DOCUMENT RESUME

ED 382 661 TM 023 096

AUTHOR DeMauro, Gerald E.; And Others
TITLE Delimiting the Verbal Domain.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-94-34
PUB DATE May 94
NOTE 115p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS *Definitions; Job Analysis; *Semantics; *Test Construction; *Verbal Ability
IDENTIFIERS *Domain Knowledge; Educational Testing Service; Rule Application; *Test Developers

ABSTRACT

A functional analysis was made of the verbal domain as it is defined by the test development process at the Educational Testing Service (ETS). The thesis of the study was that the verbal domain consists of semantic relations that examinees must interpret. Test developers used a system of rules to sample stimulus materials from among infinite possibilities that elicit an interpretive response. Many of these rules were codified and others were in different stages of formation, beginning in the use by individual test developers, and eventually becoming part of generally accepted practice. A job analysis approach was used to examine how rules are employed to present the desired semantic relations to the examinees for their interpretation. Test developers (n=8) from two major divisions of ETS were observed and interviewed. They also rated the importance of each of the rules, suggested rules that they use that were not listed, and delineated the professional activities for which these rules are most important. The study indicates that the rule system for sampling semantic relations is an evolving body, both for the individual test developer and for test development as an institution. Ten tables support the discussion, and two appendixes provide supplemental information. (Contains 26 references.) (SLD)

RR-94-34

RESEARCH REPORT

DELIMITING THE VERBAL DOMAIN

Gerald E. DeMauro
Ave Merritt
Richard Adams

Educational Testing Service
Princeton, New Jersey
May 1994

DELIMITING THE VERBAI, DOMAIN

Gerald E. DeMauro
Ave Merritt
Richard Adams

Abstract

A functional analysis was made of the verbal domain as it is defined by the test development process at Educational Testing Service. The thesis of the study is that the verbal domain consists of semantic relations that examinees must interpret. Test developers use a system of rules to sample stimulus materials from among infinite possibilities that elicit an interpretive response. Many of these rules are codified in a body of memoranda and corporate publications, and others are in different stages of formation, beginning in the use by individual test development professionals, and eventually becoming part of generally accepted practice. Because the process of test development is central to the formation of the verbal domain, a job analysis approach was used to examine how rules are employed to present the desired semantic relations to the examinees for their interpretation. This approach explains the process of delimiting the verbal domain through a comprehensive review of the uses of the rule system that defines when the items are functioning properly and when they are not. Test developers from two major divisions of ETS were observed and interviewed. They also rated the importance of each of the rules, suggested rules that they use in their work that were not listed, and delineated the professional activities for which these rules are most important. The study indicates that the rule system for sampling semantic relations is an evolving body, both for the individual test developer and for test development as an institution.

Executive Summary

This study attempted to document how the verbal domain is shaped. It concludes that the verbal domain is actually delimited from the universe of possible semantic relations according to a system of rules for sampling that universe. This rule system evolves from work activities that individual test developers find to be successful and then share with others. Memoranda recommending these activities are distributed by managers or group heads. These memoranda then become the bases of corporate publications that codify the commonly accepted rules used across programs to select material to measure verbal ability.

The task of documenting the domain-delimiting activities of test development was approached through job analysis because the methodology

1. provides great control over the amount that the study intrudes on the test development process; and

2. is most appropriate for studying the professional activities of test developers that delimit the domain of verbal skills tests

The job analysis approach used in this study was based on: Identification of an expert committee and review of documents that guide development of verbal instruments, identification and articulation of the professional activities that shape the verbal tests by the expert committee, development of a corpus of rules for defining the verbal domain, small survey validation of the importance of these rules, small survey linkage of the body of rules to the professional activities of defining the verbal domain, observations and semi-structured interviews of how the rules govern the ways in which the verbal domain is delimited, and integration of all of these sources of information.

Two major findings of the study are discussed:

1. A rule system for sampling the verbal domain is formed by the private use of sometimes implicit or informal rules that are later formalized, first by individual test developers, and then, as they gain common currency, by the corporation;

2. The evolution of rules follows a course that begins with general exclusion of material that the test developer judges to be inexpedient, contrary to sound measurement, or offensive or differentially accessible for some defined segments of the examinee population and ultimately ends with guidelines for including material that is more appropriate.

Table of Contents

Table of Contents

List of Tables

List of Tables

Introduction

Purposes of the Study

The original intent of the study was to identify and evaluate the rules for including material in and excluding material from verbal skills tests (Emmerich, 1991). The rationale for the study was that these rules determine such factors as the subject matter of reading passages and the vocabulary used in test items, which in turn affect the ability of the examinee to respond to the items. This is based on a body of cognitive research that shows that the level of cognitive skills employed to solve a problem depends on the familiarity of the context in which the problem is posed (Scribner & Cole, 1973).

The rules and guidelines used by test developers were examined to determine whether there were codified practices related to which material would be suitable for inclusion in a test and which material would not be. It became evident early in the study that rules used by test development professionals at Educational Testing Service (ETS) focus more on the skills that are required of examinees to respond to the test items than with the body of material or words that compose the items. For example, the rules for sampling reading comprehension passages are not designed to limit the range of passages to certain authors or certain time periods, but rather to assure that the examinees are not advantaged or disadvantaged in their responses by subject matter in ways that are unrelated to their abilities to interpret the passages. Ultimately the rules are concerned with best test development practices to measure the abilities examinees use in their responses to a test, and only with subject matter or vocabulary to the extent that these factors affect those responses.

The focus of this study is on the evolution of practices that are used to ensure that verbal skills tests sample the desired cognitively-mediated responses from the examinees. In that sense, the question addressed by this study is not how the test development rule system defines the vocabulary or the subject matter of verbal skills tests, but rather how the rule system functions to elicit desired examinee responses.

Purposes of Verbal Skills Tests

The three largest verbal aptitude testing programs, the Scholastic Aptitude Test (SAT), the Graduate Record Examinations (GRE), and the Graduate Management Admission Test (GMAT) all see their purpose to be to provide test scores to be used as uniform admissions criteria (see College Board, 1987; and Educational Testing Service, 1988). The importance of verbal skills for academic achievement in so many areas makes it necessary that they be considered in admissions decisions.

The original mission of the College Board (1899) to standardize the admissions requirements of colleges (Angoff & Dyer, 1971) gradually began to reflect a growing demand that students demonstrate an understanding of the relationships among facts. This goal implies assessment of such higher cognitive skills as analysis and evaluation rather than the mere recall or recognition of facts.

From the first Scholastic Aptitude Test in 1926, the verbal domain has been defined in terms of the cognitively-mediated responses of the examinee to the stimuli of the test questions. The particular combination of words in a question is a sample of infinite possibilities of representing a particular type of semantic relationship that must be interpreted by the examinee. The verbal domain is defined by the response class of the examinee, e.g. the examinee must recognize that two elements have an analogous relationship to that of two other elements. Analogy items are defined by the interpretation of that relationship, not by the particular elements that were sampled to represent it.

Test development practices for sampling stimuli to be included in or excluded from a test of verbal skills are designed not for the tightness of the structure of the test items, but rather for the tightness of the cognitive processes they elicit. The items that comprise the test are one possible vector of ways to evaluate examinee responses.

The Response Class

Interpreting semantic relations. Because the verbal domain is responses to verbal test items, than verbal ability or verbal skills is determined by knowledge of the meaning of words, the syntactic rules for their combinations, and the semantic relations between the concepts symbolized by the words (Hunt, Lunneborg, & Lewis, 1975). The materials that elicit these responses (test items, reading passages, etc.) must be chosen for their stimulus properties. In fact, the rules governing test development activities specify exactly what test developers must do to increase the reliability of eliciting a particular desired interpretive response, e.g. how to design antonyms to convey unambiguously the desired oppositional relationships, and how to avoid eliciting other responses that might contaminate the verbal construct, e.g., differential emotional responses to certain reading passage topics.

A functional analysis of the verbal domain is similar in many respects to a functional view of language. Structural properties of a communication can have infinite variety and still carry the same meaning. The formation of the communication, which amounts to delimiting a sample from infinite possibilities, is accomplished through the application of rules, or grammar. Similarly, the test as a communication is produced by sampling certain semantic relations from infinite possibilities, and the sampling is controlled by a rule or grammar system that is examined in this study.

The response class of the verbal domain might best be conceived as the capacity of the examinee to select and use problem-solving strategies from an available repertoire. Careful study of the type of cognitive tasks demanded by the test items may enable us to describe the cognitive processes involved in the test response class.

Moreover, it is reasonable that if such discrete abilities as the rapid application of cognitive skills like memory retrieval skills are important for verbal skills performance (see Sternberg, 1979 for a review of this literature), than metacognitive processes related to evaluation of those skills, selection of alternative strategies when particular skills are insufficient, and organization of different skills to meet complex cognitive demands would also advantage examinees on verbal tasks.

Developmental perspective. Several models of cognitive functioning are available to describe such higher order skills. Commons, Richards, & Kuhn (1982) hypothesize that a person capable of formal operations can formulate and test hypotheses about two classes of objects or elements (such as would be required by verbal items) without consciously reflecting on the system as a whole. A higher order of cognitive operations, termed systematic operations, would enable the person to operate on relations of whole classes of objects or elements forming systems, the type of operation that would be most valuable in reflecting on the classes of multiple abstract relationships presented by a stem, key, and options of an item. This model is one of many ways of describing how better-developed cognitive skills would advantage examinees on the verbal skills measures.

Each of the three measures of verbal aptitude has an analogy section including items involving concrete, abstract, or mixed analogies. For concrete analogies, the two terms in the item's stem and the two terms in the key are both nouns and refer to elements that can be perceived by the primary senses. For mixed analogies, some, but not all of the four elements of the stem and key refer to things that can be perceived by the primary senses. Abstract analogies are those in which none of the four elements in the stem and key can be perceived by the primary senses.

To solve concrete analogies, examinees must not simply evaluate the perceptible characteristics of the elements, but must evaluate the similarity of the relationships between the characteristics of the elements in the stem to those between the elements in each option and then judge which of the options presents a relationship most like that presented in the stem. Concrete analogies are best solved through processes of hypothesis testing and abstract reasoning, and examinees who are capable of such formal operations are clearly advantaged.

To solve abstract analogies, the examinee must discern the nature of the relationship between the elements in the stem and then between the elements in each of the options. Examinees who are most skilled can then take the further step of organizing these abstract relationships in order to compare them to determine which option represents a relationship that is most similar to that represented in the item's stem. This involves simultaneous evaluation of different abstract relationships. Clearly, an examinee who is capable of what Commons and his colleagues call systematic reasoning is at an advantage on a task of this type. Less-skilled examinees who answer the question more on the single comparison of the relationships in the stem and those in each option, one at a time, would be disadvantaged on those more difficult items in which one option is slightly better than others.

This analysis extends to all of the types of items that compose the three verbal skills measures. For example, in reading comprehension items, the examinee who is capable of comparing and organizing abstract relationships would be at a clear advantage when asked to perform such tasks as considering which of several plausible options best represents the main idea from a passage, or by simultaneously evaluating other ideas given in test item options to determine which expresses an idea that is most central to the meaning of the passage as a whole. Examinees who have better-developed cognitive skills would also be advantaged in making analogical extensions of ideas presented in passages.

Information processing perspective. A similar cognitive perspective may be drawn from information processing research. For example, from a series of six experiments, Hunt et. al. (1975) conclude that people with greater verbal skills may be distinguished from those with lower skills by their greater:

1. speed in recognizing the association between the visual stimulus of a word or letter and its conceptual meaning;

2. ability to retain information in short term memory about the order of stimulus presentation, which gives them more time to extract meaning from stimuli instead of recovering information about the order of stimulus presentation from contextual cues;

3. speed in manipulating data in short term memory, which is useful for making the transformations required in speech.

The tasks of semantic interpretation required by the analogy, antonym, sentence completion, and reading comprehension item types, give a clear advantage to examinees who can efficiently store, manipulate, and retrieve data from short term memory.

Interaction of Cognitive Skills and Context

Structural features of test items, such as the vocabulary they employ, are important intervening variables in the sense that examinees will not understand the cognitive demands if they are posed in words that are generally unfamiliar, or if these features enable higher-order cognitive responses from some groups of examinees simply because they are more familiar. For this reason, to standardize the response class, it is important to attend to the contextual features of test items.

The ability of examinees to apply their skills to the task presented in the test item interacts with their experience. Obviously, some problems or types of problems are so familiar to the examinee that cognitive processing becomes less controlled and more automatic (see Shiffrin & Schneider, 1977), conferring a particular advantage on the examinees who have used effective cognitive strategies with similar problems before. Examinees who perform at higher cognitive levels are also advantaged on novel problems that require greater control of cognitive processing.

Cross-cultural studies (e.g. Scribner & Cole, 1975) show that the likelihood that people use their repertoire of higher cognitive skills increases when the context of the task is familiar. This suggests that familiarity of context mediates the expression of verbal skills, whereby examinees are less likely to employ higher levels of cognitive functioning in contexts that are unfamiliar to them.

Part of these findings has been replicated in studies of differential item functioning, in which the familiarity of the context of items is correlated with the differentially superior performance of examinees (Scheuneman & Gerritz, in preparation). Similarly, theories such as the schema theory of reading comprehension (Anderson et al., 1977) propose that comprehension is facilitated when reading material can be assimilated into existing cognitive structures.

The cross-cultural research cited above, involved measuring the levels of cognitive skills used in responding to problems that demanded analogical thinking and formal operational problem solving. These levels were found to depend on the familiarity of the context of the problem. Hypothetical dilemmas that were unfamiliar elicited lower levels of problem solving skills than more familiar problems. Obviously, people who are incapable of formal operations do not become capable because the context of the problem changes. Rather, the context of the problem may give people information about the type of solution that is sought (Luria, 1971; Glick, 1975), enabling a more accurate measure of the level of the cognitive skills available to examinees to evaluate what an item requires and to select a strategy for answering it from their repertoires. In this sense, many more people may be capable of higher order cognitive operations under favorable conditions.

For this reason, the demands of the test items must be absolutely unambiguous. The capacity to respond correctly must depend only on the examinee's ability to interpret and evaluate the semantic relations. Obviously, there exists infinite ways of testing this ability, and a test development rule system has evolved to assure that the desired interpretive skills of the examinee and no other skills are sampled by the test items.

Test Development: Delimiting the Domain

Overview. Test developers at ETS are generally subject matter experts who both write test items and direct experts from other institutions to write test items. Generally, one such ETS expert is designated as a test assembler, that is, one who oversees the composition of a test. A second expert is designated as a test specialist, or an expert reviewer of the draft test. A third expert serves in the role of coordinator, or supervisor of the production of several tests.

For some tests, particularly those that measure achievement in a content area, one or more committees of experts outside of ETS, sometimes called committees of examiners or technical advisory committees, are convened to set the test specifications. For aptitude tests, such as those that measure verbal skills, the test specifications are predetermined from a long history of research with various content, context, and item types. Committees of experts outside of ETS generally review these specifications and make whatever changes they deem necessary to maximize measurement quality.

The givens of test design. Ongoing research, validity studies, and item analyses have contributed to the corporate experience concerning the functional properties of different types of verbal stimuli. This experience is codified principally in the form of test specifications and the procedural rules for test assembly. Generally, the test specifications for verbal skills measures are intact before the test assembler begins the process of selecting or writing the materials and items that will comprise the test. These specifications include the following:

(1) content or subject matter of the test

(2) context or setting of the test items

(3) level of use or cognitive level required for response, e.g., analysis

(4) modality or response type, e.g., short constructed response

(5) type of or means of eliciting demonstration of the cognitive skill, e.g., sentence completions, especially for multiple choice tests)

(6) difficulty, or proportion of examinees of a certain skill level that will respond correctly, or discrimination of test items

Content. The verbal sections of the Graduate Record Examinations (GRE) General Test and the Scholastic Aptitude Test (SAT) both use reading comprehension and analogy item types, described in greater detail later in the text. A third item type, antonyms, has been used in the GRE and the SAT, but is not included in the new revision of the SAT. These item types are not themselves the content domain but, rather, are ways to determine how well the examinee can interpret the semantic relations in English (memo, Woisetschlager, 1991). The content domain is the examinee's interpretive responses.

For reading comprehension, the domain is the message conveyed by written passages, including stated, logical, and inferred meanings (personal communication, L. Hecht, 1988). This meaning depends on the operation of semantic interpretation.

A fourth item type used on the GRE and SAT is sentence completion. Because a sentence may be completed on the basis of any reasonable relationship between the given part of the sentence and the response, the domain of the sentence completion items is less prescribed than those of antonym and analogy items, which are limited to semantic relations of opposition and similarity, respectively (memo, Woisetschlager, 1991).

These item types are designed to sample specific types of semantic relations, by no means all possible semantic relations, that are deemed important to understand for success as a business, graduate, or undergraduate student. One of the first undertakings of this project was review of documents that describe the semantic relations that are intended to be sampled by each item type.

