Research Report

What is Quantitative Reasoning? Defining the Construct for Assessment Purposes

Carol Anne Dwyer

Ann Gallagher

Jutta Levin

Mary E. Morley

November 2003

RR-03-30

What is Quantitative Reasoning?
Defining the Construct for Assessment Purposes

Carol Anne Dwyer, Ann Gallagher, Jutta Levin, and Mary E. Morley

Educational Testing Service, Princeton, NJ

November 2003

Abstract

In order to create fair and valid assessments, it is necessary to be clear about what is to be measured and how the resulting data should be interpreted. For a number of historical and practical reasons described in this paper, adequately detailed statements with both a theoretical and empirical base do not currently exist for the construct of quantitative reasoning for use in assessments. There is also no adequate explanation of the important differences between assessments that measure quantitative *reasoning* constructs and those that are intended to measure *achievement* in related mathematical content areas.

The literature in psychology, psychometrics, philosophy, and education, while containing much that is relevant to the construct of quantitative reasoning, unfortunately does not provide materials that can be used in research and development to address such practical issues or to explain the basic nature of quantitative reasoning assessments. This paper briefly discusses the importance and use of constructs and the quantitative reasoning and standards literature. It then presents a statement about the construct of quantitative reasoning for assessment purposes within a construct validity framework that includes both a definition of the construct and threats to valid score interpretation. These threats are based on related but distinguishable constructs and other types of construct-irrelevant variance in the assessments.

Key words: Assessment, quantitative reasoning, constructs, validity, fairness, construct validity theory

Table of Contents

Overview

Purpose

The intention of this paper is to describe quantitative reasoning in a comprehensive way within the framework of construct validity theory and to give several examples of how selected testing programs have operationalized this construct in the assessments they offer. Where the appropriate information is available, we will also compare and contrast the approaches taken by these assessments and give rationales for the differences.

This construct-validity framework requires that the construct that will form the basis of an assessment must be clearly specified, as well as differentiated from similar but distinct constructs and from other nonconstruct influences that may disrupt score interpretation. Thus for assessment purposes, having a clear construct definition is essential to validity and fairness and to the scientific integrity of the inferences drawn from the assessments. It is also a practical aid to decision-making about assessment design, development, and interpretation and provides a common framework for discussion among those involved in the assessments in various roles.

The quantitative reasoning literature is extensive, and interdisciplinary professional consensus on the nature and extent of quantitative reasoning has become very strong. This consensus is captured in detail in standards developed by the National Council of Teachers of Mathematics (NCTM) (National Council of Teachers of Mathematics [NCTM], 2000) and by other related sets of standards for all educational levels that cover both mathematical content and quantitative reasoning. The NCTM quantitative reasoning standards are broader in scope than can be captured in an assessment setting, as they encompass such elements as the creation of new mathematical knowledge.

According to the NCTM, quantitative reasoning is the developed ability to analyze quantitative information and to determine which skills and procedures can be applied to a particular problem to arrive at a solution. Quantitative reasoning, both generally and for assessment purposes, has an essential problem-solving focus. It includes the following six capabilities: reading and understanding information given in various formats; interpreting quantitative information and drawing inferences from it; solving problems using arithmetic, algebraic, geometric, or statistical methods; estimating answers and checking for

reasonableness; communicating quantitative information; and recognizing the limitations of mathematical or statistical methods. A detailed description of the quantitative reasoning problem-solving process is presented.

Quantitative reasoning requires the use of mathematical content for assessment purposes and for problem solving more generally. Quantitative reasoning is, however, fundamentally different, both conceptually and practically, from mathematical content knowledge. It is dependent upon, rather than psychologically prior to, this other form of valuable expertise. Being clear about the differences between quantitative reasoning and mathematical content knowledge relates closely to the fairness and validity of assessments. Thus it is critical to specify for any quantitative reasoning assessment what mathematical content knowledge is required to address the problem-solving tasks represented in the assessment. Four other important potential threats to the validity of assessments of quantitative reasoning discussed in this paper are calculation load, visual and spatial components, speed of responding, and reading load.

To illustrate different approaches to operationalizing the construct of quantitative reasoning, major ETS tests that contain a quantitative reasoning measure were compared and contrasted with respect to their measurement of quantitative reasoning, inclusion of mathematical content, and published statements that define a construct of quantitative reasoning for their assessment purposes. Although the target construct of quantitative reasoning appears to be highly similar among the tests we reviewed, specific variations with respect to test content and format do exist and seem related to such factors as differences in program purposes and populations. We conclude that a more comprehensive construct-centered approach is needed in order to link the assessment's core concept to the actual content of the test and the type of reasoning it assesses.

Quantitative reasoning for assessment purposes is a problem-solving process with identifiable steps, some of which can be measured by standardized tests. Quantitative reasoning is not the same as mathematical content knowledge, but knowledge of mathematical content is necessary (although not sufficient) to solve quantitative reasoning problems.

The authors of this paper aim to create a shared understanding of both the nature of constructs and of quantitative reasoning that will help guide future discussion and decision-making for research, development, and operational purposes. In this paper we address

quantitative reasoning from a theoretical perspective that allows us to link specific tests to the research literature and to other tests that share a common measurement goal. We do not discuss all facets of designing an assessment, as this would necessarily include policy considerations that are beyond the scope of the information provided here.

Nature of the Quantitative Reasoning Construct Description

In describing the construct we attempt to take into account the views of quantitative reasoning articulated by major theorists in mathematics, cognitive psychology, and education. We also consider other documents that incorporate these views into widely recognized and accepted standards for the practice of teaching and learning such as those published by the NCTM (2000). The measurement framework for this description is based on construct validation theory as articulated by Messick (e.g., Messick, 1975, 1980, 1988, 1989, 1994, 1995), and as represented in the most widely accepted standards for educational and psychological testing, the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999).

Messick's framework requires consideration of both the nature of the construct itself and of related but distinguishable constructs. A closely related element of Messick's framework is that it also requires consideration of both the adequacy of the construct representation (how much of what is important has been included in a particular test) and threats to validity introduced by elements of the assessment that are unrelated to the construct (how much of what is irrelevant has been included in a particular test).

Throughout this paper, we adhere to the meaning of "construct" as defined in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999): **"**the concept or the characteristic that a test is designed to measure" (p. 173).

Constructs and Individual Measures

A construct provides the target that a particular assessment or set of assessments is designed to measure; it is a separate entity from the test itself. Messick has always been clear that "The measure is not equated with the construct nor considered to define the construct" (Messick, 1975, p. 955). There are usually many different ways that a given construct could be measured using different methods or content, and the validation process is aimed at

establishing how well a particular test has succeeded in measuring its target construct. Two tests may be equally construct-valid, yet their scores are not necessarily, or even likely to be, interchangeable. Validity is a property of inferences about scores, and scores are a result of interactions between a specific set of test materials and a specific set of knowledge, skills, and abilities that test takers bring to the test. This point is central to distinguishing the modern view of construct validation from earlier operationalist views that viewed a test as an entity whose meaning does not extend beyond its own characteristics.

A clearly specified construct serves as the fundamental basis for judging the quality of an assessment and for the inferences that can legitimately be drawn about test takers from scores on that assessment. Scores must be interpretable in terms of the intended construct. Higher scores should be associated with greater mastery of material related to the construct, not with more of any other kind of material or testing circumstance. Comprehensive articulation of a construct is thus necessary in order to provide a reference point that can be used to address a wide range of important validity and fairness issues about specific assessments. Discussions of fairness and validity will benefit from a clearly articulated construct statement. Answers to some of the following questions will be more readily apparent with the aid of such a document:

- Which aspects of the target construct are addressed in a particular assessment and which are not?
- How should similar but distinct constructs be defined?
- How might any changes in an assessment be expected to impact the breadth of construct coverage and the exclusion of construct-irrelevant factors?
- And, perhaps most importantly, what inferences about an individual can legitimately be drawn from a particular assessment?

