in the GRE Literature Test. Regarding aspects of knowledge, we would evaluate whether the codings of items from the GRE Psychology Test are improved after implementing suggested changes in the parsing rules.

References

Bloom, B. S. (Ed.), Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. k. A taxonomy of educational objectives: Handbook I: Cognitive domain. New York: Longmans, Green, 1956.

Bowman, C. M., & Peng, S. S. A preliminary investigation of recent advanced psychology tests in the GRE program--An application of a cognitive classification system. Unpublished GRE report, 1972.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. A study of thinking. New York: John Wiley and Sons, 1956.

Fleishman, E. A., & Quaintance, M. K. Taxonomies of Human Performance. New York: Academic Press, 1984.

Glaser, R. Education and thinking: The role of knowledge. American Psychologist, 1984, 39, No. 2, 93-104.

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

Messick, S. Abilities and knowledge in educational achievement testing: The assessment of dynamic cognitive structures. In S. Elliott & J. Mitchell, Jr. (Eds.), Buros-Nebraska series on measurement and testing, Vol. 1, B. Plake (Ed.), Social and technical issues in testing: Implications for test construction and usage. Hillsdale, NJ: Erlbaum, 1984.

Metfessel, N. S., Michael, W. B., & Kirsner, D. A. Instrumentation of Bloom's and Krathwohl's taxonomies for the writing of educational objectives. Psychology in the Schools, 1969, 6, 227-231.

Mosenthal, P. Defining the expository discourse continuum: Towards a taxonomy of expository text types. Poetics, 1985, 14, 387-414.

Sigel, I. E., & McGillicuddy-Delisi, A. V. Parents as teachers of their children: A distancing behavior model. In A. D. Pelligrini & T. D. Yawkey (Eds.), The development of oral and written language in social contexts. Nor'ood, NJ: Ablex, 1984, 71-92.

Sokal, R. R. Classification: purposes, principles, progress. In Johnson-Laird, P. N. & Wason, P. C., (Eds.), Thinking: Readings in cognitive science. New York: Cambridge University Press, 1977, 185-198.

Teitelbaum, P. M. Memorandum to the GRE Sociology Committee, ETS, November, 1981.

Tucker, C. Delineation of reasoning processes important to the construct validity of the analytical test. ETS report, 1985.