Context. Context is typically determined by the time and place in which an assessed knowledge, skill, or ability will be used by successful examinees. For professional examinations, context usually is a simulation of a professional activity. On verbal aptitude tests, context most often applies to the environments that require students to make semantic interpretations. These environments for GRE antonym, analogy, and sentence completion items are: arts and humanities, social studies and practical or everyday life, science and nature, and human relationships and feelings (Carlton, 1983). For reading comprehension passages, these environments are humanities, social science, biological science, and physical science. Additionally, the SAT makes use of narrative passages or environments.

Donlon and Angoff (1971) specify other contextual consideration for the SAT verbal domain as shown in Table 1. These researchers use the word "content" to refer to the subject matter of the items (within-question content). This use should not be confused with our functional use of "content," which applies to operation of examinees on semantic relations.

Table 1

Formal Rules:
Contextual Dimensions for
Verbal Semantic relations
(from Donlon & Angoff, 1971)

Item Type	Dimension	Explanation[1]
Reading Comprehension	Passage topic	Narrative (SAT), biological science, physical science, synthesis, argumentative (GRE, SAT),humanities, and social studies, dependent, independent of stem (GRE, SAT);
Analogies	Concreteness	Abstract, concrete, mixed;
	Dependence	Dependent, independent of stem;
Antonyms	Generality	Fine or general distinction required;
	Number of words in options	Single word or phrase;
	Part of speech	Noun, verb, or adjective;
Sentence Completion	Content Number of blanks	

[1] Refers to GMAT, GRE, and SAT unless otherwise specified.

This level of specification of context is not sufficient for selecting test material. For example, what constitutes Everyday Life for the test developer may be foreign to the examinee. Rules are formed to protect examinees from contexts that are differentially familiar or just too difficult, or that measure factors that are not related to functional capacity.

In verbal skills or aptitude measures, these rules ensure that the test context is accessible to the examinees. If examinees had limited experience with an area, they would not understand the subtle nuances of the semantic relations. For example, personal foul:football:: slashing: would have no meaning to someone who was unfamiliar with football or hockey even if the concept of foul were familiar.

Cognitive function. The level of cognitive skill needed to operate on semantic relations may depend on the familiarity of the context in which the relationships are presented. For example, recognizing the main idea of a reading passage requires some degree of abstract reasoning, but the content of the passage may be familiar to the examinee, and the ability to recognize the main idea of familiar material is less difficult than the ability to recognize the main idea of unfamiliar material. Perhaps one reason for this is the fact that familiar material is overlearned, and the retrieval codes of overlearned material may have greater sensitivity to arousal by incoming stimuli (Hunt et. al., 1975).

Cognitive research has shown that some people who perform poorly on standard cognitive Piagetian tasks display great capacities for abstract thought and inference on tasks from their everyday lives (see Gleitman, 1991 for a review of this literature). Early studies of the greater abilities of people within a culture to comprehend reading passages in the context of that culture were replicated by giving college students readings from within and from outside of their fields of major (Anderson, Reynolds, Schallert, & Goetz, 1977). Similarly, the context of the semantic relations greatly affects the examinee's capacity to draw inferences, define the author's purpose, identify the main idea, or understand the task demanded by the test item.

Level of cognitive functioning is sometimes explicitly specified as proportions of the test that sample different cognitive levels such as those levels described by Bloom's Taxonomy. The level of cognitive functioning may also be indirectly prescribed by the context of the test items. It is certainly clear that familiarity with different human relationships varies a great deal from person to person, and test developers must be vigilant that the choice of material does not allow for different levels of cognitive functioning or for different semantic interpretations based on its varying familiarity to different groups of examinees.

Modality of response. There is some evidence that modality of response (essay or multiple choice), or that among essays whether a topic is required or chosen from among several options, may affect the response of the examinee or the construct being measured (DeMauro, 1992). Although there is a growing body of research in this area, there are few consistent, explicit rules governing the development of essay items to sample the verbal domain, and this is beyond the scope of the current study.

Item types. The verbal item types of all aptitude tests developed by Educational Testing Service can trace their origin to the early development of the Scholastic Aptitude Test. Experience in this area has led to modifications of the item types to tailor them to the demands of specific testing programs. Item types are decided by committee and have had great stability over the years.

In 1926, a commission led by Carl C. Brigham designed the first Scholastic Aptitude Test, consisting of arithmetic problems, classification, artificial language, antonyms, number series, analogies, logical inference, and paragraph reading. A recent principal axis factor analysis of this test revealed two scales: a verbal scale and a quantitative scale (DeMauro, in progress).

The SAT is in transition. The verbal section of the traditional version includes antonyms, analogies, sentence completion, and reading comprehension as item types, while the new revision includes Critical Reading (four passages), analogies, and sentence-based items, but not anatonyms. The current version of the GRE General (Aptitude) Test includes 186 items, of which 76 are verbal skills items, including: analogies (18 items), antonyms (22 items), sentence completions (14 items), and reading comprehension (22 items on 4 reading passages) (see Carlton, 1983, for another explanation of these item types). Through several transformations (1961, 1966, 1976, 1982) the 85-item verbal section of the GMAT has arrived at its present composition of a reading comprehension section (25 items), a sentence correction section (25 items), and a critical reasoning section (35 items).

Psychometric properties of items. Psychometric properties, such as question difficulty, are described in planning memoranda circulated to test developers before the assembly of the pretest. Based on their experience, test developers estimate the difficulty and discrimination of test items during test assembly. Items are then pretested for difficulty, discrimination, and differential difficulty for different examinee groups. The pretest process, described in greater detail below, as well as review of the item psychometric properties and advise to committees of examiners are guided by professional statisticians and statistical coordinators who are not test developers but who work in coordination with the test assembler. Items that are flawed in any of these stimulus characteristics are eliminated from the pool of items that contribute to test scores.

Test assembly. The specifications are used in test development by the test assembler, who assembles material in the test that is consistent with these specifications. In the choice of appropriate reading passages, test items, analogical relationships, sentences for sentence completions, and other stimulus materials, this test developer is continually deciding how to best elicit a response from the examinee. This necessarily means that the test developer must judge the quality of the semantic relations sampled by the test, which requires interpreting the very semantic relations that the examinee will be required to interpret.

For tests of verbal skills, the test assembler begins by gathering material from which test items may be generated to meet the requirements of the specifications to form a pretest. A pretest is a compilation of test items that can potentially form a final form or operational test. The items are generally administered as a separate section of an operational test, but it is not scored and does not affect the performance of the examinees. The purpose of pretesting is to provide assessment of the psychometric properties of possible test items before they become part of a final form.

Items in the pretest may be written by the test assembler or by other experts on staff, or may be solicited from committee members or other outside experts. Evaluating test items and assembling and reviewing draft tests require that the test developers put themselves in the shoes of the examinees and clearly interpret the semantic relations required by the test items. The less ambiguous the relationship, the better the test item. Test development rules and guidelines are desired to assure that the item samples precisely the desired relationships, and that a knowledgeable or skilled examinee will make the proper interpretation of the stimulus materials.

The draft pretest is checked and keyed by a test specialist, who is a test developer with specialties in verbal tests, and then sent to a test coordinator for review. When issues that arise from these expert reviews have been resolved, the draft pretest is sent to editors for a series of reviews to guarantee that the language is correct and clear. The pretest items are administered with an operational form of the test, to provide information about the psychometric quality of the items. With this information, informed decisions may be made about whether a test question should be included as it is, included in a modified form, or not included in an operational (sometimes called "final") form of a verbal skills test.

The functional view of test development means that the test assembler must give careful attention to the interaction of all test specifications in choice or development of test material, because the stimulus properties of the material are not determined by content or context or question type alone. For example, specifying which cognitive functions the test will assess affects the contexts chosen for test items. Cognitive psychology suggests that efficient problem-solving strategies are most commonly used with familiar problems, and that people often are capable of demonstrating higher levels of cognitive functioning on problems posed in familiar contexts than they demonstrate on problems posed in unfamiliar contexts.

Suppose the test is designed to measure analogical reasoning, and a question uses the stimulus "egg:gourami:: seed:." The intended key is "plant." One would need to know, as aquarium owners might, that the gourami is an egg-laying fish. Another item that frames the stimulus as "egg:chicken:: seed" might be more discriminable simply because the more familiar context enhances its stimulus properties for analogical skills. If this is true, the first item might be eliminated after pretest, while the second item might be used in a final version of a test. This functional analysis illustrates how decisions about what level of cognitive skills the test should elicit guide decisions about the contexts of the test items.

The experience of test developers with how these interacting considerations affect examinee performance contributes to the codified rule system. In this example, even without an explicit rule governing specialized knowledge of aquarium fish, implicit rules about cognitive functioning and question discrimination help to define the item context and clarify the analogical relationship the test question is intended to measure. Eventually, specific rules might develop to govern the types of material that is so specialized that it interferes with the examinee's ability to interpret the analogical relationship.
The operational definition of the verbal domain depends on the ability of test developers first to interpret the semantic relations required by the test specifications and then to develop material that best elicits the same interpretative responses from the examinees.

METHODS

Early Committee Meetings

Levels of definition. As described earlier, this study initially proposed the compilation of rules for either including verbal material in or excluding verbal material from Educational Testing Service's verbal skills tests. The first attempt at this included design and circulation of a broad questionnaire asking test development specialists to consider and record any such rules. Although this questionnaire had been reviewed by other specialists, the recipients were unable to complete it, because there was confusion about what could be considered to be a rule.

Subsequently, two meetings were held with an ad hoc committee of test developers to discuss the meaning of the verbal domain, in the hope that the added precision of terms would help better define the rule system. It became clear from these discussions that although there was much attention to the structural aspects of the domain (vocabulary, etc.), the verbal domain itself was actually defined by the semantic relations that comprise analogies, sentence completions, and antonyms, and perhaps even the relationships of meaning in the reading comprehension passages.

What was meant by domain, as described earlier, was not the vocabulary used to express these relationships, but the relationships themselves, and as evidence of this, the expert test developers repeatedly mentioned that the precise meaning of words was checked against the dictionary to ensure that the functional relationships among words in items were tight and not open to misinterpretation. Even the rules about the structural features of test items, such as the vocabulary to be used or the item format, do not operate to identify ways of testing knowledge about words or about formats. Instead, these rules operate on a functional level to enhance measurement of the ability of candidates to interpret the semantic relations of the test items.

Operationally Defining the Verbal Domain

The role of test developers. Given these early insights, it became clear that the verbal domain is operationally defined by test developers through selection of material for its properties to elicit semantic interpretations from examinees. The task of the test developer, then, is also to interpret the relationships expressed in the material; that is, to perform the same operations on the material as the examinees. The operational definition of the verbal domain, then, must specify these operations.

This perspective led to the careful consideration of test development as essentially a delimiting or sampling activity. The function of the rule system in this activity was examined in terms of the importance of rules used to include or exclude materials from verbal skills tests, the use of rules in the professional activities of test development, and the sequence in which rules are used. A flow chart was developed of test development activities. Each of these activities was associated with a person or group of people that bore the chief responsibility (see Table 2).

This list was sent to a number of item type experts and their supervisors to review, and as a result, was expanded considerably from its initial draft. The second draft was then sent to a director of test development, two group heads, and five other test developers, two of whom were test development supervisors.

The draft flow chart was then given to a committee of four test development supervisors with responsibility for verbal item types. The subsequent chart served as a de facto listing of the professional activities of the job of defining the verbal domain of tests.

Having defined the professional test development activities, it remained to delineate the rule system that was used to perform these activities. In job analysis methodology, the professional activities roughly correspond to the flow chart activities, and the knowledge, skills, and abilities used in performing these activities correspond to the rule system yet to be defined.

<u>Definition of the rule system.</u> This rule system was compiled from expert opinions and the memoranda and test development manuals that test developers use in performing their jobs. These materials were gathered from appropriate training manuals and test development supervisors.

The test development manuals described the nature of the semantic relations the tests proposed to sample. They also provided actual experience-derived prescriptions for what to include in and what to exclude from verbal skills tests. These prescriptions were listed separately and categorized according to item type. In addition to prescriptions that were specifically concerned with the analogy, antonym, sentence completion, and reading comprehension item types, many suggestions, prescriptions, or guidelines appeared in manuals, memoranda, and other documents that pertained to the desired cognitive response of the examinee or to desired psychometric characteristics of test items. Therefore, the list of test development rules was presented in six categories[2], representing the four item types, cognitive functioning concerns, and psychometric properties concerns. The rules were also classified as inclusion or exclusion rules, although this is often an arbitrary distinction because including some material often requires that other material be excluded.

An original list of 121 inclusion and exclusion rules was developed and sent for review to two test development supervisors, one from the

[2] Hereafter, these six categories ℓll be called "areas."

School and Higher Education Programs (responsible for GRE and GMAT) and College Board (responsible for SAT) divisions at ETS. Based on their recommendations, the list was expanded to 186 rules (Table 3), as follows: Antonyms, 28 rules; analogies, 42 rules; sentence completions, 46 rules; reading comprehension, 55 rules; ccgnitive function, 6 rules; and psychometric properties, 9 rules.

Most of the rules called for judgments to be made by the test assembler about the quality of the semantic relations expressed by the test items (See Table 3, below.). For example, Rule 96 states "Avoid using as blanks (in sentence completion items) words that are superfluous to the meaning of the sentence. Other rules involve less judgment and more clerical work on the part of the test assembler, related to institutional quality control procedures concerned with the expressing the semantic relationship as clearly as possible. For example, Rule 94 states "Avoid using as blanks the first words in sentences."

In actuality, both of these types of rules reflect a concern that the item measure the examinee's capacity to make the proper interpretation of the stimulus material. The more clerical tasks, e.g., word counts, etc., are most often related to experience with reading speeds, short term memory capacity, or other intervening variables, and have been codified because test assemblers are not in the position to make judgments about such issues and are therefore provided the guidelines validated by research or experience. Only the rules concerned with fair use or copyright are unrelated to examinee responses, e.g., Rules 140-142.

Among the judgment rules, two specifically ask the test assembler to make some judgment about the skills of the examinees. Rule 80 states, "Assure that the sentence addresses a topic that is appropriate for the examinee population," and Rule 121 states, "Assure the passage is an appropriate reading task for the educational level of the examinees."

Table 2

Content Decision Points

	Responsibility
I. Develop Test Specifications	
A. Analyze curriculum/job (optional)	A, B
B. Review existing test specs.	A, B, C
C. Develop/refine test specs.	A, B, C
D. Develop/refine detail classification	A, B, C
E. Develop section specs.	A, B, C
E. Develop test item context	A, B, C
II. Plan	
A. Prepare item writing plan	A, D
B. Write TD planning memorandum	D
C. Choose item writers and reviewers	A, E, F
III. Prepare Items	
A. Select tried items for reuse	A
B. Select item stimuli and sources	A, G
C. Develop new test items	A, G
1. Make item writing assignments	A, G
2. Write new items	A, G
(a) classify item	H
(b) review test items (peers)	H
IV. Assemble Pretest	
A. Prepare draft test	A
1. Perform global review	I
2. Perform external review (optional)	C, J
B. Edit review	
1. Edit draft test (TPS)	K
2. Perform sensitivity review	L
V. Prepare/Revise Pretest Planograph (optional)	
A. Make any recommended corrections	A
B. Review test (internal)	I
C. Review test (external, optional)	C, J
D. Review/revise and key planograph	A

Table 2 (Cont'd.)

	Responsibility
VI. Evaluate Pretest Results	
A. Evaluate performance statistics	A, I
B. Evaluate DIF statistics	A, I
C. Prepare pool for final assembly	A, I
VII. Assemble Final Test Form	
A. Prepare draft test	A
1. Perform global review	I
2. Perform external review (optional)	C, J
B. Edit review	
1. Edit draft test (TPS)	K
2. Perform sensitivity review	L
VIII. Prepare/Revise Test Planograph	
A. Resolve old PINs	A
B. Review test (internal)	I
C. Review test (external, optional)	C, J
D. Review/revise plano and key	A

Responsibility Key:
A. Item-type Specialist, Test Assembler
B. Client
C. Test Committee
D. Test Coordinator
E. Group Head (SHEP Test Development)
F. Program Director
G. Item Writer
H. Other Item Writer
I. Test Specialist
J. Outside Expert
K. Test Production Editors
L. Test Sensitivity Reviewer

Table 3

A. Antonyms

1. General

a. Inclusion rules

1. Use contrariety in antonym questions only when a strong defensible key is present.

2. Use extreme positions for polar contraries.

b. Contrariety exclusion rules

3. Avoid contrariety as a stem-key pair.

4. Avoid contrariety as distracters unless a much stronger stem-key pair is present.

c. Polar contrary exclusion rules

5. Do not put distracters for polar contraries on the same continuum as the stem-key opposition.

6. Avoid using extreme:midpoint stem-key pairs, e.g., right:middle, as opposed to right:left.

d. Converse relationship exclusion rules

7. Avoid stems and keys that are not in opposition of direction, e.g., husband:wife.

8. Avoid weak converse relationships as stem-distracter pairs unless a much stronger opposition relationship exists for the stem and key.