Historical Development of Reasoning Theory and Assessments

Before turning to the nature of the quantitative reasoning construct *per se*, we will briefly describe the development of theory and practice in the measurement of reasoning in the 20th century and comment on how this work relates to modern understandings of reasoning and of the meaning and uses of constructs.

The early part of the 20th century was an era of enormous interest in mental phenomena and optimism about the potential of measures of mental ability for improving the human condition. Studies of intelligence and "genius" (for example, Burt, 1922; Galton, 1869/1978; Terman, 1916; Thurstone, 1924/1976) fueled important assessment development efforts as well as psychological theory in a number of areas and were characterized by a clear but often implicit emphasis on the physical rather than on social or psychological characteristics of groups and individuals. Much historical analysis is available to describe these developments, but this literature makes clear that there was little interest in that era in explicating the nature of reasoning or in presenting rationales for the specific content of the assessments that were used in this research. The word "reasoning" was used frequently, but not usually given a specific, technical definition. Similarly, in these early days of assessment, test content seems to have been governed more heavily by the attractiveness of question types for their common-sense, surface relationship to the purposes of the assessment than by more abstract considerations of what question types might best measure an explicitly-defined cognitive skill. In fact, early efforts to measure what we would now call the construct of reasoning were not aimed at establishing a test's validity in the modern sense. Validity, when discussed at all, referred simply to the correlation of a particular instrument with another single measure of attainment, usually classroom grades or criteria of later attainments such as "eminence."

In this early assessment era there were strong links, and often a virtual identity, between conceptions of intelligence and reasoning in the sense that the prevailing assumption was that the variables under study were biological rather than social in nature and thus impervious to efforts to alter an individual or group's *status quo.* These links were always implicit, but often explicit as well.[1]

Question types *per se* were a major force in conceptualizing reasoning. New assessments often borrowed question types from previous assessments, usually based on the question type's statistical characteristics and face validity. Cumulative information about test scores bolstered "good" question types and fostered their repeated use in similar assessments. The same was true of face validity, especially when it was linked to the interests of contemporary psychologists such as identification of those at the extremes of cognitive functioning.

Information on individual and group differences eventually also led to alternative views of human learning in general and of what assessments were measuring in particular. This shift entailed explicit consideration of the idea that what was being measured was learned and not purely innate. One major proponent of biological views of individual and group differences, Carl Campbell Brigham, later recanted his nativist views very publicly and endorsed the idea that group differences in intelligence were susceptible to influences such as language, culture, and prior educational experiences (1923, 1932).

The legacy of earlier psychological theory and its applications to assessment have been hard to dismantle, despite decades of research and strenuous efforts by psychologists, educators, psychometricians, and others to communicate the complex social and biological antecedents of reasoning as we now understand them. These decades of research are now beginning to bear fruit. For example, in recognition of the current consensus that reasoning is a learned, culturally influenced set of knowledge, skills, and abilities, the College Board has formally disavowed the concept of "aptitude" as a descriptor of the SAT and dropped the usage of "Scholastic Aptitude Test" in favor of simply "the SAT" (Braswell, 1991, 1992). This move was largely in response to the recognition that a legacy of outdated psychological and educational thought continued to mislead many about the nature of the reasoning skills that the SAT measures and therefore about the inferences that can validly be drawn from an individual's SAT scores.

The combination of this view of reasoning as a malleable set of skills, coupled with the emergence of modern construct validity theory in the 1950s and its rise to become the consensus view of validity today, has led to a recognition of the importance of specifying constructs of both new and preexisting assessments and of using these constructs as the basis for their design, development, and use. Until the present time, however, validity theory and practice did not fully converge to provide test makers with models for translating this knowledge into actual test materials. Communications about what reasoning assessments measure have thus largely relied on descriptions of the content of the test, most often in the form of the number and types of questions contained in a test or in test specifications that focus on the content of particular types of questions.

In an ideal world, today's reasoning tests would all have been developed from a clear, modern view of validity, beginning with a complete construct definition that makes explicit

such critical issues as how scores are and are not to be interpreted; what the test is intended to measure; and what the test is seeking to avoid measuring. Assessment is, however, a practical and ongoing necessity with or without all of the desired information. Constructs will thus, in many cases, have to be fully specified well after the fact of initial test design. Such efforts can be tremendously useful in evaluating current assessments and in designing improvements for future assessments.

The Development of Quantitative Reasoning Constructs and Assessments

Although the assessment of quantitative reasoning has been a measurement goal from early in the 20th century (e.g., College Entrance Examination Board, 1926; McCall, 1923; Monroe, DeVoss, & Kelly, 1917; Thorndike et al., 1924; Yerkes, 1921), systematic treatment of quantitative reasoning as a cognitive process distinct from mathematics as content or curriculum did not begin to take shape until much later.

During the 1920s and 1930s, the views of Dewey (1933) prevailed in the analysis of quantitative reasoning. Dewey stressed making the development of reasoning ability a fundamental goal of primary and secondary schooling in America. Dewey treated quantitative reasoning as simply one aspect of the reasoning process in education and focused on the implications of reasoning for the teaching and learning process.

E. L. Thorndike was one of the first to tackle the psychology of quantitative reasoning through his research on algebra (Thorndike et al., 1924). Although his frame of reference was strictly stimulus-response, reflecting the prevailing psychological paradigms of his era, he nevertheless offered the first typology of problem types in quantitative reasoning.

Table 1

Thorndike's Typology of Problem Solving

Type I: Problems to answer in which *no* explicit equation or formula is needed or supposed to be used	1. Concerned with knowledge of meaning, e.g., of literal numbers, negative numbers, exponents
	2. Concerned with knowledge of operations
	3. Concerned with combinations of meanings and operation or with other aspects of algebra
Type II: Problems to answer in which an equation or formula is supposed to be used	1. The equation or formula is one of a group of known formulas
	2. The equation or formula is not known but must be constructed by the pupil

Note. Adapted from *The Psychology of Algebra* by E. L. Thorndike et al. (1924). New York: The Macmillan Company.

This typology was used to link question types with learners' cognitive processes for solving them. Thorndike (Thorndike et al., 1924) also explicitly addressed how such processes might be tested by teachers and others. He was unusual in his era in being clear that quantitative reasoning is essentially different from lower-level mathematical skills such as computation:

> …we need to distinguish certain types of work which might with some justification be called applications of algebraic technique to the solution of problems, but which differ notably in the psychological demands they make of a pupil, in the psychological effects they have upon him, and in their uses in the algebra course. All are different from mere computation, or the solution of equations already framed. (p. 132)

Thorndike was also prescient in noting some threats to the validity of measuring quantitative reasoning. He made numerous suggestions to test makers to avoid what we today

would call construct-irrelevant variance. For example, he advised that teachers and other test makers "be careful not to burden pupils unduly with learning physics or astronomy or engineering for the sake of having genuine formulas…[quantitative reasoning problems] should not be complicated by unfamiliar terms, or by the need of difficult inferences to secure consistency in units…" (p. 135).

Thorndike was one of the first to draw attention to the importance of hypothesis formation in problem solving that is central to the NCTM standards today. He clearly felt that he was somewhat alone in this belief, however: "…teachers of mathematics consider the mental process of making a successful hypothesis as a dubious adventure, indecorous if not immoral, and, in dignity, far below proving that a certain hypothesis is correct. Is not the contrary the case?" (Thorndike et al., 1924, p. 439)

In the 1940s, many authors addressed critical thinking, often from a perspective of mathematical thinking, but not with a focus on quantitative reasoning *per se*. A major exception to this trend is Polya (1957, 1962, 1965) who described quantitative reasoning as a problem-solving process and advocated teaching that process explicitly. Polya did not, however, deal directly with the assessment of quantitative reasoning. Presseisen (1986), in a history of critical thinking, attributes the surge of interest in reasoning, including quantitative reasoning, in the 1940s, to a 1938 report by the National Education Association (NEA) called *The Purposes of Education in American Democracy*, which stressed the role of schools in helping students develop skills to resist political propaganda. Polya's central theme of problem solving and how to teach it was vigorously taken up again during the 1960s and 1970s. The earlier work of Glaser, for example, continued this emphasis on quantitative reasoning as a problem-solving process, but also provided a much broader theoretical framework which encompassed a general theory of knowledge and the learning process (e.g., Glaser, 1976). Resnick and Resnick (1977) added another dimension that is represented in the current NCTM worldview, that of the social construction of quantitative reasoning. They argued that there had been a change in attitude toward higher-level cognitive processes that had a social class dimension. That is, quantitative reasoning had traditionally been part of the mathematics curriculum for the privileged, while education for the masses stressed training and practical lower-level skills.[2] Resnick and Resnick say that what is new in modern thinking about quantitative reasoning is not its inclusion in the curriculum, since it has been there for

the schooling of the elite for many years, but rather its inclusion in curriculum intended for "the masses."

In this same era, educationally oriented cognitive psychologists such as Greeno (1978; Greeno & Simon, 1988) were interested in showing that various mathematical task content knowledge and procedures were prerequisites to many classroom activities and tasks, making the point that these are not usually directly taught. Theoretical advances in understanding the cognitive basis of quantitative reasoning emerged (e.g., Newell, 1980; Newell & Simon, 1972; Schoenfeld, 1983, 1985, 2002). Strong efforts were made to translate these insights and empirical data into classroom and assessment practice, with an emphasis on the domain specificity of quantitative reasoning (e.g., Ball, 1991; Glaser, 1984; Glaser & Baxter, 2002).

John Carroll (1993) presented evidence on quantitative reasoning in people from early childhood through adulthood. His main conclusion was that there are only three major reasoning abilities: sequential (deductive), inductive, and quantitative. This conclusion, interestingly, holds for children age five through adulthood. For very young children, reasoning factors cannot be meaningfully separated.

Lohman (in press) shares with Carroll an emphasis on the importance of inferential and deductive processes that he applies to analyses of the nature of quantitative reasoning. According to Lohman, when looking for the uniquely reasoning aspects of quantitative reasoning, one should focus on tasks that tap either deductive or inferential reasoning with quantitative concepts, and he offers useful definitions with some attention to evaluative criteria:

1) *infer[ing]* (either automatically or deliberately) concepts, patterns, or rules that best (i.e., most uniquely) characterize the relationships or patterns they perceive.... Better reasoning is characterized by the use of concepts or rules that simultaneously satisfy the opposing needs for abstraction (or generalization) and specificity. Such concepts or rules tend to be at least moderately abstract yet precisely tuned. Put differently, a poor inference is often vague and captures only a subset of the relationships among the elements in the set. The judgment of what constitutes better reasoning is in part dictated by the shared knowledge and conventions of particular communities of discourse and in part by the precision and generality of the inference.

2) *deduc[ing]* the consequences or implications of a rule, set of premises, or statements using warrants that are rendered plausible by logic or by information that is either given in the problem or assumed to be true within the community of discourse…. Better reasoning involves providing warrants that are more plausible or consistent with the rules of logic or the conditions embodied in a comprehensive mental model. More advanced deductive reasoning involves providing either multiple (possibly divergent) warrants for a single claim or an increasingly sophisticated chain of logically connected and separately warranted assertions.

Thus by the early 1970s, cognitive psychologists, mathematics educators, and mathematicians had all begun to converge on the importance of understanding the *process* of quantitative reasoning and its complex interactions with the mathematical content by which the reasoning takes place. Although this work stresses the importance of the *interactions* between process and content, recognizing the distinction between the two has made possible important advances in the teaching and learning of mathematics and has focused considerable attention on the quantitative reasoning processes of students at all educational levels. Markman and Gentner (2001) provide a detailed review of the reasoning process, including an analysis of issues of domain specificity and domain generality that have been at the heart of many debates about the nature of quantitative reasoning. Markman and Gentner do not, however, address quantitative reasoning as a reasoning factor *per se*, nor do they address measurement issues in quantitative or other aspects of reasoning.

The history of theory and research on quantitative reasoning shows few attempts at definitive conceptualizations of quantitative reasoning specifically within the context of assessments of individual differences, and nothing specifically within the modern construct validity framework. From a contemporary perspective, early attempts were rather general. Although the NCTM standards are, it could be argued, the first comprehensive attempt since Thorndike to define quantitative reasoning for the purposes of teaching and learning that integrates assessment issues, the standards do not offer a comprehensive view of quantitative reasoning for the specific purpose of assessment.

The NCTM (2000); the Mathematical Association of America (MAA) (Mathematical Association of America [MAA], 2003; Sons, 1996); the American Mathematical Society (AMS) (Howe, 1998); and the American Mathematical Association of Two-Year Colleges

(AMATYC) (American Mathematical Association of Two-Year Colleges [AMATYC], 1995), in their statements about the goals of mathematical education, all discuss quantitative reasoning as an ability that all high school and college students can and should develop. These documents also discuss a great deal of curricular material in addition to quantitative reasoning. They have certain differences in scope as well,[3] but there is substantial agreement among them as to what constitutes quantitative reasoning and what constitutes the mathematics content on which the reasoning is based. The following material reflects that broad and strong consensus in the world of mathematics.

Definition of Quantitative Reasoning Construct for Assessment Purposes

The consensus on differentiating reasoning skills and content that exists in the world of mathematics creates the basis for defining the construct of quantitative reasoning for assessment purposes.

In the realm of measurement theory and practice conceptually parallel work in validity theory by Messick and others now enables us to understand that it is critical to the interpretation of reasoning tests to differentiate between elements of the reasoning construct itself that is the target of the assessment and the common core of content knowledge that all test takers are assumed to bring to the test. The distinction between the *functions* of reasoning (as the construct of interest) and the enabling content that the test takers are assumed to bring to the testing task is an important one to bear in mind in creating a fair and validly interpretable assessment. It is the basis for the assumption that scores reflect construct-relevant variance in quantitative reasoning rather than differences in the construct-irrelevant variance related to the tools needed to approach the reasoning tasks.

In practice it is important to recognize that it is impossible to assess reasoning and content in complete isolation from one another. For the purpose of defining the construct of quantitative reasoning for assessment purposes we have defined mathematics content as that which is required for reading and understanding mathematical problems, such as knowledge of mathematical terminology and mathematical procedures. This includes knowledge of numbers, operations, transformations, and other mathematical topics. Quantitative reasoning itself, on the other hand, is defined as the ability to analyze quantitative information, including the determination of which skills and procedures can be applied to a particular problem to arrive at a solution. It is possible to assess mathematics content while only assessing a

minimal level of quantitative reasoning (e.g., by asking the test taker to implement a specific procedure or to recall the definition of a mathematical term that a person has been asked to memorize). It is not possible, however, to assess quantitative reasoning without the content since it is the manipulation and application of the content that allows test takers to demonstrate their reasoning.

We define quantitative reasoning broadly as the ability to analyze quantitative information. It is, therefore, not restricted to skills acquired in mathematics courses, but includes close up reasoning abilities developed over time through practice in almost all high school or college courses, as well as in everyday activities such as budgeting and shopping.