Table 3 (Cont'd.)

2. General rules for writing antonyms

a. Inclusion rules

9. Consider the connotative and denotative meanings of words.

10. Use familiar words.

11. Substitute the key in sentences using the stem, to determine if the sentence expresses the opposite meaning.

12. If words and phrases are in combination in an item, two must be of one type and three must be of the other type (words or phrases).

13. Phrases as options may have 2-3 words, either adjectival, adjective modifying noun, adverb modifying adjective, verbal, or prepositional.

14. Use parallel parts of speech as the stem and options in single-word antonyms.

15. When the stem may be different parts of speech, make the first option unambiguously one part of speech to establish the part of speech intended for the stem and all the options.

16. Antonyms may test the ability to define words or make fine distinctions among similar distracters.

b. Exclusion rules

17. Avoid specific determiners, or a key which is different in some dimensions from other options.

18. Avoid synonyms as distracters.

19. Avoid antonyms that require specialized knowledge.

Table 3 (Cont'd.)

20. Check GRE antonym stems and keys for overlap in GRE antonym file.

3. General rules for reviewing antonyms

   a. Inclusion rules

      21. Consider the stem in several contexts.

      22. Key the item considering each option.

      23. Check the exactness of the key against the dictionary.

      24. Examine usage and dictionary cross references.

      25. Suggest improvements for the key and distracters.

   b. Exclusion rules

      26. Examine possible obscure uses of distracters that may make them keyable.

      27. Eliminate tricky, frivolous, or implausible distracters.

      28. Eliminate distracters that do not conform to the general rules for writing antonyms.

Table 3 (Cont'd.)

B. Analogies

1. Rules for writing analogies

a. Inclusion rules

29. Use vocabulary that is familiar to the intended examinees.

30. Base the analogy on relationships that are familiar to the examinee.

31. Make the relationship between the first and second word of the stem the same as the relationship between the first and second word of the key.

32. Use only analogies with concise rationales.

33. Allow stem and key pairs to be from different realms.

34. Check all analogies with a dictionary, making sure that the key is the best answer.

35. Check the definitions of the words in the stem and the key in at least one dictionary.

36. Assure that the relationships given in at least two distracters are strong enough to stand as stems in other questions.

37. Assure that parts of speech in the options are parallel to parts of speech in the stem, except where the stem is a pair of nouns or verbs and the options are attributive adjectives or adverbs.

38. Make the words of the first option unambiguous in terms of part of speech.

39. State the rationale succinctly on the back of the item sheet.

Table 3 (Cont'd.)

40. Base the key on the relationship alone and not the subject area of the stem.

41. Use single word analogies whenever possible.

42. Assure that any mix of content areas in the option is appropriate.

43. Assure that at least one distracter has a negative rationale if the stem and key has a negative rationale.

44. Assure that every option has an immediately apparent relationship.

45. Assure that the logical fit between the stem and the key is "tight."

46. Assure that option A unambiguously establishes the part of speech if it is not clearly established in the stem.

47. Assure that the stem clearly establishes the intended rationale.

b. Exclusion rules

48. Avoid analogies which incorporate cliches.

49. Avoid analogies in which the key or stem omits an intermediate step expressed in the other.

50. Check all of the words in the item against the list of overused words and replace all overused words.

51. Check the stem and the key against the word overlap data base and initial the item sheet when the check is completed.

52. Avoid analogies in which the relationship depends entirely on synonyms or antonyms.

Table 3 (Cont'd.)

53. Do not use a reverse key or other trick distracters.

54. Avoid proper names and brand names.

55. Check GRE analogies for overlap in GRE history file.

56. Replace distracters that fit alternate unintended rationales for the stem.

57. Revise options that have merely associational relationships.

58. Check all the options for inappropriately overlapping rationales.

2. General rules for reviewing analogies

a. Inclusion rules

59. Check the relationship in the stem.

60. Check each option to determine whether it fits the relationship.

61. Check rationales for the key and for each option.

62. Suggest ways to improve the key and distracters.

63. Assure that all options have the same parts of speech.

64. Assure that the part of speech used in option A is unambiguous.

b. Exclusion rules

65. Eliminate possible obscure uses of distracters that make them keyable.

66. Eliminate options that can be keyed under a second rationale.

Table 3 (Cont'd.)

67. Eliminate tricky, frivolous or implausible distracters.

68. Eliminate options that do not have strong relationships.

69. Eliminate options that rely on identical relationships but have unintended second-level factors.

70. Eliminate options that are synonym or antonym pairs.

Table 3 (Cont'd.)

C. Sentence Completions

1. Rules for writing sentence completions

a. Inclusion rules

71. Select either original or published sentence.

72. Select sentences of standard and gramatically correct style and vocabulary.

73. Select sentences having independent parts.

74. Select sentences with meanings that are self-contained.

75. Use sentences that are no more than 35 words in length.

76. Use single word options except when including articles or prepositions enables writing plausible distracters.

77. Use options that are parallel in use of acceptable English.

78. Keep gender and racial references in balance.

79. Assure that the sentence meets length requirements.

80. Assure that the sentence addresses a topic that is appropriate for the examinee population.

81. Assure that the sentence conveys a thought as succinctly and clearly as possible.

82. Assure that the sentence incorporates all necessary contextual information.

83. Note all sources for unfamiliar topics on the back of the item sheet.

Table 3 (Cont'd.)

84. Assure that the options are parallel in format and in part of speech.

85. Assure that race and gender codes have been completed for every item.

b. Exclusion rules

86. Exclude sentences that use colloquial expressions, contractions, and/or slang.

87. Avoid cliches, foreign, or familiar sentences.

88. Avoid metaphorical sentences.

89. Avoid sentences in which vocabulary is the only skill tested.

90. Avoid sentences which require specialized knowledge outside of the sentence for completion.

91. Avoid sentences which require subtleties of formal English usage for completion.

92. Avoid using as blanks words on which the meaning of the sentence depends.

93. Avoid writing options and keys that enable the key to be determined by vocabulary level, length, etc.

94. Avoid using as blanks the first words in sentences.

95. Avoid using as blanks words that can only function as prepositions.

96. Avoid using as blanks words that are superfluous to the meaning of the sentence.

97. Avoid identifying the key as being different in any other way from the distracters than in meaning.

98. Do nct use words in options that appear in other options or in the stem.

Table 3 (Cont'd.)

99. Check all options against the list of overused words.

100. Check keys for definitional questions against the word overlap database.

101. Avoid questions with two blanks which can be keyed using only one of the two blanks.

102. Do not use distracters in questions with one blank that are antonyms of the key.

103. Do not write stems of more than 20 words in length for definitional sentence completion questions.

3. Rules for reviewing sentence completion questions

a. Inclusion rules

104. Key each item, considering each option.

105. Reconcile discrepancies with the official key.

106. Compare the options to words that would correctly complete the blanks.

b. Exclusion rules

107. Identify violations of the guidelines for writing sentence completion questions.

108. Identify weak or idiomatically misfitting questions.

109. Check the key for unusual characteristics.

Table 3 (Cont'd.)

4. Rules for classifying sentence completion questions.

a. Inclusion rules

110. Assure that definitional sentence completions depend on linking the definitions of specific words to the sentence.

111. Assure that definitional questions enable examinees to have a specific idea of the word needed before reading the options.

112. Assure that distracters of definitional questions draw on vocabulary knowledge rather than on logical characteristics of the stem.

113. Assure that keying regular sentence completion depends on analyzing logical relationships within the sentence.

114. Assure that the stem of a regular sentence completion item presents a complex and sophisticated thought.

115. Assure that the distracters of a regular sentence completion item can be eliminated more on the basis of logical relationships than on vocabulary knowledge.

b. Exclusion rules

116. Exclude definitional sentences longer than 20 words.

Table 3 (Cont'd.)

D. Reading Comprehension

1. Passage choice rules

a. General inclusion rules

117. Set the proper level of sophistication and density for the examinees.

118. Choose complex passages combining exposition and argument.

119. Choose self-contained passages.

120. Assure proper representation of women and minority groups.

121. Assure the passage is an appropriate reading task for the educational level of the examinees.

b. General exclusion rules

122. Do not choose passages that require specialized information for interpretation.

123. Do not choose passages from current journals or texts.

c. Format inclusion rules

124. Choose passages of about 450 words for long sets, about 150 words for short sets, and passages that fit into desired categories for total number of words.

125. Add orientation phrases as appropriate.

126. Number every fifth line except when it is the last line.

Table 3 (Cont'd.)

127. Assure that the passage represents no more than one-tenth of the original source.

128. Assure that passages drawn from anthologies or collections of essays are less than one-tenth of the original individual essay or literary piece.

d. Procedural inclusion rules

129. Note material use in source documents.

130. Pre-check sensitivity and subject matter accuracy with specialists.

131. Supply relevant information to reviewers about sources that have been used previously.

132. Identify and address specialized terms or assumed background knowledge in contextual cues such as footnotes and the introduction.

133. Assure that the passage is accessible with engaging features, examples, and breathing space.

134. Assure that the introduction to the passage supplies information that is helpful and not merely summarizing.

135. Assure that the passage functions as a complete unit of thought, with a sense of beginning, middle, and end.

136. Choose passages with enough complexity to yield a high proportion of Extended Reasoning questions.

137. Attach a rationale for pairing for a pair of passages.

138. Have 4 - 8 stems and keys submitted for pairs of passages.

139. Assure that the passage is reasonably up to date for rapidly changing fields.

Table 3 (Cont'd.)

140. Attach photocopies of relevant pages from the original source.

141. Attach completed officially initialed (approved) copyright information card for reviewers (2 cards for pairs).

142. Attach photocopy of the copyright page of the original source (for each passage).

e. Procedural exclusion rule

143. Check passage overlap in author file.

144. Check that the author of a passage does not appear on the list of overused writers.

2. Rules for writing reading comprehension questions

a. Inclusion rules

145. Choose the best item and option format.

146. Design EXCEPT questions to have statements that are all true and stems that include the statement "all of the following."

147. Base item difficulty on understanding the passage, not on understanding the question.

148. Refer to the passage and author properly in the stem.

149. Use specific line references in the stem.

150. Use directed stems.

151. State the stem clearly and concisely.

152. Include in the stem words that must otherwise be repeated in each option.

Table 3 (Cont'd.)

153. Have 75-80% of the questions measure Extended Reasoning skills.

154. Assure that overlap questions are critical and distinct and do not extend the number of questions beyond the minimum.

155. Assure that the questions in a set cover all sections of the passage.

156. Include at least a few easy questions in a set, especially early on.

157. Assure that Literal Comprehension questions cover points that are essential to understanding the passage as a whole.

158. Assure that questions testing the main idea and style or tone cover points that are essential to understanding the passage as a whole.

b. Exclusion rules

159. Limit the number of Roman numeral format questions to one or less per set (two for long sets and one for short sets in College Board tests).

160. Limit the number of "EXCEPT" questions to one or less per set.

161. Limit the number of negative stem questions to one for short sets and two for long sets (College Board tests).

162. Avoid revealing test answers in the stem.

163. Keep the numbers of Vocabulary in Context and Literal Comprehension questions within guidelines.

164. Restrict the number of overlap questions in a set to two.

165. Eliminate all unnecessary qualifiers from the stem.

Table 3 (Cont'd.)

166. Eliminate from questions any factors that would allow at least one reviewer to answer correctly without reading the passage.

3. Rules for writing reading comprehension questions based on paired passages

a. Inclusion rules

167. Assure that 25-50% of the questions cover comparisons between the passages.

168. Assure that questions that do not cover comparisons between the passages are reasonably divided between the passages.

169. Cover significant aspects of the pair of passages in comparison questions.

170. Cover distinct, non-overlapping points in comparison questions.

b. Exclusion rules

171. Eliminate factors on the stems of questions in one pair that give away keys to questions on the other passage or on the pair.

Table 3 (Cont'd.)

E. Cognitive Function

1. Contextual rules

a. Inclusion rule for antonyms, analogies, and sentence completions

172. Include Arts and Humanities, Social Studies and Everyday Life, Science and Nature, and Human Relationships and Feelings contexts for questions.

b. Inclusion rules for reading passages

173. Include culturally diverse reading selections.

174. Include passages on biological sciences, humanities, and social studies.

175. Include passages that represent women.

c. General exclusion rules

176. Avoid controversial subjects, e.g., religion and theoretical treatment of evolution.

177. Avoid subjects that are abundant in the item pool.

Table 3 (Cont'd.)

F. Psychometric Properties

1. Difficulty and discriminability

a. Inclusion rule

178. Focus on the field of accomplishment in questions portraying women, rather than portraying success as gender-based.

b. Exclusion rules

179. Exclude Roman numeral format questions from LSAT.

180. Exclude Roman numeral format questions from GRE pretests.

181. GRE use of Roman numeral format questions must control overlap of material in I and in other options.

2. Differential Item Functioning (DIF)

a. Inclusion rules

182. Use a vertical format for analogy item options whenever possible.

183. Use external DIF reviewers.

b. Exclusion rules

184. Reduce non-construct related difficult language in questions.

185. Reduce speededness.

186. Reduce the use of contexts or settings, e.g. sports, war, violence, rural life, that may be differentially interesting or familiar but are not related to the construct.

Five test developers from the College Board and the School and Higher Education Programs Divisions, the two major divisions at Educational Testing Service that produce verbal skills tests, were given the modified rule lists. Each was asked to complete two tasks: rate each of the rules in terms of its importance for developing verbal skills tests, and list for each rule up to two of the flow chart activities for which the rule was important. The test developers chosen for this job had been recommended by their supervisors as those who would best represent the ways in which the rule system is applied in test development.

For each of the verbal item types, the test developers were also asked to add any implicit or explicit inclusion or exclusion rules they use that were not listed. These tasks are described below.

The sequence dimension. The first two dimensions of the operational definition of the verbal domain, then, are the importance of the inclusion and exclusion rules, and the uses of the rules in terms of the test development flow chart. The third dimension of the operational definition is the sequence of application of the rules. Obviously, this is largely defined by the flow chart. Still, some validation is expected from observation of the application of the rules.

A test development specialist who does not work on verbal skills tests provided independent records of test development rule application through observing development of a verbal skills test in each of the two major ETS divisions that produce these instruments. This person also used a series of semistructured interviews to gather the impressions of test developers about how they design the verbal domain.

Survey Instrument Development

Rating the importance of each of the rules used to define the verbal domain and then specifying which activities each rule is important for would be an enormous undertaking. Rating the importance of each of the 186 rules for each of the 30 separate tasks listed on the flow chart would involve completing a 5,580-cell matrix of ratings. This potentially enormous task was reduced on the survey rating instrument as follows:

(1.) Each respondent was asked to rate from 0 (not important) to 4 (critical) the importance of each rule for defining the verbal domain

(2.) Each respondent was also asked to list up to two of the the 30 activities identified in the flow chart, for which each rule most important. The respondents were told to list only activities for which the rule would have been rated as 2 or higher, had it been rated for each activity.

In this way, a very difficult rating task was reduced to a manageable task of 186 importance ratings and up to 372 activity-linkage ratings. The first page of the final form of the survey instrument is provided in Appendix A.

Survey Procedures

Following job analysis procedures, five test developers rated the importance of each of the 186 rules for carrying out the professional activities of defining the verbal domain. The ratings were 4 (critical), 3 (very important), 2 (important), 1 (of little importance), and 0 (not important).

Focus of Observations and Interviews

Clearly, the major part of defining the verbal domain for test developers is in preparing items (III), assembling pretests (IV), and assembling final forms (VII). The observations and interviews focused primarily on these activities,and, as described above, the sequence in which rules were applied to delimit the verbal domain. The rules that were mentioned by the test developers during the course of the interviews had all been linked by the outcomes of the survey to the activity of "preparing items."

Several themes were part of the initial discussions regarding the verbal domain and its definition. One concerned whether or not different contexts of reading passages systematically differed in their comprehension demands and how best to determine what the differences were, if any. A second concerned external verification of the importance of the skills tested by verbal items, and a third concerned the need to document the actual process of developing verbal skills tests, focusing in particular on the rules of inclusion and exclusion that test developers use in creating items. The documentation would require an examination of both the formal and the informal written guidelines of the programs and of the actual procedures test developers use. The current project focuses especially on the second half of the third concern--the rules of inclusion and exclusion that test developers actually use in creating items--with some preliminary exploration into what might be required to address the first and second concerns.

Interview Methodology

The primary means of gathering data was the semi-structured interview. Thirteen interviews were conducted with test developers in School and Higher Education Programs and in the College Board. Eight test developers were interviewed, five twice and three once. The test developers were selected by their supervisors foı participation in the project. They represented different degrees of experience as test developers.

The interviews focused on three areas of verbal item development:

1. the selection and revision of passages for testing reading comprehension

2. the writing of analogy items

3. the writing of sentence completion items

The GRE and the SAT (old and new) were the main focus of the interviews although other tests, in particular the PreProfessional Skills Tests (PPST), served as points of comparison.