Quantitative reasoning includes the following six capabilities:

- reading and understanding information given in various formats, such as in graphs, tables, geometric figures, mathematical formulas or in text (e.g., in real-life problems);
- interpreting quantitative information and drawing appropriate inferences from it;
- solving problems, using arithmetical, algebraic, geometric, or statistical methods;
- estimating answers and checking answers for reasonableness;
- communicating quantitative information verbally, numerically, algebraically, or graphically;
- recognizing the limitations of mathematical or statistical methods.

These capabilities are included as competencies that all college graduates should have in the report *Quantitative Reasoning for College Graduates: A Complement to the Standards of the Subcommittee on Quantitative Literacy Requirements of the MAA* (Sons, 1996).

As noted above, there are many different ways that a construct such as quantitative reasoning can be measured. Because quantitative reasoning for assessment purposes is, according to the NCTM and related standards, the ability to solve problems, it is important to understand this process in some detail. The following outline of the problem-solving process, which is consistent with the *Principles and Standards for School Mathematics of the NCTM* (NCTM, 2000), illustrates the use of quantitative reasoning skills in problem solving for

assessment purposes. Note that this conception of quantitative reasoning is independent of any particular mathematical content or level of mathematical achievement.

Outline of the Components of the Problem-solving Process

Step 1. Understanding and defining the problem

- a. Understanding the underlying situation described in the problem
  - i. Understanding and analyzing both the explicit and the implicit information in the problem (this includes defining the constraints of the problem)
  - ii. Finding instances that fit the constraints (e.g., if x is an integer that is divisible by 7, then x could be 14)
  - iii. Recognizing patterns and formulating conjectures about the situation
- b. Translating a real situation into a mathematical representation
- c. Translating a real or pure (that is, not contextualized) mathematical problem into another mathematical problem (e.g., changing a geometric problem to an algebraic problem by using analytical geometry)

Step 2. Solving the problem

- a. Recognizing when to stop trying to understand the problem and when to start solving it
- b. Drawing on a variety of specific techniques that can be used to solve problems (e.g., drawing a picture, using a formula, solving an equation)
- c. Deciding which techniques are most useful for solving the given problem
- d. Applying the techniques correctly to solve the problem

Step 3. Understanding results

- a. Recognizing when to stop solving the problem and start trying to understand what has already been done (this includes monitoring and understanding intermediate results)
- b. Translating results arrived at mathematically into the language of the problem (e.g., if an equation is used to solve a word problem, interpreting what the solution to the equation means in terms of the word problem)
- c. Recognizing implications of the answer or of intermediate results
- d. Checking an answer or a result for reasonableness and recognizing when suggested answers or intermediate results are not reasonable

e. Recognizing when errors have occurred in carrying out the reasoning process

f. Communicating the results, by presenting in writing (or in other modes) clear explanations of the solution process

The problem-solving process described above represents a restriction of the domain of larger quantitative reasoning due to practical constraints of testing such as the testing time available and the number of questions each test taker must answer. In practice, most tests are designed to assess only a portion of the quantitative reasoning process.

Mathematics Content Issues

As noted earlier, the intention of a test of quantitative reasoning is not to assess mathematics content knowledge *per se*, but to use the mathematical content knowledge as a tool with which to reason. In order to engage in quantitative reasoning, however, test takers must use mathematics content. A further distinction should be made between mathematical *content* and *context*. Context refers to the settings that are described in the applied questions on quantitative reasoning tests. Content, in contrast, refers to the knowledge and skills that are needed to answer the questions.

It is of the utmost importance to ensure an assessment's validity and fairness that solutions to test questions should not depend on a level of knowledge that goes beyond the level that is *explicitly assumed* to be common throughout the testing population; otherwise, to the extent that this occurs, the test is unfair to the test takers who do not have that enabling knowledge. Its construct validity is diluted because content knowledge becomes a source of variance extraneous to the intended reasoning construct. If some test takers have less than the assumed level of content knowledge, then they do not have the basic tools they need to solve the problems in the manner intended by the test author, and they are at an unfair disadvantage relative to other test takers. For example, a person with strong quantitative reasoning skills may score low due to their lack of the specific mathematical content knowledge needed to answer particular reasoning questions. As a result, the intended inferences about their quantitative reasoning proficiency cannot validly be drawn from their scores. This is a critical issue for test design and for later interpretation of test scores. Quantitative measures, therefore, are generally aimed at a specific population whose mathematical content knowledge can be precisely determined. This is the reason that when quantitative reasoning

measures are aimed at different populations the exact level of content knowledge required by each measure is usually different.

Despite the strong professional consensus to the contrary,[4] it seems to remain very difficult for many people to think in terms of quantitative reasoning as a conceptually separate entity from mathematical content topics. This creates the potential for difficulties in communicating important information about quantitative assessments to important stakeholders, including students and teachers. Although among knowledgeable professionals such as mathematics educators and mathematicians there is no question that a knowledge base of concepts and procedures is necessary but not sufficient for quantitative reasoning (e.g., Glaser, 1984; Nickerson, 1988; Stodolsky, 1988), research by Wilson, Fernandez, and Hadaway (2002) concludes that students, and even many teachers, tend to be largely unaware of the reasoning processes involved in their problem-solving activities. Wilson et al. express concern that the professional consensus on the centrality of quantitative reasoning is not yet well accepted by students, their families, and many teachers who are uncomfortable with pedagogical techniques that stray from known algorithms and traditional course-oriented mathematics content structure. Thus the language of content rather than reasoning processes continues to pervade discussions of assessments of quantitative reasoning, providing a source of much confusion and misunderstanding (Ma, 1999).

Another related content issue is that of the implications of novelty-of-test-content to the individual test taker. As noted above, from the time of Thorndike through the present-day NCTM standards, there has been a strong consensus that quantitative reasoning is the process of problem solving, not simply memorization or the rote application of known algorithms. Thus test materials that require only these skills cannot be said to be testing quantitative reasoning. For example, among adults, basic arithmetic operations such as multiplication are performed by rote rather than by employing quantitative reasoning skills. When questions in tests of quantitative reasoning are those that have, in fact, been explicitly taught to some or all of the test takers, they can no longer be considered true measures of quantitative reasoning for those students.

Mathematical test content should also be distinguished from quantitative reasoning for assessment purposes in terms of difficulty level. Quantitative reasoning tasks, being conceptually independent of the mathematical content with which test takers reason, can be

made relatively easier or more difficult for a given population within a given level of mathematical content. Other sources of valid differences in reasoning difficulty include, for example, the number of independent reasoning processes needed to arrive at a correct solution. Test development experience has shown that quantitative reasoning test questions can be made very difficult using only an elementary basis of mathematical content by employing such strategies as increasing the number and complexity of the reasoning processes involved in reaching a solution. Higher content requirements can be of practical value because they permit a wider range of items. It also seems logical to assume, although we are not aware of the existence of any data specifically on this point, that such content requirements might have the positive consequence of promoting additional study of important mathematical concepts. The overriding concern from the point of view of fairness, however, should be the point we have repeatedly emphasized: that the mathematical content in an assessment of quantitative reasoning should include only that which all test takers can be assumed to possess.

The interpretation of quantitative-reasoning test scores rests on the assumption that all students have the mathematical content knowledge necessary to solve the required reasoning tasks on the test. Therefore, care must be taken to ensure that all students have the necessary knowledge before test makers increase the level of mathematical content knowledge required. If test takers do have the necessary content knowledge, of course, then increasing the content requirements alone will not appreciably increase the difficulty of the questions.

If the test takers do not all have the appropriate higher content knowledge—that is, if the mathematical content is *beyond* that which is explicitly assumed to be necessary to answer the reasoning questions—then a source of construct-irrelevant variance has been introduced into the test to the detriment of its validity, its fairness, and its utility.

Mathematics Content Categories Used in Quantitative Reasoning Assessments

In describing the mathematics content of quantitative reasoning assessments in the sections below, we will follow the structure used in the NCTM standards for Grades 9-12 (NCTM, 2000). The first five of the NCTM standards are the content categories: Number and Operations, Algebra, Geometry, Measurement, and Data Analysis and Probability. In the discussion below, we combine Geometry and Measurement. We will also refer to the *Crossroads in Mathematics* position paper by the AMATYC (1995); and the MAA report

Quantitative Reasoning for College Graduates (Sons, 1996). The MAA report does not separate content into categories, but it specifies that students should be able to "Use arithmetical, algebraic, geometric and statistical methods to solve problems." (www.maa.org/past/ql/ql_part2.html)

- *Numbers and Operations.* The NCTM standard includes understanding numbers, understanding meanings of operations, and being able to compute fluently. This category matches the Number Sense category from the AMATYC standards, where it is defined as follows: "Number Sense includes the ability to perform arithmetic operations, to estimate reliability, to judge the reasonableness of numerical results, to understand orders of magnitude, and to think proportionally" (AMATYC, 1995, chapter 2). Some topics included in this category are percents, ratios, place value, and multiples and remainders.
- *Algebra.* The NCTM standard is Algebra, the AMATYC standard is Symbolism and Algebra. Both standards include the ability to translate problem situations into algebraic representations and to use algebraic representations to solve problems. Some topics included in this category are linear equations, algebraic manipulation, and the translation of a word problem to an algebraic expression.
- *Geometry and Measurement.* The NCTM document contains separate standards for Geometry and for Measurement. The AMATYC includes measurement in its Geometry standard. Both standards include traditional plane and three-dimensional geometry as well as analytic geometry. Some topics included in this category are areas and perimeters of two-dimensional objects, volumes of three-dimensional objects, parallel lines, and angles in the plane.
- *Data Analysis and Probability.* This is the last of the NCTM content categories. AMATYC's corresponding standard is Probability and Statistics. Both standards include reading and understanding tables and graphs, as well as simple statistics and probability. The MAA states that a quantitatively literate college graduate should be able to "interpret mathematical models such as formulas, graphs, tables, and schematics, and draw inferences from them" (http://www.maa.org/past/ql/ql_part2.html). Some topics included in this category are

tables, line graphs, bar graphs, circle graphs, counting, simple probability, mean, and median.

Considerations in Measuring Quantitative Reasoning: Threats to Validity

In specifying the construct of quantitative reasoning, it is important to describe not only the target construct, but also elements that, although not part of the construct itself, may form part of the measurement experience. In addition to the case of mathematical content discussed above, four more of these elements are discussed below. Each is a potential source of construct-irrelevant variance. The intrusion of these elements may be justified to a greater or lesser extent in a particular testing situation, depending on such logical factors as the degree to which they could be construed as relevant to the construct and the extent to which they can be shown to have minimal effects on score differences. Practical considerations also include such factors as the composition of the target population, considerations relating to the physical environment of the test, and anticipated costs and benefits of testing.

In order to reach a judgment about the overall quality of the assessment, the degree to which inferences based on the test scores are free of construct-irrelevant variance must be weighed, in each case against considerations such as the utility and expense of the test and the test-takers' access to the assumed level of content knowledge.

The Role of Calculation in Assessment

In the Number and Operations standard of the *Principles and Standards for School Mathematics*, the NCTM (2000) states that instructional programs should enable all students to compute fluently and to make reasonable estimates. In their comments, reviewers for the MAA and AMS strongly support this view. Computational fluency assumes the ability to perform mental arithmetic and to carry out paper-and-pencil calculations or to use a calculator when appropriate, as well as to have a sense of the relative size of numbers and their orders of magnitude.

In a quantitative reasoning test, however, the focus should not be on testing computational skills *per se* (which may be an important construct itself in other settings), but rather on assessing

- the ability to select and use appropriate and efficient computational strategies to solve a problem and

- the ability to determine the necessary degree of accuracy of a calculation or the appropriateness of using estimations.

Given the constraints of a timed test, problems should not require complicated and lengthy calculations, whether or not a calculator may be used. If some computation is necessary to arrive at an answer, the required accuracy should be clearly indicated, either by the wording of the question or by a judicious selection of the options in a multiple-choice question. Questions that can be answered either by lengthy calculations or by other, less time-consuming methods should be avoided as much as possible since the availability of two very different approaches affects validity and introduces an additional source of error. For example, if the answer choices for a question are sufficiently far apart to allow a student to solve the problem by estimation, then the wording of the question should indicate to students that they are being asked to make an approximation, not a lengthy calculation.

Issues Related to Spatial and Visual Abilities

Certain spatial and visual concepts are often included in discussions of quantitative reasoning, but although they may be important constructs themselves in other settings, they are not a critical part of our basic definition. Among the aspects frequently discussed are spatial visualization and the use of visual representations of quantitative information, such as graphs and charts in mathematical problem solving.

Spatial visualization is a multifaceted ability to mentally manipulate representations of two-and three-dimensional objects and to perceive an object from different perspectives. It can be used very effectively to solve certain types of mathematical problems. Although the ability to manipulate a visual representation mentally is useful in many settings, mental visual manipulation is not in itself a quantitative reasoning process, but rather one of several ways in which certain quantitative reasoning problems can be solved. Thus it is not an essential component of quantitative reasoning. In addition, spatial visualization is not a skill in which all test takers can be assumed to be equally proficient. Research has demonstrated that certain forms of spatial visualization, particularly three-dimensional mental rotation, are significantly more difficult for some test takers than for others (e.g., Casey, Nuttall, & Pezaris, 1997; Casey, Nuttall, Pezaris, & Benbow, 1995; Linn & Peterson, 1985, 1986; Masters & Sanders, 1993; Shea, Lubinski, & Benbow, 2001).

In a similar fashion, visual displays of information such as charts and graphs are used along with other formats to communicate information about quantitative material, but they are not a critical part of the construct of quantitative reasoning. As noted above, an important aspect of quantitative reasoning is to be able to read, understand, and communicate about quantitative material given in a variety of formats. Frequently used formats include numerical, verbal, symbolic, and graphical representations. No particular format can in itself, however, be said to be essential to the process of quantitative reasoning. As is the case with spatial visualization, some graphical displays are significantly more difficult to utilize for some test takers than for others, including (but not limited to) people who are blind.

It should be noted that the concern about factors such as spatial and visual representations goes beyond item format. One should also be concerned with the extent to which other item formats show evidence of requiring spatial visualization strategies. Such information can probably best be gleaned from special studies, as inspection and routine item analyses are likely to be ineffective for detecting it.

Reading Load and Quantitative Reasoning

Reading ability *per se* is also not part of the construct of quantitative reasoning, beyond the ability to comprehend questions in contexts related to real-life situations. It should be recognized, however, that a certain level of reading ability is required in order to understand and correctly answer test questions.

For a test to be valid and fair, test makers must be able to make certain assumptions about the common knowledge that test takers bring to the testing situation. In assessing quantitative reasoning, an important assumption is that all test takers understand the intended meaning of the test materials, just as it is important to be able to assume that all test takers have the content knowledge with which to demonstrate their reasoning capabilities. Orr (1997) argues that certain linguistic features of the language of mathematics disadvantage African-American learners. For example, some test takers may not recognize the distinction between "twice" and "half" and/or use different methods of expressing quantities versus differences among quantities.

As noted earlier in the basic definition of quantitative reasoning, an important aspect of such reasoning is the ability to translate standard English into the "language" of mathematics, and vice versa. The vocabulary and sentence structure of the questions should,

however, be at an appropriate level for the population being tested. Questions that involve verbal contexts should be written using the clearest possible English that accurately conveys the intended meaning.