A series of questions appropriate to each of the three areas--passage development, the writing of analogy items, and the writing of sentence completion items--was prepared prior to the start of the interviews. These questions served as a basis for maintaining the focus of the interviews, for ensuring comparability in the types of questions asked of the test developers, and for ensuring that an adequate range of topics was covered. Although the interviewer referred directly and conspicuously to the questions periodically, the interviews were conducted in as conversational a manner as possible. Each interview focused on just one of the areas covered in the project--on passages or on analogy items or on sentence completion items.

The interviewer took notes that were later transcribed into a paraphrased question-answer format. If there were points that seemed later to need clarification, the interviewer telephoned the test developer. In one case, the interviewer showed the test developer the transcribed results in order to corroborate the accuracy of the note-taking. The strategy of the semi-structured interview format was used to ensure that the interview was informal and flexible enough to permit the interviewer to obtain both directly and indirectly the kind of information that had bearing on the rules the test developers use in creating verbal items. The test developers were encouraged to add any information or state any concerns that seemed to them worthy of expression. In addition, the interviewer occasionally asked questions that were not part of the prepared questionnaire.

Three sets of draft questions served as the bases for the interviews: one for reading comprehension items, one for sentence completion items, and one for analogy items. The questions focused on the following:

1. the test developers' individual processes for writing, which included obtaining the source materials needed to construct passages or items, identifying a stimulus, making decisions about how to revise and polish the stimulus, writing the test question, and writing the key and the distracters

2. the impact of the specifications on the foregoing

3. the kinds of sources the test developers used for locating appropriate stimulus material

4. the constraints the review process imposed on the manner in which passages and items were developed and, related to that, the relative ease or difficulty with which passages and items having certain characteristics, as identified by the test developer, seemed to survive the review process

5. the relative ease or difficulty with which passages and items in the subject areas--humanities, social sciences, science, and human interest--survived the review process

6. the test developers' thoughts about what the items measure, based on their experiences in exercising the skills required to create the items

Additional means of gathering information included observation of one session of a passage review meeting in the SAT program and subsequent perusal of the review comments for one of the three passages reviewed at the meeting.

In the early phases of the project, two test developers--one representing science and the other humanities--were asked to select an article or section of a textbook that they judged to be representative of the field and to "track" their comprehension processes--that is, to take conscious note of the strategies they were using in order to comprehend the text. The results of this initial exercise indicated that such an exploration was beyond the scope of the current project.

Limitations of Interviews

In a few instances, the test developers who were interviewed indicated that their jobs now require more coordination and training than actual development of passages or items. Their answers to questions consequently were based on their recollection of the procedures they used in writing and on their overall understanding of the workings of development in their programs. These test developers, however, especially in their role as trainers of new test developers, regularly review passages and items written by others. Thus, the interviews reflect a range of types of involvement with the test development process and, in conjunction with that, a mix of "takes" on the request that they describe their individual procedures. At one end, one test developer initially answered questions from a programmatic rather than an individual perspective, and it was apparent that recall of the actual procedures this person used required some pause and reflection. At the other end, one test developer allowed the interviewer to observe while she constructed an item from start to finish. That test developer's list of reasons that items are lost at first review was drawn directly from item copy she had retained. The other 11 interviews--the two persons just described were each interviewed only once--consisted of recollections made by test developers without looking at item copy, test developers for whom writing items is a regular part of the demands of their jobs.

The interviews and observations do not include the development of antonyms. It was decided to focus the interactive activities on the item types that were in common to the large testing programs reviewed. Antonyms are no longer used on the SATs. Similarly, the cognitive and psychometric areas were not treated separately, but were incorporated into the investigations of development of the sentence completion, analogy, and reading comprehension item types.

Document Review Results

General Descriptions of the Semantic relations of the Verbal Domain

Antonyms. Four types of oppositional relationships have been distinguished among antonyms: contradiction, contrariety, polar contrariety, and converse relationship. Contradiction is the strongest kind of opposition because the relationship requires that two statements are made about something that cannot both be true and both be false. Contrariety is not as strong as opposition and should only be used in antonym items when there is a strong, defensible key. Polar contraries are permissible only if the distracters are not on the same continuum as the stem and key. Polar contraries should involve extreme positions, e.g., right:left rather than an extreme and a midpoint, e.g. right:middle.

In this study, the antonym rules had varying use, because antonyms are no longer an item type of the new Scholastic Aptitude Test. The reader is cautioned that many of the rules discovered in the study regarding antonym development did not have very high average ratings of importance by test developers because they have limited relevance to the work of test developers in the College Board division.

Analogies. Fifteen types of analogical relationships are sampled in verbal skills tests: class:subclass; quality or subject:associated object or symbol; word:defining, descriptive, or associated action or trait; word:rough synonym; cause:effect; place:inhabitant; measuring units:object; action:action's meaning; word:antonym; occupation:characteristic of occupation; part:whole; object:components of object (or smaller sized object); object:source of object; description or described object:intensified description; object:function of object.

For the most part, analogy rules are concerned with making the vocabulary and structure of the items easy to interpret. In this regard, rules for analogy items require that they are presented in familiar language and represent familiar relationships and that the relationships between the elements presented in the item stem and the elements in the correct options (the keys) are parallel.

Sentence completions. Sentence completion items are designed to sample three major types of semantic relations: contrasting, in which two halves of a sentence have opposite meanings; definitional, in which the rest of the sentence defines one or more of its key words; and causal, in which one half of a sentence is the cause of an effect in the second half.

Reading comprehension. Reading comprehension items are different in structure from other verbal item types. For other verbal item types, the information needed to respond correctly must come from the item stem. For reading comprehension items, the information needed to respond correctly cannot be located in the stem (memorandum from Adams, April 1991) but must be drawn from the passage. Thus, the nature of the stimulus (the passage)and the nature of the task (answering a question about information) are both unique to these items, as is the ability that is measured.

For reading passages, some selection rules are inclusion rules because they require reading passages to be selected from the realms of biological science, humanities, and social studies (Donlon & Angoff, 1971), and to represent cultural diversity and women. Other rules, concerned with excluding material from reading comprehension sections of tests, fall into three broad categories (memorandum from Adams, August 1991): Differential Item Functioning (DIF, discussed below); controversy including religion and theoretical treatment of evolution, and topics overrepresented in the item pool, including the Harlem Renaissance, thé Asteroid Impact Hypothesis, Stellar Evolution, Greenhouse Warming, Pre-Columbian American-Indian Diseases, Native American passages (LSAT only), and Painters (LSAT only).

General Considerations for Delineating Procedural Rules

Familiarity and interest. Recent contributions of cognitive science have increased attention to the effects of contextual variables like familiarity and interest on test performance. Traditionally, research had been concerned more with how such structural features as sentence and word length and usage affect reading comprehension, and had ignored cognitive processing variables such as linguistic knowledge, that are so important to semantic interpretation (Anderson & Davison, 1988). Higher cognitive processes are sensitive to examinee variables such as interest and background.

The fact that every examinee is not equally familiar with or interested in the context of every test question raises a problem of how context might be standardized to avoid systematically advantaging or disadvantaging examinees. Three possible approaches to this problem are:

1. Make the context as universally accessible or equally familiar and interesting to all examinees as possible.

2. Develop context-specific tests that are functionally equivalent in familiarity and interest for different populations.

3. Balance contexts, so that those that are less familiar or interesting to some examinees appear in the same test with those that are more familiar or interesting.

Solution (3) assumes that the degree of familiarity or interest can be gauged and balanced for every examinee. Solution (1), on the surface, seems more promising if a corpus of cultural knowledge exists that is accessible and interesting enough to all examinees and if variations in familiarity or interest in this corpus do not affect performance on measures of verbal skills. Solution (2) is under study for the GRE analytical measure (Emmerich, Enright, Rock, & Tucker, 1991), and offers some promise.

For now, though, the problem is monitored through DIF techniques, and research on solutions continues. The rules that exist concerning DIF values and material associated with high DIF values are listed as psychometric concerns (see below).

It is important, here, to note that sensitivity concerns are not necessarily DIF concerns. Sensitivity is more cognitive in orientation, because the goal is to maintain consistency across examinees in their familiarity with the subject matter of the test or in their emotional response to the test items. The original sensitivity guidelines used at Educational Testing Service (Hunter & Slaughter, 1980) contained many strictures against uses of certain words or portraying people in stereotypical or pejorative ways. There was also an attempt to have material included in test items that reflects "the multicultural nature of our society" (p. 5). Over the years, these considerations became internalized on a corporate level, and the latest version of the guidelines contain relatively less discussion about words or stereotypes to be avoided and relatively more discussion on how to represent various population groups in tests that measure different constructs, e.g., content and skills tests.

Item psychometric properties. Specifying desired item psychometric properties helps define the verbal domain through including items with certain characteristics and excluding items without those characteristics. The three characteristics that are routinely investigated are difficulty, discrimination and differential difficulty. These can be measured either by conventional methods--for example item delta values, item biserial correlations, and Mantel-Haenszel delta values--or by Item Response Theory parameters. Moreover, estimates of these characteristics can be made after pretests to define the item pool for final test forms, after administration of final test forms but before scoring (preliminary item analysis or PIA), or after final form scores are reported (item analysis or IA).

It is important to note that for all the verbal item types of the GRE, GMAT, LSAT, and SAT, the criterion measure used both for biserial correlations and for DIF analyses is the total verbal raw score. As a consequence, items that remain within the tolerable limits for these statistics, and hence contribute to final form test scores, are those that support a unidimensional verbal score, that is, that elicit examinee responses based on the same skill. Items that have biserial values below .20 or are very hard or very easy are flagged by statistical consultants for review by test developers and outside experts. Items that elicit differential responses from examinees who have the same level of skills, as observed in Mantel-Haenszel absolute delta values exceeding 1.5 and significantly greater than 1.0, are also flagged for review by a DIF committee.

There are no inclusion or exclusion rules that operate a priori on the basis of item difficulty or of discrimination. However, there are rules that control the use of Roman numeral format items, presumably to maximize their measurement properties (Adams, October 1990) and rules that relate to the inclusion of material representing the accomplishments of various population groups.

The Mantel-Haenszel technique is the standard analysis tool for estimating and reviewing DIF. Another techniques, the standardization procedure, is also used with the SAT items. Both procedures rely on matching two groups of examinees who achieve the same overall level of performance but who belong to different population subgroups (Holland & Thayer, 1986). Generally, these groups are referred to as the focal group and the reference group. Although these groups can be theoretically any two defined populations, in practice, for DIF analyses comparing male and female performance, the focal group is female examinees (even if they are the majority of test takers), and the reference group is male examinees. For DIF analyses comparing ethnic groups, the focal group is an ethnic minority group, e.g., African American, Asian American, or Hispanic examinees, and the reference group is White or European American examinees.

Research (Kingston, Schneider, & Briel, 1988; Stricker & Rock, 1985) generally supports using the verbal scores to match examinees in order to evaluate DIF of constituent items. Because the groups have been matched on the relevant skills, observed differences in group performance may sometimes be interpreted as examinees responding to item characteristics other than those intended to elicit the response (e.g., differences in the experience of different groups of examinees). In the pure sense, DIF is not only concerned with the question of the differential difficulty of the items, but also, like a microcosm of factor analytic studies, it is sensitive to differences in what the item measures among different groups of people.

Although test items that function differently for different subgroups have been empirically identified, the factors causing the differences may well be related to differential interest and familiarity. Research is still being conducted, however, and factors may be identified in the future which will give rise to item writing rules about which topics, structures, or kinds of semantic relations to avoid. The many research findings regarding DIF are beyond the scope of the current study, but specific research outcomes related to formation of the verbal domain are briefly reviewed below.

African American and Hispanic examinees are more likely to perform poorer than matched groups of European American examinees on analogy items than on the other three verbal item types. Wild, McPeek, and Koffler (1988) found that any three of the four item types in the verbal measure would be as valid as are all four together and that analogy items may contribute slightly less than the other three item types to concurrent validity. Cole (memo, June 1990) reports that easy analogies in particular tend to have elevated DIF values. DeMauro (1990) found that focal groups are more likely disadvantaged when they are a small proportion of the population and the item is easy. Remember, an item will be easy in general if it is easy for a large proportion of the population. If the reference group is a very large proportion of the examinee population, and the item is easy in general, it is likely that the item is easy for reference group members, and the item being easy for reference group members is a necessary but not sufficient, condition for finding DIF.

General exclusion rules pertaining to DIF are summarized in a memorandum from Cole (June 1990):

1. Use a vertical format for analogy item options, unless impossible.

2. Use external DIF reviewers.

3. Reduce visual material that is not construct related.

4. Reduce non-construct related difficult language in questions.

5. Reduce speediness.

6. Reduce use of settings or contexts that may be differentially familiar or interesting but are not related to the construct.

A memorandum from Adams (December 1990) and a second memorandum (Cole, February 1993) went on to specify material that was to be avoided, based on DIF analyses:

1. horizontal formats for analogy question options

2. unfamiliar vocabulary or contexts related to legal, political, financial, scientific/technological, farming or rural life, construction, or transportation matters

3. subject matter related to war, violence, or weaponry

4. Subject matter related to sports

This memorandum also advised test developers to avoid certain question characteristics whenever these were not required specifically by test content specifications:

1. convoluted, difficult, or unfamiliar vocabulary or syntax

2. charts, maps, and graphs

3. material requiring spatial skills

Remember that Mantel-Haenszel delta values, the DIF index, sum to 0 over the criterion measure (in this case the total verbal score) as a function of matching. It follows that if a group of examinees performs differentially better on one segment of the criterion, the rest of the examinees must perform differentially better on the rest of the criterion.

Survey Results

Criteria of Importance

There are two methods for analyzing the survey data--examining the mean rating and examining the distribution of ratings. We are particularly interested in those rules that did not attain mean ratings of at least 2.5 (the upper bound of "important") and those rules that were given ratings of less than "important" (2) by a majority of the respondents. Actually, with five raters, if the "importance" rating is not given by a majority of raters (3 or more), it is not possible to achieve a mean rating of 2.5.

Ratings (Appendix B provides the survey results; see also Table 4)

Of the 28 antonym rules, 9 (28 percent) failed to meet either the majority criterion of being rated important by 3 or more respondents or the 2.5 mean criterion, and 4 (14 percent) failed only to meet the 2.5 mean criterion. Of the 42 analogy rules, 4 (10 percent) failed to meet either criterion, and 3 (7 percent) failed only to meet the 2.5 criterion. Of the 46 sentence completion rules, 3 (7 percent) failed to meet either criterion, and 9 rules (20 percent) failed only to meet the 2.5 criterion. Of the 55 reading comprehension rules, 3 (6 percent) failed to meet either criterion, and 13 (24 percent) failed only to meet the 2.5 criterion. Table 4 shows the rules that failed to meet one or both of these criteria.

Many of the rules that involved the least semantic judgment from the test developers failed to meet one or both criteria of importance. For example, Rules 50 and 99 require test developers to remove words that are overused from the items. Neither of these rules met either criterion for importance. Similarly, Rule 144 requires test developers to check passage authors against a list of overused writers, and the mean rating for this (2.2) failed to meet the importance criterion. The importance ratings, and indeed, the rules themselves, affirm our impression that the test developers view their job as judging the semantic relations involved in the stimulus materials, and they view the opportunity to make these judgments as more important to the p[erformance of their jobs than the opportunity to carry out judgments already made about material that best presents the semantic relations.

The most interesting of the results were in the areas of cognitive functioning and psychometric properties. All six cognitive functioning rules met both criteria of importance, but none of the nine psychometric properties rules met both criteria. In fact, five of these rules (56 percent) failed to meet either criterion, and the other four mean ratings below 2.5 but did meet the majority criterion. The reader will note that the cognitive functioning rules concern the familiarity and interest of the subject matter of test items, but the psychometric properties rules concern item features that have been associated with some psychometric properties.

A mixed hierarchical general linear model analysis of variance was made to evaluate whether rules that applied across testing programs and item types were rated higher than were rules that were more specifically concerned with an item type or testing program. All rules were classified as being general or specific, and variance was partitioned by raters, and this classification of the rules (see Myers, 1972, for explanations of models of this type). The mean ratings for the 54 general rules (3.15) was significantly higher than the mean rating for the 132 specific rules (2.82) (F (df=1,4) = 9.57, p .04). When the description of ratings of this magnitude is considered, however, it is not clear that these difference have any meaning, since both mean ratings are within the range of the "very important" rating (3).

Reliability of Ratings (Table 6)

The reliability of the ratings was assessed over rules and over judges. This was accomplished by partitioning the ratings made by the judges into three components of variance: judges, rules, and judges by rules. The mean square for each component was computed, and reliability was given by the following two formulas:

Interjudge: $$\frac{MS(j) - MS(j \times r)}{MS(j)}$$

Interrule: $$\frac{MS(r) - MS(j \times r)}{MS(r)}$$

Where MS(j) is the mean square for judges, MS(r) is the mean square for rules, and MS(j x r) is the mean square of the interaction of judges by rules. The reliabilities were computed for each of the four item types, and for cognitive functioning and psychometric properties. These reliabilities are given in Table 6.