Speed of Responding

Discussions about the proper timing of quantitative tests, particularly at the whole-test level (as opposed to the single-question level), include clearly differing points of view on psychometric and psychological aspects of test speed. The following discussion does not directly address the issue of the *nature* of speed of response or how it is measured. For example, no distinction is made between speed defined as reaching the end of a test[5] versus speed defined as working at an individually optimal or preferred rate.

The following discussion also does not address practical constraints that may have to be set with regard to the time available for administering tests of quantitative reasoning. Instead, our discussion focuses on the relationship between test speed and the construct that the test is intended to represent. The rationale for looking first at the relationship between test speed and construct information is to enable a consensus to be reached on a full and technically acceptable definition of the target construct of quantitative reasoning that reflects policies of the test maker. Once that consensus has been reached, policy decisions must be made regarding how best to represent this construct while balancing constraints imposed by concerns for issues such as practicality, fairness, and usefulness.

Technical definitions of quantitative reasoning and taxonomies of mathematics content and abilities that relate to tests of quantitative reasoning do not include the element of speed of response as one of their core elements. Therefore, although a test may show speededness, this should be considered an irrelevant source of variance and attempts should be made to control it.

Arguing in favor of retaining a significant level of test speededness, some point out that speed of problem solving is an indicator of proficiency. Some highly competent mathematicians are in fact known for their extraordinary ability to solve problems quickly, although most of the profession agrees that this skill is incidental to quantitative reasoning ability. Further probing from a cognitive point of view elicits the observation that rapidity of response is associated with overlearning of basic concepts. This overlearning is a tool used to

enable people to engage in higher-level quantitative reasoning, and which some research has shown to be characteristic of experts as opposed to novices.

Some argue in favor of considering speed of problem solving to be irrelevant to the core construct of quantitative reasoning that speed of response includes personality, cultural, and individual stylistic components.[6] Some of those who take this point of view argue that this is exclusively so, while others who are also in favor of eliminating the speed factor argue that these noncognitive sources of variability in problem-solving speed coexist with cognitive sources of variability as indicated above. It is generally agreed, however, that those whose work is based on a high level of quantitative reasoning ability, such as mathematicians, engineers, and scientists, show a great diversity in their ability to work quickly, and in their preferences to do so in their daily work. Thus we conclude that speed and quantitative reasoning ability are likely to be independent of one another, at least among specialists and probably in the general population as well.

There is considerable consensus about the relationship of speed to the most common criteria, the work that needs to be done at the high school, college, and university level, and later in professional situations. Many would agree that although it may be desirable to be able to work quickly (or to be in the habit of doing so), a more significant determinant of success is the motivation to spend sufficient time to complete work successfully.[7] Thus, in most cases, it is not highly significant whether a student or professional using mathematics *can* work quickly or not, as long as he or she is willing to "put in the time" needed to complete a given task successfully.

At the level of the single question, in contrast to the total test, there is more consensus about the role of speed. The desirability of controlling the time requirements across questions is increasingly recognized in test specifications and in instructions to question writers and reviewers. Questions that allow different solution paths requiring different average completion times are increasingly seen as problematic for most testing situations, in that they introduce additional measurement error. It is difficult to predict, however, what specific methods individual candidates will use to answer a question. As a practical matter, it may be more feasible to attempt to create time limits broad enough that different solution paths do not become problematic rather than to attempt to ensure that questions have only one solution path.

Given the nature of this discussion, and a consideration of the following factors, the preponderance of evidence indicates that the construct of quantitative reasoning does not include speed of response.

1) Standards and other well-supported definitions of quantitative reasoning that do not include speed as a component;
2) The multiplicity of sources of speed differences across individuals and population groups (cognitive, personality, motivational, and sociocultural);
3) Evidence of variability of speed in other criterion measures such as the rate-of-work styles of experts; and
4) The time requirements of high school, college and graduate study and the majority of professional work settings.

It is desirable that time requirements of question types and individual questions be controlled through test specifications and in test construction. Monitoring time requirements of individual items has become feasible in computer-administered testing.

Descriptive Materials for Tests

In order to illustrate some of the complexities inherent in defining and operationalizing the quantitative reasoning construct, and to consider what are currently provided as construct definitions, we reviewed several quantitative reasoning assessments developed by ETS. First we review how each testing program describes its tests; then in a later section we compare the contents of these assessments of the quantitative reasoning construct. The tests we reviewed are the GMAT® (Graduate Management Admission Test®); the GRE® (Graduate Record Examinations®) General Test's quantitative measure; the PPST® (Praxis Pre-Professional Skills Test for beginning teachers); the PSAT (for high school juniors, although it is also taken by many high school sophomores); and the SAT® I reasoning test (primarily for high school juniors and seniors).

A number of sources provide information to the public about the quantitative reasoning tests that we reviewed, including information about their content and how to prepare to take them. What follows is a description of quantitative reasoning, mathematics content, and test preparation advice that can be found in a wide variety of published materials available to the general public. This material is the primary means by which the intended

construct of the tests is conveyed to test takers and other interested parties who will make inferences based on the scores from these tests.

Descriptions of Quantitative Reasoning

For each test we consider in this paper, the respective program's published descriptive materials state that the test measures basic mathematical skills and the ability to reason quantitatively and/or to solve problems involving mathematical concepts. The actual wording and the order in which mathematical content areas are listed varies from test to test and also across the various publications that describe a given test. All testing programs whose materials we reviewed mention that the skills measured are assumed to be learned over time, and that the mathematics content is limited to that which is usually covered in high school or to that studied by all examinees. Some programs state this the same way in every publication; others do so in only a few of their publications. No more than single-sentence descriptions are provided to explain the nature of quantitative reasoning or its component skills.

The distinction between quantitative reasoning and mathematics achievement, on the other hand, is made by only a few of the programs. The GMAT, for example, states in several of its publications that the quantitative measure does not assess achievement in specific fields such as economics or accounting. The GRE *Technical Manual* (Briel, O'Neill, & Scheuneman, 1993) states that the test is "intended to assess quantitative reasoning rather than mathematical achievement" (p. 10), but such explicit statements do not appear in other GRE publications.

The SAT, PSAT and GRE programs have all worked to describe the attributes that constitute quantitative reasoning. The attributes identified by the GRE diagnostic service, which grew out of this research, are primarily content-based, with just a few attributes directly targeting reasoning. As of 2001, the PSAT provides students with enhanced score reports that include information about their performance on various attributes (or question types). Like the GRE, the PSAT attributes are based on a combination of reasoning categories, content categories, and response formats (e.g., "multiple-choice").

Identifying reasoning attributes for any test is a very time- and labor-intensive activity. In order to be useful to test developers and score users, the attributes must be objective and clearly defined to allow for reliable coding; they must accurately describe the complete test; and they must be statistically verifiable. They must also be meaningfully defined for test

takers, and advice should be provided to test takers on how to improve performance on questions assessing these attributes. At present, we know of no testing program that completely achieves this goal.

Descriptions of Mathematics Content Necessary for Quantitative Reasoning

None of the descriptive materials that we reviewed delineates explicitly the entire scope of the mathematical knowledge candidates are assumed to have in order to answer the questions and what mathematical content is actually to be measured by the test. In fact, some of the descriptions of content are to some degree contradictory with respect to the role that content plays in the measure. For example, the description of the GMAT quantitative measure in the GMAT *Guide* (Graduate Management Admission Council [GMAC], 2000) states that the "questions require knowledge of arithmetic, elementary algebra, and commonly known concepts of geometry" (p. 8), while the Math Review section of the GMAT *Guide* (GMAC, 2000) lists very similar mathematical content topics, describing them as "the kinds of topics that are tested [sic]." (p. 31) These two descriptions leave some doubt as to whether the content areas listed constitute baseline knowledge that all test takers are expected to bring to the task of answering the reasoning questions, or whether, on the other hand, the test is intended to assess mathematics content *per se* rather than assessing the ability to *reason* with that content.