Comments of Survey Respondents

Comments about rules. The respondents wrote many comments in the provided spaces and throughout the survey document. Some were either paraphrases of a rule itself or of the description associated with the numerical rating given to it. These types of remarks seemed to be a type of verbal mediation used by the respondents to help them complete the survey and are not summarized below. However, a large number of comments are summarized below that provide information about how test developers interpret and apply the rules.

One of the respondents commented that at least one of the rules involving antonyms might have several interpretations. For example, having a strong defensible key of opposing elements is very important if distracters are used that have only part of their meanings in opposition to the meaning of the stem word (Rule 1). For this respondent the importance of the rule rests on the interpretation of strong key being that the key is more oppposite in meaning to the stem word than is any of the distracters.

This respondent also remarked that the familiarity of words could be checked against the Breland/Jones Word List (Rule 10). With regard to the psychometric property rule about using outside Differential Item Functioning (DIF) reviewers, the same respondent stated that this was unnecessary on the tests she works on because items with large DIF pretest DIF values are eliminated (Rule 183). This practice may become a rule. Current DIF rules require a review of items with large DIF values.

Table 4

Mean and Median Ratings
by Item Type and for
Cognitive Functioning
and Psychometric Properties
(0=not important, 4= very important)

Item Type	No. Rules	Mean	S.D.	Median
Analogy	42	3.08	1.23	4.00

High Rule[3]: 31. Make the relationship between the first and second words of the stem the same as between the first and second words of the key.

39. State the rationale succinctly on the back of the item sheet.

51. Check the stem and the key against the word overlap data base and initial the item sheet when the check is completed.

59. Check the relationship in the stem.

Low Rule[4]: 69. Eliminate options that rely on identical relationships but have unintended second-level factors.

[3] The high rules were rated as "very important" (4) by all five respondents.

[4] The low rules all failed both criteria of importance and were the lowest-rated rules for the item type, for cognitive functioning or for psychometric properties.

Table 4 (cont'd.)

Antonym	28	3.72	1.59	3.00

High Rule: 9. Consider the connotative and denotative meanings of words.

28. Eliminate distracters that do not conform to the general rules for writing antonyms.

Low Rule: 2. Use extreme positions for polar contraries.

Cognitive Func.	6	3.67	0.60	4.00

High Rule: NONE (All were rated at least "important" by all five respondents, but none achieved mean ratings of 4.0.)

Low Rule: NONE

Psychometrics	9	1.49	1.63	0.00

High Rule: NONE

Low Rule: 181. GRE use of Roman numeral format questions must control overlap of material in I and in other options.

Table 4 (cont'd.)

<u>Reading Comp.</u> 55 2.99 1.29 3.00

High Rule: 117. Set the proper level of sophistication and density for the examinees.

120. Assure proper representation of women and minorities.

121. Assure the passage is an appropriate reading task for the educational level of the examinees.

122. Do not choose passages that require specialized information for interpretation.

136. Choose passages with enough complexity to yield a high proportion of Extended Reasoning questions.

143. Check passage overlap in author file.

145. Choose the best item and option format.

Low Rule: 154. Assure that overlap questions are critical and distinct and do not extend the number of questions beyond the minimum.

<u>Sentence Comp.</u> 46 2.99 1.29 3.00

High Rule: 90. Avoid sentences which require specialized knowledge outside of the sentence for completion.

104. Key each item, considering each option.

105. Reconcile discrepancies with the official key.

Low Rule: 73. Select sentences having independent parts.

Table 5

Rules That Failed to Have
at Least Three Respondents
Rating Them as Important or
That Failed to Achieve Mean Ratings
of at Least 2.5
(Table entries are rule numbers.)

		Rated Important by		
Area	Mean	at Least 3 Raters	Fewer than 3 Raters	Total
Analogies	Less than 2.5	8, 10-12	1-7, 13	12 rules
	2.5 or Greater	9, 14-28		16 rules
	Total	20 rules	8 rules	28 rules
Antonyms	Less than 2.5	33, 41, 52	30, 50, 69-70	7 rules
	2.5 or Greater	29, 31-32, 34-40, 42-49, 51, 53-68		35 rules
	Total	38 rules	4 rules	42 rules
Reading Comp.	Less than 2.5	119, 123-124, 134, 137, 144, 149, 153, 155-157, 166	146, 154, 164	15 rules
	2.5 or Greater	117-118, 120-122, 125-133, 135-136, 138-143, 145, 147-148, 150-152, 158-163, 165, 167-171		40 rules
	Total	52 rules	3 rules	55 rules

Table 5 (cont'd.)

		Rated Important by		
Area	Mean	at Least 3 Raters	Fewer than 3 Raters	Total
Sentence Comp.	Less than 2.5	74, 83, 85, 88, 92, 103, 112, 116	73, 89, 99	11 rules
	2.5 or Greater	71-72, 75-82, 84, 86-87, 90-91, 93-98, 100-102, 104-111, 113-115		35 rules
	Total	43 rules	3 rules	46 rules
Cognitive Funct.	Less than 2.5			0 rules
	2.5 or Greater	172-177		6 rules
	Total	6 rules	0 rules	6 rules
Psychomet. Prop.	Less than 2.5	178, 182, 184, 186	179-181, 183, 185	9 rules
	2.5 or Greater			0 rules
	Total	4 rules	5 rules	9 rules

Table 5 (cont'd.)

		Rated Important by		
Area	Mean	at Least 3 Raters	Fewer than 3 Raters	Total
All Areas Prop.	Less than 2.5	31 rules	23 rules	54 rules
	2.5 or Greater	132 rules		132 rules
	Total	163 rules	23 rules	186 rules

Table 6

Reliabilities of Ratings of the Importance of Inclusion and Exclusion Rules, for Four Item Types, Cognitive Functioning, and Psychometric Properties

	Anal.	Ant.	Cog. Func.	Psych. Prop.	Read. Comp.	Sent. Comp.
Number of Rules	42	28	6	9	55	46
General Linear Model Regression						
Mean Square raters	0.98	6.13	1.42	5.87	1.96	6.44
Mean Square rules	2.99	5.46	0.21	2.06	3.12	3.39
Error (raters x rules)	1.16	1.65	0.20	2.42	1.29	1.14
Reliability (rules)	-0.18	0.73	0.86	0.59	0.34	0.82
Reliability (raters)	0.61	0.70	0.07	-0.18	0.59	0.66

A second respondent noted that the rules used for reviewing antonyms often involve procedures that item writers follow before items are reviewed, so that there is some redundancy in the application of the rules. This comment implies that rules governing the review process are being incorporated into item development. Similarly, the respondent remarked that rules for item development of one item type have already been formally incorporated into development of other item types. For example, some analogy and sentence completion rules overlap, e.g., analogy rules concerning the relative quality of the distracters and the key (e.g., Rule 36), and those sentence completion rules concerning the relative quality of distracters and the key (e.g., Rule 106). These comments illustrate two stages in rule development, implicit extension of rules followed by formal extension of rules.

This respondent also stated that it is acceptable to use some overused words on analogy items in final test forms (Rule 50; see below, the SHEP respondent stated that there is no check of overused words.) and that, among sentence completion items, some logical inference is always necessary (Rule 112).

The second respondent noted that reading comprehension items that use the word "EXCEPT" in the stem are more varied than what is implied by the rule (Rule 146) about their form. This respondent also explained that line references are not always used, particularly for forms for examinees who have disabilities (Rule 149). The respondent also felt that it should not be claimed that "no reading comprehension passages can be keyed independently of the passage." Rather, it is the aim of test developers to eliminate clues and cues from the items (Rule 166). The third respondent also felt that the item reviewers used by ETS test development are able to key at least some test items without reading the passage.

The third respondent, who is the first among the three from the SHEP Division, stated that there is no overused word list against which to check either analogies (Rule 50) or sentence completion items (Rule 99) and that no list exists of overused writers against which to check reading comprehension passages (Rule 144). The respondent also had two comments about rationales for Analogy items: Rationales for distracters are not written down (Rule 61), even though they might exist (the rule implies they do not have rationales), but distracters should not have rationales better than the rationale for the key (Rule 66).

The third respondent also felt that the rule requiring that sentences be grammatically correct is always guaranteed by the editing process (Rules 72 through 75). In sentence completion items, blanks should not be conjunctions or prepositions (Rule 95). The respondent also stated that, for definitional sentence completion items, assuring that the solution depends on linking the definitions of specific words to the sentence (Rule 110), eliminates the necessity of assuring that distracters draw on vocabulary knowledge rather than on the logic of the stem (Rule 112).

The third respondent claimed that material from which science passages are drawn should be current, but not from the month in which the test is being assembled (Rule 123). She also felt that passage lengths vary considerably both between programs and within the GMAT program (Rule 124).

The fourth respondent, also from the College Board Division of ETS, stated that there are no exceptions for attributive adjectives or adverbs to the rule (Rule 37) that states that the parts of speech in the analogies options should be parallel to the parts of speech in the stem. This respondent also noted that the new SAT specifications enabled analogies to have antonym or synonym relationships (Rule 52). This was also mentioned by the first and second respondents. In sentence completion items, the respondent felt that selection of either original or published sentences was a very important rule, but that sentences change over the course of test development (Rule 71). This respondent also stated that exposition should be avoided in reading comprehension passages (Rule 118).

The fifth respondent to the survey was also from the SHEP division. She had no comments about specific rules.

The respondents' comments are all consistent with our functional analysis of the verbal domain. That is, the rules are all evaluated in terms of their effect on the ability of test developers to perform the required semantic interpretations.

General comments and suggested rules. The respondents were invited to add whatever rules they use for each item type that were not listed on the survey, to rate the importance of the added rules, and to list up to two test development activities for which the rule was most important. This was done by the first, third, fourth, and fifth respondents.

The first respondent stated that elitist, religious, or insensitive words should be excluded from antonyms, and rated this rule as 3, or "very important" (this was also added by the fourth respondent; see below). This rule would be used for reviewing test items of peers (decision point IIIC2a) and reviewing draft tests (IVA1). The respondent added a rule for test developers to avoid analogies that depend on certain attitudes, e.g., political attitudes, and rated this rule as "critical" (4). This rule was held important both for writing or designing test specifications and for writing new items. For reading comprehension, the respondent stated that only passages approved by the passage review panel should be used, and rated this as "critical" (4), but did not specify particular decision points. Finally, the respondent added two psychometric property rules, and rated them both as "critical" (4), and said they were both important for assembling final draft tests (VIIA) and performing global reviews of them (VIIA1). These rules were to meet the point by point delta (item difficulty) specifications for the test and to meet the biserial correlation specifications for the test.

The third respondent added an antonym rule: Eliminate all distracters that are keyable. She rated this rule as "critical" (4), and stated that it was important for developing new test items (IIIC) and preparing draft tests (IVA). The respondent also added two sentence completion rules. The first, rated as "very important" (3), was to exclude sentences containing thoughts that go too strongly against readers' expectations or common knowledge. This was held to be important for selecting item stimuli or sources (IIIB) and developing new test items (IIIC). The second new rule was held to be important for the same two areas (IIIB and IIIC). It was not to use sentences that contain libelous assertions about people who are alive and can sue and not to use sentences that refer to living political figures, corporate products by brand name, or corporations. This was rated as "critical" (4).

The fourth respondent also stated that antonyms should avoid using elitist terms or words that have a religious association and gave this a "very important" (3) rating, but did not specify the decision point for which this was most important. This respondent also added that overused words should be avoided, and rated this new rule as "critical" (4), but again did not list a specific decision point for which it was important.

The fifth respondent added two rules to the psychometric properties rules and rated each of these as "critical" (4). These were to perform a passage precheck and to perform an informal sensitivity review. These rules were not associated with any specific decision points.

More global comments were made by the first and fourth respondents, as well. The fourth respondent felt that the association of decision points with the rules was arbitrary since the test development process involved an integration of several rules and their sometimes iterative and sometimes simultaneous application in several decision points. If this comment is true, it would follow that the decision points might not be thought of as discrete entities, but rather as hierarchically organized goals for test development, e.g., to evaluate how well the item samples the domain.

Choice of most important rules. The first respondent felt that it might be a more informative to the study to have test developers choose the ten rules for each item type that are most important to the overall work of test development. The surveys were sent out again, with new instructions to choose the ten most important rules for each item type and for cognitive functioning an for psychometric properties (obviously, where there are fewer than ten rules, it was expected that only a few of the rules would be identified as most important). A senior examiner, an examiner, and an associate examiner from the College Board Division completed the assignment, but the respondents did not all identify ten rules for each area as most important for that area (item type or cognitive functioning or psychometric properties). Some listed more rules, and one did not list any rules for antonyms.

Because the Scholastic Aptitude Test has dropped antonyms as an item type, one respondent did not identify the ten most important rules for antonyms. This respondent also did not identify the 10 most important rules for either psychometric properties or cognitive functioning. Table 7 shows the mean ratings that had been given earlier to the rules chosen as most important in each area. Table 8 shows the agreement among the three test developers on the 10 most important rules. General linear regression reveals that the rules identified as most important by the first respondent for antonyms, cognitive functioning, and reading comprehension; the rules identified as most important by the second respondent for analogies, antonyms, and reading comprehension; and the rules identified as most important by the third respondent for reading comprehension all had significantly higher mean importance ratings than those rules that were not identified as important.

Linkage to Job Activities (see Table 9)

Each of the five original respondents was asked to link each of the 186 rules to two activities from the revised flow chart of decision points (table 2) for which it was at least "important." This leaves a possibility of 1,860 linkages. In fact, there were 739 blanks and 1,121 linkages. Of these, 433 (38.6 percent) of the linkages were made to job activity IIIC2, which is writing new test items, and another 231 (20.6 percent) of the linkages were made to reviewing test items developed by peers, activity IIIC2b. In all, category IIIC, developing new test items (including writing, classifying, and reviewing new test items) accounted for 817 (71.9 percent) of the 1,121 linkages. The other linkages of rules to activities for which they were important were as follows: IVA1, performing global reviews of draft tests, accounted for 82 (7.3 percent) of the linkages; IIIB, selecting item stimuli and sources accounted for 61 (5.4 percent); IVA, preparing draft tests, accounted for 52 (4.6 percent); VIIA1, performing global reviews of draft final test forms, accounted for 49 (4.4 percent); and VIIA, preparing draft tests, accounted for 24 (2.1 percent).

In reviewing these linkages, it is clear that the major activities of the test developers in which the rules are important are those of preparing items (III), assembling pretests (IV), and assembling final test forms (VII). In fact, all but 16 of the 186 rules were linked to some activity by at least three of the respondents, and all but 31 were linked to a specific type of activity (activities at the Roman numeral level on the flow chart) by a majority of respondents. Table 8 shows the linkages.

Table 7

Mean Importance Ratings Given
by All Five Respondents to the Most Important Rules
for Defining the Verbal Domain, As
Chosen by Three Test Developers[5]
from the College Board Division
(Numbers of Important Rules Identified
Appear in Parentheses)

Area		Senior Exam.	Exam.	Associate Exam.
Analogies	most imp.	3.18 (8)	3.58 (10)	3.16 (10)
	not	3.05	2.92	3.05
Antonyms	most imp.	3.38 (12)	3.13 (9)	----
	not	2.23	2.53	----
Cognitive	most imp.	3.72 (5)	3.67 (6)	----
	not	3.40	----	----
Psychomet.	most imp.	1.67 (3)	1.20 (1)	----
	not	1.40	1.53	----
Reading Comp.	most imp.	3.32 (15)	3.36 (10)	3.32 (20)
	not	2.87	2.91	2.80
Sentence Comp.	most imp.	3.22 (11)	3.06 (10)	3.12 (10)
	not	2.91	2.97	2.95

[5] The mean ratings are for the rules identified as most important and for the remaining rules of that item type, for cognitive functioning, or for psychometric properties.

Table 8

Agreement Among Three Test Developers from the College Board Division on the Most Important Rules for Defining the Verbal Domain (Proportion of Rules Agreed Upon)

		Senior		Associate
Area		Exam.	Exam.	Exam.
Analogies				
	With Sr.	----		
	With Exam.	0.67	----	
	With Assoc.	0.76	0.62	----
Antonyms				
	With Sr.	----		
	With Exam.	0.68	----	
	With Assoc.	----	----	----
Cognitive Functions				
	With Sr.	----		
	With Exam.	0.83	----	
	With Assoc.	----	----	----
Psychometric Properties				
	With Sr.	----		
	With Exam.	0.78	----	
	With Assoc.	----	----	----
Reading Comprehension				
	With Sr.	----		
	With Exam.	0.73	----	
	With Assoc.	0.80	0.71	----
Sentence Completion				
	With Sr.	----		
	With Exam.	0.67	----	
	With Assoc.	0.76	0.70	----

Table 9

Numbers of Rules Linked by Three or More Respondents to the Eight Superordinate Test Development Activities[6]

Linked by Three or More Respondents to	Rule Numbers	Total Number of Rules
No Activity	2-6, 50, 70, 73, 80-81, 164, 178-179, 183, 185-186	16
No Specific Activity	1, 7, 61, 66, 69, 78, 106, 124-128, 172, 176, 182, 184	16
Activity Type III Only (Prepare Items)	8-21, 23-49, 51-60, 62-65, 67-68, 71-72, 74-77, 79, 82-104, 109-123, 129-163, 165-171, 177, 180-181	147
Activity Type IV Only (Assemble Pretest)	108	1
Activity Type VII Only (Assemble Final Form)	173-175	3

[6] If a respondent linked a rule to one or two activities with the same superordinate (Roman numeral) category, it was counted as one link.

Table 9 (cont'd.)

Activity Types III and IV Only	22, 105, 107	3
Activity Types III and VII	none	
Activity Types IV and VII	none	

Summaries of the Outcomes of the Interviews

Overview

The first set of interviews focused on having the interviewed test developers members explain how passages are chosen to serve as stimuli for reading comprehension items. The summaries of the responses of these test developers are organized according to the six areas of the interview questionnaires (four item types, cognitive functioning, and psychometric properties), described earlier in the Interview Methodology section.

A constant concern of the test developers is the capacity of the item to pass all of the required reviews. The rules governing reviews are focused on the clarity with which the items present material that must be interpreted by the examinees.

Reading Comprehension

Individual test developers' procedures. The process of looking for a passage begins with skimming or "flipping through" materials the test developer considers to be of appropriate level of difficulty. Five or six potential candidates for passages are noted, based on how interesting they are to the test developer; whether they are well written and clear; whether they have elements of "tension"--that is, argument, comparison, contrasting points of view, pros and cons, and the like. The test developer looks for relatively self-contained treatments, ones that would require a minimum of cutting and pasting to become a unified whole. Though this last consideration seems most important to the new SAT, it also figures importantly in the selection process in both the GRE and the PPST. Total rewrites of sources are seldom done, even by experienced test developers. Note that these concerns are addressed by rules 133-136.

Having focused on the best possibilities according to the criteria above, the test developer begins to clarify and to tighten each potential passage, eliminating digressions and ensuring that outside knowledge is not required to comprehend the text (Rule 132). The passage that best stands up under this process in the test developer's judgment is the one that is submitted to a reviewer.

The above process is not consciously guided by a consideration of the test specifications--the kinds of items that must be asked based on the passage content. Items about the main idea or main purpose of a passage may influence passage selection, especially for easier reading comprehension tests (like the PPST in Reading, another instrument developed by this same group of professionals), because such items give the test developer an early indicator of the unity of the passage. In general, test developers figure that if a passage contains sufficient tension, as described above, the passage will support the required number and types of items.

Kinds of sources. Test developers use a variety of materials for sources of passages. For the GRE they use popular magazines "of appropriate level of difficulty," literary magazines, journals, reviews, and critical commentary. Textbooks are not used for any of the testing programs at ETS. Textbook content is considered to be too specialized or likely to be familiar to examinees. Materials for a general audience are preferred, with some restrictions, depending on the program. For example, SAT test developers avoid popular magazines altogether (Rule 123) and for science passages avoid works by popular writers, however well written, because these writers tend to be overused.

Books as sources for passages have a mixed reception, owing to the different work requirements concerned with fair use of passages. When looking for sources, test developers use libraries, their personal book collections, and magazines at hand. They are likely to have a few favored sources for passages about women and ethnic minority groups because it is most difficult to obtain materials in these areas that are interesting but not controversial or inflammatory (Rule 176).

Some test developers avoid using books as sources because books tend to be more discursive in treating subject matter than are studies or reviews; books are not as likely to come quickly to the point and to develop the point without digressions.

Impact of the review process. Content for passages that is sensational and "troublesome," in that it violates or comes close to violating sensitivity guidelines (Rule 130), does not fare well in the review process. On the other hand, test developers try to avoid content that the reviewers may perceive as bland.

The GRE (and PPST) review process favors concreteness--facts and a discussion of facts, opinions and counter opinions, theory and counter theory, all supported with specific examples and evidence. In particular, the GRE review process demands that passages have considerable "density"--that every sentence represent some substantial development or extension or refutation of previous ideas (Rule 118); that description, narration, and restatement be kept at a minimum.

Reviewers avoid passages that are very abstract. While test developers differed in their criteria of "abstract," they stated that highly theoretical sociology, history of science, philosophy of science, or philosophy in general would be not acceptable to reviewers. GRE reviewers tend to avoid science subject matter that does not contain factual and concrete processes and descriptions (Rules 117 and 122). Because the new SAT uses passages of 850 words in length, reviewers are receptive to and welcome passages that contain a mix of styles, including narrative and theoretical for science (Rule 118), as long as the passage can support the number of analytical and literal items the specifications require (Rule 136).

As stated under the test developers' procedures for finding content, passages that are discursive and "rambling"--that don't come quickly to the point--do not pass review. The requirements of density and tension are somewhat more relaxed for the new SAT, because of the extended length and because of a program mandate that passages adhere as closely as possible to the original source, with a minimum of editing.

Other review requirements for passages include the following: Most obviously, the text should be well written; it should represent a unified whole with a beginning, a middle, and an end (Rule 135). The presentation of ideas should be essentially linear, with sufficient transitions to make it easy for the reader to track the development of the ideas. There should be a central focus, with elaboration in the manner described above--not redundancy or varied ways of making a single point. The text should be self-explanatory; it should not require specialized knowledge.

Impact of the specifications. The test developers place most of the focus on securing a good passage, as described above, and are confident that such a passage will support the items they need to meet the specifications. They rarely turn to the specifications once they have a passage and say to themselves, "I need to write a literal question," then look at the passage to see what is possible. Rather, the majority of the items are formed based on the particular content of the passage that suggests items. A test developer will know while polishing a passage that certain parts will lend themselves to specific kinds of items, e.g., analytical. The passage content drives the choice of items at the outset. The specifications are most likely consulted during development of the final items, when the test developer must ensure that all of the required types of items have been asked. At this stage, when items have already been developed to cover most of the passage content, application items can usually be developed.

Once a set of items has been created, test developers vary widely in their perceptions of whether or not that set provides "a thorough analysis and evaluation of the substance of the passage." The only generalization that appears safe to make is that the test developers for the new SAT are confident that it fulfills this requirement of thoroughness better than the old test did. The more advanced the test and the more items needed, the greater is the need for fine distinctions among the items. The attempt of test developers to develop items that assess examinee's abilities to make fine distinctions in their semantic interpretations may sometimes mean that certain features of the passage are ignored. Test developers call this a failure to "use up" passage content.

Analogies

Individual test developers' procedures. As with selection of passage content, test developers begin by looking for words and relationships that are interesting to them. They have individual strategies for accomplishing this, including free association, perusal of a word list, thumbing through a dictionary, and skimming a specialized text or a magazine. One test developer interviewed keeps a notebook in which she writes words that suggest to her some relationships she might want to explore. In looking for appropriate words, the test developers reflect more on the difficulty of the words than on specific content requirements of the specifications (Rule 29).

After a relationship has been established, the first judgment the test developers make regarding its fit is intuitive. Then they check the dictionary meanings of the individual words to reevaluate the accuracy of the relationship (Rule 35). This initial process centering on the dictionary definition leads them at times to abandon the original relationship and establish another. In no case is the approach formulaic. Most often, the rationale is written after the stem is written, though one test developer interviewed sometimes does not write the rationale until after the key and a full set of distracters have been written. In general, however, once the stem is written, the rationale drives the rest of the development. Stems and keys are checked for overlap with analogies that have appeared in previous test forms. According to the test developers, few items are lost because of overlap.

Content constraints considered in the review process. The constraints mentioned here are those that relate not to the technical merits of the analogy, but to the content or the type of analogical relationship or the requirements of the pool. Certain kinds of content will be rejected in review because of the possibility that it might be unsettling for examinees; the specific topics may or may not be features of the sensitivity guidelines. Examples include those topics described earlier as being related to DIF (see Rule 186). Concrete analogies--those whose parts can be perceived though the senses, that is, seen, felt, or heard--are more likely than abstract or mixed analogies to violate either corporate or program sensitivity guidelines because they are more likely to be differentially familiar to some population groups. Archaic, obscure, or highly specialized words are also avoided. Humorous words and relationships are not acceptable (Rule 29). Relationships that are defensible from a lay perspective but not from the perspective of science are rejected. Finally, highly unusual and idiosyncratic or "weird" relationships are rejected. A relationship may have survived the review process to the point of the coordinator's review or the planograph review, but if that reviewer or any other just does not "get" the relationship, perhaps because of its difficulty or obscurity, the item will die.

Technical constraints considered in the review process. From a technical standpoint--how well formed the item is--desirable characteristics for analogies focus primarily on the tightness of the fit between the stem and key (Rule 45). Tightness or necessity of fit is a judgment that the relationship between a pair of elements is "what a lot of people will agree is analogous," without further explanation or justification. This concern for reliability of the examinee's interpretive response receives the bulk of review attention. False, loose, and strained relationships between the stem and the key and the stem and distracters are eliminated by reviewers. Related to the strength of the relationships are the suitability and applicability of the rationale. The rationale must indisputably describe the relationship between the stem and key (Rule 47), and no other rationale must exist that would justify any of the distracters as keys.

Closely related to the above is the requirement of parallelism. The parts of speech of the words in the pairs represented in the options must be parallel to the parts of speech in the stem, except where the stem is a pair of nouns or verbs and the options are attributive adjectives or adverbs (see Rule 37). Also, the distracters must come from the same "realm" as the stem and key--e.g., the distracters for a science stem and key must also have a science origin (Rules 31 and 37). In addition, all of the elements of the item--distracters and stem/key pair--must have tight or necessary relationships.

Given satisfaction of all of the considerations described above, individual test developers also obey a body of rules for writing and reviewing with different degrees of rigor. Examples of these rules include the general avoidance of opposites and synonyms (Rule 52), although unusual antonym relationships are acceptable in the new SAT, which no longer has an antonym type of item; avoidance of items that have complicated rationales; avoidance of "keying down"--items in which one word in the stem and key are synonymous. One test developer stated a rule that requires one distracter to be "almost" a synonym of one of the words in the stem. According to the test developers, these rules, applied within the review process, may or may not be written in a test developers' manual at any given time; and the strictness of the application may depend on the individual reviewer, especially regarding what is or is not acceptable for distracters. Rules that are persistent in their visibility within the review process will likely become incorporated into memoranda and later into a manual. An example of the latter in the SAT program is the history of the origin and acceptance of rules intended to lessen the extent to which analogy items are coachable.

Test developers believe that vocabulary mainly determines the level of difficulty of an analogy although an item may have easy vocabulary and a difficult relationship. The test developers who had a sense of the types of analogies examinees find most difficult stated that, to their knowledge, abstract analogies produced the highest degree of difficulty. Not every test developer interviewed had a sense of what type of analogy proved most difficult for examinees. All were fairly specific as to the ease or difficulty of generating ideas for the various types of analogies, a factor that was not necessarily significant in the context of the review process. Abstract items were identified as the easiest to write, mixed next, and concrete hardest. According to the test developers, concrete items--though most troublesome to write because of the difficulty inherent in finding four "things" that relate analogously--have the tightest or most necessary relationships. Coming up with ideas for an abstract analogy may be easiest, but the concrete relationships produce tighter items. The responses were mixed as to which of these types of items fair best in the review process.

A factor that affects the ease of generation of analogies is the relative size of the universe--at least in the minds of the test developers--of content from which the analogies are drawn. Science analogies must be general and not specialized, a requirement that dictates a small arena from which science analogies can be drawn. To a slightly greater or lesser extent, depending on the judgment of the individual test developer, the same is true of the humanities. Social science and human relationships appear to have the broadest available universes from which analogies can be drawn. The test developers indicated that science and humanities analogies are least likely to survive the review process, for the reason stated above: their (apparent) shrinking universes for non-specialized relationships and the fact that science analogies must be scientifically accurate. The obligatory review of science analogies by a science specialist causes the loss of items that may have seemed defensible prior to the reviews.

The likelihood that items will be lost after the first review depends on whether the first reviewer used by the particular program is the more or less experienced reviewer. SAT first reviewers are often the less experienced reviewers, and the opposite is true for the GRE. SAT test developers stated that they sometimes lost items at second review, whereas those GRE test developers interviewed indicated that they never lost items at the second review. Defensibility becomes a progressively strong factor in determining whether an item will end up in a final form as the item moves along the review continuum. The movement along the continuum is from subjectivity--a requirement that the content and the relationship be of interest to the writer--to objectivity--a requirement that the item content and relationship be defensible under legal scrutiny.

Sentence Completions

Item development. In the first stage of writing a sentence completion item, test developers flip through the pages of a magazine, newspaper, or book that has subject content the item should represent (rule 71). They look for a sentence that contains a subordinate clause and has an element of "tension"--a negation in one half of the sentence, an if-then formulation, or another type of contrastive quality. In some cases, they will find two sentences that together establish the desired tension and will combine the two into one complex sentence.

The one test developer interviewed who writes the new type of SAT sentence completion items that assesses vocabulary often consults word lists to generate ideas for that particular type of item. Except for the one-blank vocabulary type of sentence completion item, test developers begin not with a word or words, but with an appropriate sentence--one that has the necessary strength of tension and that plausibly presents the subject content: science and nature, arts and humanities, social studies and practical or everyday life, human relations and feelings (Rule 82). Their first decisions are, in this way, semantic. They decide which words to leave blank after editing the sentence to their satisfaction, although during the process of polishing the sentence they may be aware of a word or words that are potential candidates for blanking (Rules 92, 94, 95, and 96). They might, for example, spontaneously generate distracters for one or two words in the sentence they are shaping. Having focused certainly on a word, the test developer revises the sentence to ensure that the sentence logic points directly and unambiguously to the concept that the word represents--to ensure that the sentence meaning unequivocally suggests the particular relationship that the blanked words are intended to make explicit. When writing distracters, test developers may consult a thesaurus or depend on their own associational abilities.

Sources. Since sentence completions are fewer than 35 words (Rule 75), the content must have a tightly focused logic if it is to support the conditions that make blanking successful. The sentence must have a clear meaning that is apparent before the blanked words are added. At the same time, the sentence must represent the particular content requirement of the specifications that, as an item, it is intended to meet. Because of these constraints, test developers tend to use specialized books and magazines as sources for sentence completions. Or, depending on the content required for the particular item, they might peruse the better written popular magazines or make sentences up.

The rule governing the number of words permitted for sentence completion items is a good example of how a seemingly structural rule that involves little judgment on the part of the test developer is really a functional rule at base. The number of words require that the logic of the sentence is tightly focused, thereby decreasing the likelihood that the examinee will be misdirected by extraneous information. Moreover, such rules as these also illustrate the evolution of the rule system. Judgments about the tightness of the sentence are at first implicitly made by the individual test developer and then are guided by a rule system based on common experience of test development professionals.

Constraints considered in the review process. Test developers avoid content that violates corporate sensitivity guidelines or violates good taste in writing sentence completions. Other content restrictions include easily dated content, cliches, accepted opinions, and content that makes reference to living figures (Rule 87). The content of the items must be plausible, serious--even "textbooky"--but must not be so technical or specialized as to require outside knowledge for keying (Rule 90).

Except for content, the most notable reason a sentence completion fails to pass first and second reviews is that the blanks are not well defined; the content of the sentence, its logic, does not precisely demand the intended key (Rule 104). This failure may mean that there is no key or that one or more distracters is as keyable as the intended key. Other major reasons include unclear or poorly written sentences, faulty sentence logic, and obviousness (only one word will fit the blank; there are no plausible distracters). The key must not repeat any word in the sentence (Rule 98), the examinee should have to read the entire sentence to key the item, and all of the options must have the appropriate syntactic fit within the sentence. In addition to fitting in terms of syntax, the distracters should be as plausible as possible without being keyable (Rule 97). Distracters have to be revised when reviewers can key them by making atypical interpretations of the meaning of the sentence. Clearly, the failure to pass reviews is most directly related to lack of precision in the possible interpretations examinees can make of the sentences.

The key words that are blanked in sentence completion items are usually nouns and verbs, while items in which adjectives, adverbs, and conjunctions are blanked would not pass review (Rule 96). In the formation of sentence completion items, it is the semantic relations among the words, rather than the words themselves or the structure of the sentences that in every step of the development process determines the final item.

Rules specifically related to distracters include the following: distracter pairs should not include a synonym and an antonym (Rule 102); distracters should not all be positive or negative; distracters should not be opposites of the key or of one another. If one of the blanks in a two-blank sentence completion does not lend itself to plausible distracters, reviewers will suggest that the sentence completion become a one-blank type (Rule 101). As with analogy items, the rules for sentence completion distracters are applied more with the intent of eliciting a reliable interpretive response from the examinees than with the intent of reproducing a particular physical form of the item.