Each program we reviewed does give an explicit list of mathematical content topics, usually in the program's test preparation materials. These topics are referred to as "[mathematics that] it is expected examinees are familiar with" (p. 22) by the GRE (Graduate Record Examinations [GRE] Board, 1998); "topics that are tested" (p. 8)by the GMAT (GMAC, 2000); "topics covered" (p. 57) for the PPST (Educational Testing Service, 2001); and "general mathematics concepts" (p. 15) for the SAT (College Board, 2001).

No listing that we reviewed makes any claims of being exhaustive or of containing all the topics that a test taker needs to know. There are substantial differences, however, in the level of detail supplied; the GRE and GMAT listings are fairly high-level and general, while the SAT and PPST listings are more specific and detailed. Some descriptions mention topics that are *not* included. For example, the GRE, GMAT, and PSAT each specify that they do not expect test takers to be able to construct geometric proofs. The SAT and the PSAT also provide a list of commonly used formulas in the test book itself, so that test takers do not have

to memorize them. This sends a clear signal to the test taker about the reasoning nature of the test.

None of the lists of content topics indicates the level of difficulty of the content used in test questions. All materials do, however, refer the reader to sample questions and/or tests, and strongly recommend studying these. The GRE, GMAT, and SAT also publish samples of the content of their quantitative tests as part of their test preparation materials. Altogether, these materials provide a fairly complete description of the various tests.

It should be noted that the content descriptions included in the test specifications for test developers are essentially the same as those published for the general public. Together with other materials, such as question writing guides, style guides, and *ad hoc* memoranda, they define the mathematical content permissible for each test. In addition, test developers use their own experiences, in high school or college teaching and in test development, to judge the difficulty and appropriateness of test questions. Other experienced mathematics educators who are in regular contact with the test-taking population are also usually involved in the writing of the questions.

Test Preparation

The programs that we reviewed offer test takers a range of free and priced test preparation materials in print and online formats, including full sample tests. The free materials that we reviewed include advice on how to take tests, descriptions and reviews of the content that tests cover, and a sample of actual test questions.

As noted above, examples of the types of problems examinees may encounter in various areas of mathematics are provided in preparation materials, but no exhaustive listing of the mathematics content that every test taker is assumed to have is provided, nor is quantitative reasoning explicitly defined. Preparation materials also provide detailed explanations of solutions to problems, which include descriptions of the quantitative reasoning that required. They do not, however, cluster these descriptions into categories of reasoning that may be found on the test. Publicly available materials for all the tests discussed here state in some form that, aside from familiarization with test content and format, the best preparation for taking the test is to study hard in school. This is very consistent with the nature of the construct of quantitative reasoning as defined in this paper, and explicitly

supports the important point that the skills measured in tests of quantitative reasoning are developed in response to the quality of one's education and are thus malleable.

Comparisons Among Quantitative Reasoning Tests

Because quantitative reasoning is a complex construct that can be measured in different ways, and because there are practical constraints on any test (e.g., time and the number of questions that can be asked), no test of quantitative reasoning should be expected to cover the entire construct, or to cover the construct in exactly the same way. The target population for each test differs in the kinds of mathematics they have studied and in the length of time they have been away from instruction on the mathematics content that is used in the test. As a result, each test that we reviewed has a slightly different focus or emphasis that has been designed specifically for the target population. Table 2 displays characteristics of the tests discussed here in terms of the intended test-taking population, the test length (number of questions and time), the delivery mode, the content, and the types of questions that make up the test.

Table 2

Characteristics of Tests

Test	Intended population	Testing time	# of questions	Delivery mode	Description of content	Question types
GMAT	College graduates	75 min.	37 questions, one section	CBT CAT	Basic mathematical skills, understanding of elementary concepts, the ability to reason quantitatively, solve quantitative problems, and interpret graphic data	Standard MC, DS

(Table continues)

Table 2 (continued)

Test	Intended population	Testing time	# of questions	Delivery mode	Description of content	Question types
GRE	College graduates	45 min.	28 questions, one section	CBT CAT	Basic mathematical skills and understanding of elementary mathematical concepts, as well as the ability to reason quantitatively and to solve problems in a quantitative setting	Standard MC, QC, QS
PRAXIS PPST	First or second year college undergrads	60 min.	46 questions, one section	Paper & pencil or CBT CAT	Key concepts of mathematics and the ability to solve problems and to reason in a quantitative setting	Standard MC
PSAT[a]	High school sophomores or juniors	Two sections of 25 min.	40 questions (two sections of 20)	Paper & pencil	Basic arithmetic, algebra, and geometry, plus miscellaneous	Standard MC, QC, SPR
SAT[a]	High school juniors or seniors	Two 30 min. sections, one 15 minute section	60 questions (two sections of 25 questions, one of 10)	Paper & pencil	Basic arithmetic, algebra, and geometry, plus miscellaneous	Standard MC, QC, SPR

Note. CBT = computer-based testing, CAT = computer adaptive testing, DS = data sufficiency, MC = multiple-choice, QC = quantitative comparisons, QS = question sets, SPR = student-produced response.

[a]According to PSAT and SAT program sources (J. S. Braswell, personal communication, May 31, 2003), the QC item type is being dropped and only standard MC and SPR item types will be used. For the PSAT, the change will take place in the fall of 2004; for the SAT, the change will take place in the spring of 2005.

Content Categories and Their Distribution

In general, it can be seen from Table 2 that the mathematics content of these tests is very similar. There are, however, differences in emphasis among the tests. None of the tests provides an explicit rationale for the distribution of questions and content categories, but it is clear from their development that the current designs result from a combination of factors such as time available for testing and a tendency toward equal weighting of categories as the "default" position. On the SAT I and PSAT tests, the questions are approximately evenly divided among arithmetic, algebra, and geometry, with a smaller number of questions that are classified as miscellaneous. There is also some tendency for geometry to be more heavily weighted in tests for high school students, who can be expected to have studied geometry more recently. Test development staff point out that quantitative questions for high school students also carry a lower reading load than do those aimed at populations that are of college age or older.

Like the PSAT and the SAT, the GRE also has questions on arithmetic, algebra, and geometry. It is assumed, however, that many GRE test takers have been away from instruction on this mathematics content for several years, and therefore many of the questions are less demanding in terms of recall of mathematics facts. These three categories are approximately evenly represented in the GRE, with eight to nine arithmetic questions, six to seven algebra questions, and five to six geometry questions. The GRE test also adds another category: Data Interpretation. This is included because data interpretation is considered by the GRE Board to be an important skill that students should have when they enter graduate school. Data interpretation questions are considered a content category rather than a question type, but they have some uniform qualities. They are all based on graphical or tabular information, and frequently appear in sets (otherwise some of the questions would be too time-consuming, given the "investment" of time needed to understand the data to be interpreted). Six to nine of the 28 GRE questions are data interpretation questions. On the SAT I, PSAT, and GMAT tests, there may be questions concerning data, but data does not appear as a separate classification.

On the GMAT test, questions are classified as arithmetic, algebra, or geometry. Arithmetic is considered to be the most important category for performance in business

school, and geometry is considered to be the least important. On this test, 16 out of 28 questions are arithmetic, nine are algebra, and only three are geometry.