Other Issues

The words "concrete" and "abstract" are used differently with regard to different item types to categorize text and the perceived demands of text as well as to describe types of content. How particular subject content fares in the review process has more to do with its rhetorical characteristics--e.g., whether or not there is argumentation or some other form of tension--than with whether or not the content derives from humanities, science, social sciences, or practical affairs. In general, concrete content is considered to be easier for examinees and tighter--less open to interpretation--in its form and substance than abstract content, but to sample a smaller, and perhaps shrinking, universe of appropriate material for testing purposes. Abstract content in general is considered to be harder for examinees but often easier to write and more abundantly available as source material. It is apparent that these considerations are evolving and have not yet been formalized as rules.

Test developers mentioned the science review sufficiently often to indicate the importance of this review in the survival of passages and items. Because test developers working on verbal items are likely to have a humanities background, their orientation may affect the particular type of science passage that is selected. One program, at the suggestion of test developers who are science specialists, monitors itself to ensure that the humanities passages are not more specialized than the science. A future study might examine what impact there is, if any, on development of passages and items of systematically assigning test developers of varying backgrounds to determine the content of verbal materials. In addition, several test developers indicated that they select science materials based on whether or not the content can be visualized. How widely imaging is as a cognitive mediator of the examinee's response may be worth exploring.

A third issue concerns the relative importance of style and logic in review decision making. Every test developer interviewed for this project stressed that logic should be the only determinant of keyability, but style may be an issue of at least minor currency in the development process (see Rules 96 and 113, for example).

Importance of Rules Mentioned in Interviews

Table 9 shows that the rules that were explicitly mentioned in interviews had higher mean ratings than did those that were not explicitly mentioned. It is clear that in the course of item development, rules are evolving both for the institutional process itself and for the individual test developer. As the use of the rule system increases, both for the individual and across individuals, the rules become formalized as memoranda or as different sources of guidelines. This formalization provides a starting point for new test developers, who gradually and inevitably internalize the formal rules as they continue to evolve idiosyncratic working procedures that might then gain in popularity and ultimately become codified.

Table 10

Importance Ratings of Rules
Mentioned in Interviews,
by Content Area

	Mentioned			Not Mentioned		
	Mean	N	S.D.	Mean	N	S.D.
Analogies	3.17	7	1.18	3.06	35	1.24
Psychomet.	2.40	1	1.67	1.38	8	1.61
Read. Comp.	3.34	7	1.08	2.94	48	1.31
Sent. Comp.	3.32	13	1.03	2.85	33	1.36

Discussion

Definition of the Job

One of the major distinctions among job analysis methodologies is whether the focus is on how the work is accomplished or what the work produces (Fine, 1986). This project incorporated both of these approaches because the product of the domain is so large and dynamic that one cannot begin to define it without looking at how its components are selected. In a sense, the work of the test developer is to delimit the verbal domain of the tests rather than to create the verbal domain, because the verbal domain already exists as the infinite relationships of meanings in human language.

One can view this process as Michelangelo viewed his art, as of freeing the form from the stone which surrounds it. In this case, the freeing is accomplished through the application of certain codified procedures.

The sequence of rule formation seems to have two parallel and cyclical components, personal use and institutional use. Individual test developers first learn a formal rule system, propagated by memorandum and by mentoring practices, in how to delimit the verbal domain. In their professional activities, they come to find and internalize effective practices for developing test items, even when they cannot formalize these practices as rules. That is, it is often much more difficult for test developers to describe what they do than to demonstrate what they do. With greater experience, they can codify these practices as rules to be shared with others. When these rules are shared, they become institutionally codified in publications and memoranda. The publications then become training guidelines for new test developers and the references for all test developers, thus starting the cycle again. In this way, the individuals and the institution have parallel sequences of rule formations, going from formative practices to formalized systems. This evolution explains why many of the test development rules are clerical and, on the surface, seem to focus on structural considerations, e.g., use a vertical format for analogy item options whenever possible (Rule 182). These clerical tasks may often be supported by research (see Carlton & Harris, 1989 for this rule), and formalize the practices that individual test developers use to assure that the item presents the same interpretation problem to every examinee.

This analysis also illustrates how formation of the verbal domain for testing is parallel to formation of language, both for the individual and for the language itself. In both cases individuals select from infinite possibilities structures to represent semantic relations. This selection is governed by rules of grammar. In a very real sense, the rules governing the formation of the verbal domain for aptitude tests are the grammar of the domain.

Evolving Rules

Perhaps the best example of evolution of the rule system is in the choice of materials for reading comprehension passages. The material has to be interesting enough to engage the examinee but cannot be controversial. The test developers know the rules of exclusion, e.g., no sports, no war, no rural life, but they also exclude material based on past experiences with what fails to engage the readers or fails to generate items that have adequate psychometric properties. Test developers who are trained in humanities are often concerned that science passages are engaging for science majors and are accurate in the information they communicate. The current informal practice to address these concerns is to have the science passages evaluated by science specialists. It may well be that this practice will ultimately evolve into a formal system for checking science passages.

In an earlier stage of development are implicit rule systems. For example, one can see how the issue of engagement and differential experience of examinees with the reading material may be bound to social class. What is engaging for the well educated test developer may not be engaging for the high school junior or senior with a different background. This is recognized in the formalized systems having to do with gender and ethnic representation, and is in the early formative stages for such concerns as region (much attention has been given to national differences, for example) and social class. Thus, the ETS sensitivity guidelines caution against elitism, ethnocentrism, and inappropriate underlying assumptions at a formal level, prescribing that certain material be excluded, e.g., speaking about polo or junk bonds. As the rule system continues to evolve, it will continue to exclude elitist material, and it will probably begin to set standards for including material that is familiar and engaging to the working class student, for example. In functional terms, some rules have been codified that limit irrelevant affective responses by examinees based on class issues, but there are not as yet any rules to govern the stimulus properties of material to elicit interest and engagement from readers from different social classes.

This seems to be a discernible pattern of rule evolution, as well, that material that is clearly not accessible to large segments of the population is excluded from the verbal domain because it does not allow some examinees to fairly evaluate the semantic relations. Examinees may understand the vocabulary and structure of the material of a reading passage, for example, but might not find the material familiar enough to demonstrate the high level of interpretive skills they possess.

The next step in this evolution is to identify contexts that are more accessible to the disenfranchised group, or more universally accessible to all examinees, and to have inclusion rules for this material. This next step may be observed in the formal rules now governing inclusion of passages about the accomplishments of women in different fields, or the inclusion of material that focuses on humanities or on science or on areas that will engage some definable segment of the population. Another, more subtle characteristic of this second step in the evolution of rules is that exclusion rules become more focused as technology such as the Mantel-Haenszel procedure allows test developers to identify with greater accuracy the kinds of materials that are differentially accessible or are offensive to some examinees (A. Simpson, personal communication), and exclusion rules that were explicit in the past (such as those listed in the first version of the Sensitivity Guidelines) become internalized.

Clearly, test developers cannot categorize all examinees in terms of their most important demographic characteristics, but once test developers identify material that is differentially interesting or familiar to different segments, the evolution of rules begins. This evolution begins by identifying material to be excluded and making some general recommendations for including material, and it advances (thus far) in recommending that material be included that experience and research has identified as being familiar and fair to examinees.

How this material is included is a matter of concern for cognitive functioning, as described earlier. Material may be sampled that is universally accessible, or material that is differentially accessible may be balanced either within or across tests. Mantel-Haenszel analyses have provided great potential for identifying types of contexts that are differentially familiar to certain segments of the examinee population.

Summary and Conclusions

The functional orientation and job analysis methodology used in this study have been very successful and productive. We have identified a formal rule system for delimiting the verbal domain from the infinite variety of relationships among words. Along with the rules, we have estimated the importance of their use and have described the activities for which test developers find them most useful.

Future Inquiry

A number of future investigations are recommended:

1. The study should be expanded to examine the rules governing sampling the verbal domain for constructed response items, which are now in the process of formalization;

2. A longitudinal investigation should be made of rules that have not yet been formally codified, to better define the evolutionary mechanism.

3. The role of outside expert committees has only been tangentially mentioned as part of the formation of test specifications. More research needs to address other roles that outside experts can assume in helping to evaluate items and define the content domain.

4. Further functional analyses of item stimulus characteristics and examinee response characteristics may prove most fruitful for improving measurement. Tools such as Mantel-Haenszel analyses are available to make such analyses possible.

References

Anderson, R. C., & Davison, A. (1988). Conceptual and empirical bases of readability formulas. In A. Davison & G. M. Green (Eds.) Linguistic complexity and text comprehension: Readability issues reconsidered. Hillsdale, NJ: Erlbaum.

Anderson, R. C., Reynolds, R. E. Shallert, D. L. & Goetz, E. T. (1977). Frameworks for comprehending discourse. American Educational Research Journal. 4 (pp. 361-381).

Angoff, W. H., & Dyer, H. S. (1971). The Admissions Testing Program. In W. H. Angoff (Ed.) The College Board Admissions Testing Program: A technical report on research and development activities relating to the scholastic aptitude test and achievement tests. New York: The College Board (pp. 1-13).

Carlton, S. T. (1983). The GRE verbal scale. New York: The College Board, unpublished manuscript.

Carlton, S. T., & Harris, A. M. (1989). Female-male performance differences on SAT: Causes and correlates. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

DeMauro, G. E. (1990, April). Effects of representation of gender groups in the examinee population on the Mantel-Haenszel procedure. Paper presented as part of a symposium at the annual meeting of the American Educational Research Association. Boston (April).

DeMauro, G. E. (in progress). Specifying the content domain of aptitude tests. Princeton, NJ: Educational Testing Service.

Donlon, T. F. & Angoff, W. F. (1971). The Scholastic Aptitude Test. In W. H. Angoff (Ed.), The College Board Admissions Testing Program: A technical report on research and development activities relating to the scholastic aptitude test and achievement tests. New York:
The College Board (pp. 15-47)

Educational Testing Service (1992). Sensitivity Review Process: Guidelines and Procedures. Princeton, NJ: Author.

Educational Testing Service Manual for item writers. Unpublished manuscript.

References (cont'd).

Emmerich, W. (1991). Vocabulary and Reading Topics in Verbal Admissions Tests: PRPC Work Plan. Princeton, NJ: unpublished manuscript.

Emmerich, W., Enright, M., Rock, D., & Tucker, C. (1991). The development, investigation, and evaluation of new item types for the GRE analytical measure. (GRE No. 87-09P, ETS RR-91-16). Princeton, NJ: Educational Testing Service.

Fine, S. A. (1986). Job analysis. In R.A. Berk, (Ed.). Performance assessment: Methods and applications. Baltimore, MD: The Johns Hopkins Press (pp. 53-81).

Gleitman, H. (1991). Psychology (Third Edition). New York; W. W. Norton & Co.

Graduate Record Examinations (1988). 1988-89 GRE information bulletin. Princeton, NJ: Educational Testing Service.

Haladyna, T. M. & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. Applied Measurement in Education 2(1), pp. 37-50.

Hecht, L. W. & Schrader, W. B. (1986). Graduate Management Admission Test: Technical report on test development and score interpretation for GMAT users. Princeton, NJ: Graduate Management Admission Council and Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the Mantel-Haenszel procedure. (PSRTR-86-69, ETS RR-86-31). Princeton, NJ: Educational Testing Service.

Hunt, E., Lunneborg, C. & Lewis, J. (1975). What does it mean to be high verbal? Cognitive Psychology, 7, pp. 194-227.

Hunter, R. V. & Slaughter, C. D. (1980). The Educational Testing Service sensitivity review process. Princeton, NJ: Educational Testing Service.

Kingston, N. M., Schneider, L. M., & Briel, J. B. (1988). Using empirical Bayes methods to investigate the potential sex and ethnic bias of the GRE General Test. Paper presented at the annual conference of the American Psychological Association. Washington, D. C.

Myers, J. L. (1972). Fundamentals of experimental design. Boston, MA: Allyn and Bacon.

References (cont'd.)

St?ffensen, M. S., Joag-dev, C., & Anderson, R. C. (1979). A cross-cultural perspective on reading comprehension. Reading Research Quarterly, 15, (pp. 10-29).

Sternberg, R. J. (1979). The construct validity of aptitude tests: An information-processing assessment. In Construct Validity in Psychological Measurement: Proceedings of Colloquium on Theory and Application in Education and Employment. U.S. Office ·f Personnel Management, Educational Testing Service, pp. 67-75.

Stricker, L. J. & Rock, D. A. (1985). Factor structure of the GRE General Test for older examinees: Implications for construct validity. (RR 85-9, GREB No. 83-10R). Princeton, NJ: Graduate Record Examinations Board, Educational Testing Service.

Wild, C., McPeek, W. M., & Koffler, S. (1988). Co::current validity of verbal item types for ethnic and gender subgroups. (GREB No. 84-10). Princeton, NJ: Graduate Record Examinations Board, Educational Testing Service.

Appendix A

Survey Instrument of Use of Test Development Rules (First page of instrument)

Group:____________________ Title:____________________

Verbal Tests:______________________________________

A list of explicit rules for forming test specifications, and choosing questions and materials for verbal measures follows. Please rate the importance of each of the rules for defining the domain of verbal tests. Use the following scale: 4 (critical), 3 (very important), 2 (important), 1 (of little importance), or 0 (not important). After you make the 0-4 rating for each rule, kindly consult the decision point list and write the two decision points for which the use of the rule most important, e.g. "IC" for Refining Test Specifications and "IF" for Developing Test item Context. Only include decision points that are rated as important (2) or higher. There may be rules for which you will not write any decision points or rules for which you will write only one decision point. In each section, please feel free to add any other implicit or explicit rules that you use in test development that are not listed.

	Rating	0-2 Uses

A. Antonyms

1. General

   a. Inclusion rules

      1. Use contrariety in antonym questions only when a strong defensible key is present. _____ _____ _____

      2. Use extreme positions for polar contraries. _____ _____ _____

Appendix B

Mean Importance Ratings and Numbers of Respondents (n=5) Who Rated Each Rule as At Least Important (2) (Underlined values failed to meet the criterion)

A. Antonyms	Mean Rating	No. Rating 2 or More
1. General		
a. Inclusion rules		
1. Use contrariety in antonym questions only when a strong defensible key is present.	1.6	2/5
2. Use extreme positions for polar contraries.	0.8	2/5
b. Contrariety exclusion rules		
3. Avoid contrariety as a stem-key pair.	1.0	2/5
4. Avoid contrariety as distracters unless a much stronger stem-key pair is present.	1.6	2/5
c. Polar contrary exclusion rules		
5. Do not put distracters for polar contraries on the same continuum the stem-key opposition.	1.6	2/5
6. Avoid using extreme:midpoint stem-key pairs, e.g., right:middle, as opposed to right:left.	1.6	2/5

A. Antonyms	Mean Rating	No. Rating 2 or More
d. Converse relationship exclusion rules		
7. Avoid stems and keys that are not in opposition of direction, e.g.,husband:wife.	1.6	2/5
8. Avoid weak converse relationships as stem-distracter pairs unless a much stronger opposition relationship exists for the stem and key.	2.0	3/5
2. General rules for writing antonyms		
a. Inclusion rules		
9. Consider the connotative and denotative meanings of words.	4.0	5/5
10. Use familiar words.	2.4	4/5
11. Substitute the key in sentences using the stem, to determine if the sentence expresses the opposite meaning.	1.6	4/5
12. If words and phrases are in combination in an item, two must be of one type and three must be of the other type (words or phrases).	2.4	4/5
13. Phrases as options may have 2-3 words, either adjectival, adjective modifying noun, adverb modifying adjective, verbal, or prepositional.	1.2	2/5
14. Use parallel parts of speech as the stem and options in single-word antonyms.	3.2	4/5

A. Antonyms	Mean Rating	No. Rating 2 or More
15. When the stem may be different parts of speech, make the first option unambiguously one part of speech to establish the part of speech intended for the stem and all the options.	3.6	5/5
16. Antonyms may test thc ability to define words or make fine distinctions among similar distracters.	3.8	5/5
b. Exclusion rules		
17. Avoid specific determiners, or a key which is different in some dimensions from other options.	3.8	5/5
18. Avoid synonyms as distracters.	3.4	5/5
19. Avoid antonyms that require specialized knowledge.	3.4	5/5
20. Check GRE antonym stems and keys for overlap in GRE antonym file.	3.2	4/5
3. General rules for reviewing antonyms		
a. Inclusion rules		
21. Consider the stem in several contexts.	3.4	5/5
22. Key the item considering each option.	4.0	5/5
23. Check the exactness of the key against the dictionary.	3.4	5/5

A. Antonyms		Mean Rating	No. Rating 2 or More
24.	Examine usage and dictionary cross references.	3.4	5/5
25.	Suggest improvements for the key and distracters.	3.6	5/5
b.	Exclusion rules		
26.	Examine possible obscure uses of distracters that may make them keyable.	3.8	5/5
27.	Eliminate tricky, frivolous, or implausible distracters.	3.4	5/5
28.	Eliminate distracters that do not conform to the general rules for writing antonyms.	4.0	5/5
Added Antonym rules:			
a.	Avoid elitist, religious, or insensitive words in antonyms.	3.0	2/2
b.	Eliminate all distracters that are keyable.	4.0	1/1
c.	Avoid overused words.	4.0	1/1

B. Analogies	Mean Rating	No. Rating 2 or More
1. Rules for writing analogies		
a. Inclusion rules		
29. Use vocabulary that is familiar to the intended examinees.	2.8	5/5
30. Base the analogy on relationships that are familiar to the examinee.	2.2	2/5
31. Make the relationship between the first and second word of the stem the same as the the relationship between the first and second word of the key.	4.0	5/5
32. Use only analogies with concise rationales.	2.8	5/5
33. Allow stem and key pairs to be from different realms.	1.8	4/5
34. Check all analogies with a dictionary, making sure that the key is the best answer.	3.0	4/5
35. Check the definitions of the words in the stem and the key in at least one dictionary.	2.8	4/5
36. Assure that the relationships given in at least two distracters are strong enough to stand as stems in other questions.	2.6	4/5
37. Assure that parts of speech in the options are parallel to parts of speech in the stem, except where the stem is a pair of nouns or verbs and the options are attributive adjectives or adverbs.	3.0	4/5

B. Analogies	Mean Rating	No. Rating 2 or More
38. Make the words of the first option unambiguous in terms of part of speech.	3.0	4/5
39. State the rationale succinctly on the back of the item sheet.	4.0	5/5
40. Base the key on the relationship alone and not the subject area of the stem.	3.0	4/5
41. Use single word analogies whenever possible.	2.4	5/5
42. Assure that any mix of content areas in the options is appropriate.	3.0	5/5
43. Assure that at least one distracter has a negative rationale if the stem and key has a negative rationale.	3.6	5/5
44. Assure that every option has an immediately apparent relationship.	3.6	5/5
45. Assure that the logical fit between the stem and the key is "tight."	3.8	5/5
46. Assure that option A unambiguously establishes the part of speech if it is not clearly established in the stem.	3.4	5/5
47. Assure that the stem clearly establishes the intended rationale.	3.4	5/5

b. Exclusion rules

48. Avoid analogies which incorporate cliches.	3.4	5/5

B. Analogies		Mean Rating	No. Rating 2 or More
49.	Avoid analogies in which the key or stem omits an intermediate step expressed in the other.	3.0	4/5
50.	Check all of the words in the item against the list of overused words and replace all overused words.	1.4	2/5
51.	Check the stem and the key against the word overlap data base and initial the item sheet when the check is completed.	4.0	5/5
52.	Avoid analogies in which the relationship depends entirely on synonyms or antonyms.	2.4	4/5
53.	Do not use a reverse key or other trick distracters.	3.8	5/5
54.	Avoid proper names and brand names.	4.0	5/5
55.	Check GRE analogies for overlap in GRE history file.	3.2	4/5
56.	Replace distracters that fit alternate unintended rationales for the stem.	3.8	5/5
57.	Revise options that have merely associational relationships.	3.0	5/5
58.	Check all the options for inappropriately overlapping rationales.	3.0	4/5

B. Analogies

2. General rules for reviewing analogies

a. Inclusion rules

		Mean Rating	No. Rating 2 or More
59.	Check the relationship in the stem.	4.0	5/5
60.	Check each option to determine whether it fits the relationship.	3.8	5/5
61.	Check rationales for the key and for each option.	3.4	5/5
62.	Suggest ways to improve the key and the distracters.	3.6	5/5
63.	Assure that all options have the same parts of speech.	3.6	5/5
64.	Assure that the part of speech used in option A is unambiguous.	3.2	5/5

b. Exclusion rules

		Mean Rating	No. Rating 2 or More
65.	Eliminate possible obscure uses of distracters that make them keyable.	3.6	5/5
66.	Eliminate options that can be keyed under a second rationale.	2.8	4/5
67.	Eliminate tricky, frivolous or implausible distracters.	3.4	5/5
68.	Eliminate options that do not have strong relationships.	2.6	5/5
69.	Eliminate options that rely on identical relationships but have unintended second level factors.	0.4	1/5

B. Analogies	Mean Rating	No. Rating 2 or More
70. Eliminate options that are synonym or antonym pairs.	1.6	2/5
Added Analogy rules:		
a. Avoid analogies that depend on certain attitudes, e.g., political attitudes.	4.0	1/1

C. Sentence Completions	Mean Rating	No. Rating 2 or More
1. Rules for writing sentence completions		
a. Inclusion rules		
71. Select either original or published sentence.	3.8	5/5
72. Select sentences of standard and grammatically correct style and vocabulary.	3.0	4/5
73. Select sentences having independent parts.	0.4	1/5
74. Select sentences with meanings that are self-contained.	2.4	3/5
75. Use sentences that are no more than 35 words in length.	2.8	4/5
76. Use single word options except when including articles or prepositions enables writing plausible distracters.	3.0	4/5
77. Use options that are parallel in use of acceptable English.	3.4	5/5
78. Keep gender and racial references in balance.	2.8	4/5
79. Assure that the sentence meets length requirements.	3.4	5/5
80. Assure that the sentence addresses a topic that is appropriate for the examinee population.	3.8	5/5
81. Assure that the sentence conveys a thought as succinctly and clearly as possible.	3.6	5/5

C. Sentence Completions

		Mean Rating	No. Rating 2 or More
82.	Assure that the sentence incorporates all necessary contextual information.	3.6	5/5
83.	Note all sources for unfamiliar topics on the back of the item sheet.	1.4	3/5
84.	Assure that the options are parallel in format and in part of speech.	3.8	5/5
85.	Assure that race and gender codes have been completed for every item.	2.4	5/5
b.	Exclusion rules		
86.	Exclude sentences that use colloquial expressions, contractions, and/or slang.	2.6	5/5
87.	Avoid cliches, foreign, or familiar sentences.	3.2	5/5
88.	Avoid metaphorical sentences.	2.2	4/5
89.	Avoid sentences in which vocabulary is the only skill tested.	1.4	2/5
90.	Avoid sentences which require specialized knowledge outside of the sentence for completion.	4.0	5/5
91.	Avoid sentences which require subtleties of formal English usage for completion.	3.4	5/5
92.	Avoid using as blanks words on which the meaning of the sentence depends.	1.8	3/5

C. Sentence Completions		Mean Rating	No. Rating 2 or More
93.	Avoid writing options and keys that enable the key to be determined by vocabulary level, length, etc.	3.8	5/5
94.	Avoid using as blanks the first words in sentences.	3.2	5/5
95.	Avoid using as blanks words that can only function as prepositions.	3.4	5/5
96.	Avoid using as blanks words that are superfluous to the meaning of the sentence.	3.2	5/5
97.	Avoid identifying the key as being different in any other way from the distracters than in meaning.	3.6	5/5
98.	Do not use words in options that appear in other options or in the stem.	3.8	5/5
99.	Check all options against the list of overused words.	1.4	2/5
100.	Check keys for definitional questions against the word overlap database.	2.6	4/5
101.	Avoid questions with two blanks which can be keyed using only one of the two blanks.	3.6	5/5
102.	Do not use distracters in questions with one blank that are antonyms of the key.	3.0	4/5
103.	Do not write stems of more than 20 words in length for definitional sentence completion questions.	2.2	3/5

C. Sentence Completions	Mean Rating	No. Rating 2 or More
3. Rules for reviewing sentence completion questions		
a. Inclusion rules		
104. Key each item, considering each option.	4.0	5/5
105. Reconcile discrepancies with the official key.	4.0	5/5
106. Compare the options to words that would correctly complete the blanks.	2.2	3/5
b. Exclusion rules		
107. Identify violations of the guidelines for writing sentence completion questions.	3.2	5/5
108. Identify weak or idiomatically misfitting questions.	3.6	5/5
109. Check the key for unusual characteristics.	3.6	5/5
4. Rules for classifying sentence completion questions.		
a. Inclusion rules		
110. Assure that definitional sentence completions depend on linking the definitions of specific words to the sentence.	3.6	5/5
111. Assure that definitional questions enable examinees to have a specific idea of the word needed before reading the options.	3.4	5/5

C. Sentence Completions	Mean Rating	No. Rating 2 or More
112. Assure that distracters of definitional questions draw on vocabulary knowledge rather than on logical characteristics of the stem.	2.4	4/5
113. Assure that keying regular sentence completion depends on analyzing logical relationships within the sentence.	3.6	5/5
114. Assure that the stem of a regular sentence completion item presents a complex and sophisticated thought.	2.6	4/5
115. Assure that the distracters of a regular sentence completion item can be eliminated more on the basis of logical relationships than on vocabulary knowledge.	3.2	5/5
b. Exclusion rules		
116. Exclude definitional sentences longer than 20 words.	2.0	3/5
Added Sentence Completion rules:		
a. Exclude sentences containing thoughts that go too strongly against readers' expectations or common knowledge.	3.0	1/1
b. Avoid sentences that contain libelous assertions about living people, and sentences about political figures, corporate products by brand name, or corporations.	4.0	1/1

D. Reading Comprehension	Mean Rating	No. Rating 2 or More
1. Passage choice rules		
a. General inclusion rules		
117. Set the proper level of sophistication and density for the examinees.	4.0	5/5
118. Choose complex passages combining exposition and argument.	3.0	5/5
119. Choose self-contained passages.	2.4	3/5
120. Assure proper representation of women and minority groups.	4.0	5/5
121. Assure the passage is an appropriate reading task for the educational level of the examinees.	4.0	5/5
b. General exclusion rules		
122. Do not choose passages that require specialized information for interpretation.	4.0	5/5
123. Do not choose passages from current journals or texts.	1.8	3/5
c. Format inclusion rules		
124. Choose passages of about 450 words for long sets, about 150 words for short sets, and passages that fit into desired categories for total number of words.	2.4	3/5
125. Add orientation phrases as appropriate.	2.6	3/5

D. Reading Comprehension	Mean Rating	No. Rating 2 or More
126. Number every fifth line except when it is the last line.	3.0	5/5
127. Assure that the passage represents no more than one-tenth of the original source.	3.2	5/5
128. Assure that passages drawn from anthologies or collections of essays are less than one-tenth of the original individual essay or literary piece.	3.2	5/5
d. Procedural inclusion rules		
129. Note material use in source documents.	3.0	4/5
130. Pre-check sensitivity and subject matter accuracy with specialists.	2.8	4/5
131. Supply relevant information to reviewers about sources that have been used previously.	2.8	4/5
132. Identify and address specialized terms or assumed background knowledge in contextual cues such as footnotes and the introduction.	3.0	4/5
133. Assure that the passage is accessible with engaging features, examples, and breathing space.	3.6	5/5
134. Assure that the introduction to the passage supplies information that is helpful and not merely summarizing.	2.4	4/5

D. Reading Comprehension	Mean Rating	No. Rating 2 or More
135. Assure that the passage functions as a complete unit of thought, with a sense of beginning, middle, and end.	3.8	5/5
136. Choose passages with enough complexity to yield a high proportion of Extended Reasoning questions.	4.0	5/5
137. Attach a rationale for pairing for a pair of passages.	2.2	4/5
138. Have 4 - 8 stems and keys submitted for pairs of passages.	3.2	5/5
139. Assure that the passage is reasonably up to date for rapidly changing fields.	3.2	5/5
140. Attach photocopies of relevant pages from the original source.	3.6	5/5
141. Attach completed officially initialed (approved) copyright information card for reviewers (2 cards for pairs).	3.8	5/5
142. Attach photocopy of the copyright page of the original source (for each passage).	2.8	4/5
e. Procedural exclusion rule		
143. Check passage overlap in author file.	4.0	5/5
144. Check that the author of a passage does not appear on the list of overused writers.	2.2	3/5

D. Reading Comprehension		Mean Rating	No. Rating 2 or More
2. Rules for writing reading comprehension questions			
a. Inclusion rules			
145.	Choose the best item and option format.	4.0	5/5
146.	Design EXCEPT questions to have statements that are all true and stems that include the statement "all of the following."	1.4	2/5
147.	Base item difficulty on understanding the passage, not on understanding the question.	3.6	5/5
148.	Refer to the passage and author properly in the stem.	2.8	4/5
149.	Use specific line references in the stem.	1.4	3/5
150.	Use directed stems.	3.4	5/5
151.	State the stem clearly and concisely.	3.8	5/5
152.	Include in the stem words that must otherwise be repeated in each option.	2.8	5/5
153.	Have 75-80% of the questions meas[illegible] Extended Reasoning skills.	2.0	3/5
154.	Assure that overlap questions are critical and distinct and do not extend the number of questions beyond the minimum.	0.8	2/5
155.	Assure that the questions in a set cover all sections of the passage.	2.0	4/5

D. Reading Comprehension	Mean Rating	No. Rating 2 or More
156. Include at least a few easy questions in a set, especially early on.	2.4	4/5
157. Assure that Literal Comprehension questions cover points that are essential to understanding the passage as a whole.	2.4	4/5
158. Assure that questions testing the main idea and style or tone cover points that are essential to understanding the passage as a whole.	3.4	5/5
b. Exclusion rules		
159. Limit the number of Roman numeral format questions to one or less per set (two for long sets and one for short sets in College Board tests).	3.6	5/5
160. Limit the number of "EXCEPT" questions to one or less per set.	3.6	5/5
161. Limit the number of negative stem questions to one for short sets and two for long sets (College Board tests).	3.6	5/5
162. Avoid revealing test answers in the stem.	3.2	4/5
163. Keep the numbers of Vocabulary in Context and Literal Comprehension questions within guidelines.	3.6	5/5
164. Restrict the number of overlap questions in a set to two.	1.4	2/5
165. Eliminate all unnecessary qualifiers from the stem.	3.2	5/5

D. Reading Comprehension	Mean Rating	No. Rating 2 or More
166. Eliminate from questions any factors that would allow at least one reviewer to answer correctly without reading the passage.	2.2	4/5
3. Rules for writing reading comprehension questions based on paired passages		
a. Inclusion rules		
167. Assure that 25-50% of the questions cover comparisons between the passages.	3.8	5/5
168. Assure that questions that do not cover comparisons between the passages are reasonably divided between the passages.	3.2	5/5
169. Cover significant aspects of the pair of passages in comparison questions.	3.0	5/5
170. Cover distinct, non-overlapping points in comparison questions.	3.2	5/5
b. Exclusion rules		
171. Eliminate factors on the stems of questions in one pair that give away keys to questions on the other passage or on the pair.	3.6	5/5
Added Reading Comprehension rules:		
a. Use only passages approved by the passage review panel.	1.0	1/1

E. Cognitive Function	Mean Rating	No. Rating 2 or More
1. Contextual rules		
a. Inclusion rule for antonyms, analogies, and sentence completions		
172. Include Arts and Humanities, Social Studies and Everyday Life, Science and Nature, and Human Relationships and Feelings contexts for questions.	3.8	5/5
b. Inclusion rules for reading passages		
173. Include culturally diverse reading selections.	3.8	5/5
174. Include passages on biological sciences, humanities, and social studies.	3.8	5/5
175. Include passages that represent women.	3.8	5/5
c. General exclusion rules		
176. Avoid controversial subjects, e.g., religion and theoretical treatment of evolution.	3.4	5/5
177. Avoid subjects that are abundant in the item pool.	3.4	5/5

Added Cognitive Function rules:

F. Psychometric Properties	Mean Rating	No. Rating 2 or More
1. Difficulty and discriminability		
a. Inclusion rule		
178. Focus on the field of accomplishment in questions portraying women, rather than portraying success as gender-based.	1.8	3/5
b. Exclusion rules		
179. Exclude Roman numeral format questions from LSAT.	0.8	1/5
180. Exclude Roman numeral format questions from GRE pretests.	1.0	2/5
181. GRE use of Roman numeral format questions must control overlap of material in I and in other options.	0.6	1/5
2. Differential Item Functioning (DIF)		
a. Inclusion rules		
182. Use a vertical format for analogy item options whenever possible.	2.2	4/5
183. Use external DIF reviewers.	1.4	2/5
b. Exclusion rules		
184. Reduce non-construct related difficult language in questions.	2.0	3/5

F. Psychometric Properties		Mean Rating	No. Rating 2 or More
185.	Reduce speededness.	1.2	2/5
186.	Reduce the use of contexts or settings, e.g.,sports, war, violence, rural life, that may be differentially interesting or familiar but are not related to the construct.	2.4	4/5
Added Psychometric Properties rules:			
a.	Meet the point by point delta (item difficulty) specifications for the test.	4.0	1/1
b.	Meet the point by point biserial (item discrimination) specifications for the test.	4.0	1/1
c.	Perform a passage precheck.	4.0	1/1
d.	Perform an informal sensitivity review.	4.0	1/1