The PPST uses a different classification scheme to represent similar content. The PPST tests basic mathematics skills for prospective teachers. The classifications used are:

Conceptual Knowledge—Number sense and operation sense (6 questions)

Procedural Knowledge—Ways to represent quantitative relationships and solve problems (12 questions)

Representations of Quantitative Information (12 questions)

Measurement and Informal Geometry (12 questions)

Reasoning in a Quantitative Context (4 questions)

Question Contexts

The ability to solve applied problems (problems posed in a real-life or familiar context) is included as an important part of quantitative reasoning by all of the programs whose materials we reviewed. This ability cannot be assessed without contextualized questions. Most of the other aspects of quantitative reasoning can be assessed in either a purely mathematical or an applied setting. Accordingly, all of the quantitative reasoning tests contain some applied questions, but the proportion of such questions varies substantially from program to program. In the GRE, there are 10 out of a total of 28 questions on each test form that are applied questions; in the GMAT, there are 13 out of a total of 28 questions; in the PPST, there are 20 out of 40 questions; and in the SAT, 13-16 out of 60 questions.

As noted above, context should be distinguished from mathematical content. Context refers to the settings that are described in the applied questions on quantitative reasoning tests. Content, in contrast, refers to the mathematical knowledge and skills needed to answer the questions. As used here, context is a concept closely related to face validity.

The contexts used in applied questions are restricted to settings that can reasonably be assumed to be familiar to all test takers; therefore, they vary somewhat among the programs. For example, there are more business-related questions on the GMAT than on the GRE, and the PPST frequently uses school-related settings because its test takers are prospective teachers. The emphasis in these measures is on settings that may be encountered by adult professionals, rather than on settings typically found in high school mathematics textbook questions.

The language used in applied questions is nontechnical, and information useful for a specific context is given in the question, unless it can be assumed to be generally known. Scientific or other highly technical settings are avoided in all programs reviewed here. In addition, all real-life settings must conform to the ETS *Standards for Quality and Fairness* (Educational Testing Service, 2002).

In the GRE at least four of the applied questions on each form concern data interpretation; in these, some real-life data are represented in graphical or tabular form and one or more questions about the data are posed. Questions that require making inferences from the data and applying statistical concepts are included in this category.

Question Types

The tests discussed here use a variety of question types that are described below. Although there have been some intensive efforts to study the cognitive aspects of question types (e.g., Gallagher, 1992; Messick, 1993; Traub, 1993), a great deal of research could still be done to further explore the aspects of quantitative reasoning that are associated with different question types in current operational testing programs.

Table 2 refers to five question types present in the programs we reviewed. They are:

Standard multiple-choice. Standard multiple-choice (five-choice MC) is by far the most common question type and it appears in all of the tests that we reviewed. This question type is generally considered to be useful for testing a wide range of quantitative problem-solving skills. The format consists of a question followed by a set of five unique options from which the test taker must select an answer. As indicated in Table 2, the number of questions of this type varies across tests, and it is not clear that this variation is directly related to such construct-related reasons as the population served or the basic intention of test use. On the SAT I, 35 out of 60 questions are of this type; on the PSAT, there are 20 out of 40 questions; on the GRE 14 out of 28 questions; and on the GMAT, 16 out of 28 questions. On the PPST, all 40 questions are five-choice multiple choice.

Question sets. The GRE includes two sets of two questions each. The questions in a set share one data stimulus, such as a table or graph. This allows for questions about a stimulus that may be too long or complicated for use with only one question (an example of this is the data interpretation content category discussed above) and thus can add cognitive complexity to the assessment. The questions in these sets are five-choice multiple choice, and

are included in the 14 GRE multiple-choice questions. Question sets may also appear on other tests, but their use is not required by test specifications.

Quantitative comparison. The quantitative comparison (QC) question type is found on the SAT I, the PSAT, and the GRE. As noted above, this question type is useful for testing quantitative concepts without requiring detailed calculations, which are not part of the target quantitative reasoning construct. Questions display quantities in two columns, which test takers are asked to compare by size. The test taker must determine if one quantity is larger than the other, whether they are of equal value, or whether the relationship between the two quantities cannot be determined. As with other question types, the number of these questions varies by test. Of this question type there are 15 on the SAT I, 12 on the PSAT, and 14 on the GRE.

Data sufficiency. The data sufficiency (DS) question type is now unique among the programs examined here to the GMAT, although it was previously used in other programs, including the SAT, and is currently used in the Swedish SAT. Like the QC question type, these questions are well suited for testing understanding of quantitative concepts without requiring the test taker to perform detailed calculations or algebraic manipulations. In this question type, the test takers are given a question for which they need to determine whether the given information is sufficient to answer the question. On the GMAT, 12 questions are of this type.

Student-produced responses. Like the five-choice multiple-choice question, student-produced responses (SPRs) can be used to assess a wide range of quantitative reasoning skills, and the active production of answers very likely adds an additional dimension to the quantitative reasoning skills that can be assessed. SPR questions can be found on the SAT I and the PSAT. In this question type, the answer is always a number, which the student must fill in (no choices are supplied). On the SAT I there are 10 questions of this type, and on the PSAT eight questions of this type.

See the appendix for examples of the standard multiple-choice, quantitative comparison, data sufficiency, and student-produced responses question types.

Probably the most important distinction to be made among the quantitative reasoning question types is between the SPRs and the other question types, which are all multiple choice. The SPRs clearly have great face validity as measures of quantitative reasoning. They

also require somewhat different quantitative reasoning skills in that an answer must be independently calculated and then written (gridded) in, rather than recognized from a list of given options (such options can also be used to check one's answers [Gallagher, 1992]). In this way, the SPRs extend the range of the aspects of the quantitative reasoning construct that are measured. It should be recognized, however, that many students attempt to solve the multiple-choice question first and then match their answer to one of the given choices. Thus differences between SPRs and standard multiple-choice questions occur primarily for those students who eliminate choices or "backsolve" as a test-taking strategy.

To a certain extent, the QC and the DS question types, although they are multiple choice in format, differ from traditional multiple-choice questions. Because both of these question types require test takers to evaluate given quantities, it is likely that they require somewhat different cognitive skills than do the other multiple-choice questions, which pose different tasks. There is not, however, a clear record of why these question types are and are not included in particular tests. Most of the rationales provided (e.g., Donlon, 1984) suggest strongly that the decisions are related to practical matters rather than to using these question types to expand the aspects of the quantitative reasoning construct that can be measured. These practical needs include providing more questions in the same testing time, reducing the total testing time without lowering the tests' reliability (all other things being equal, tests with more questions are more reliable) and controlling the effects of short-term coaching. The QC and DS question types also differ in that their responses are fixed, that is, always the same set of responses; other multiple-choice questions almost invariably have response options that are unique to a particular question. Such question types have generally been shown to be more susceptible to short-term coaching efforts than the question types whose answers are free to vary (Donlon, 1984, p. 195).

There is much still to be learned in construct terms about the effects of including different types of questions in differing proportions for different assessments of quantitative reasoning. There is currently some variation in the types and numbers of quantitative reasoning questions in the programs we reviewed, but this variation is not well documented, or even well understood, with respect to the aspects of the quantitative reasoning construct that are to be covered in each test.

Summary and Conclusions

A complete view of quantitative reasoning from a construct perspective requires consideration of both what quantitative reasoning is and what it is not. We have followed the consensus in the mathematics research and practice communities that is embodied in the NCTM standards and the related postsecondary standards of the AMATYC, the AMS, and the MAA in defining quantitative reasoning. In this view, quantitative reasoning is a problem-solving process with components that can be specified in detail and that can be analyzed empirically and logically for instructional, research, and measurement purposes. Quantitative reasoning is related to, but distinct from, traditional content categories of mathematics curriculum and instruction. The quantitative reasoning process for measurement purposes is more limited than the entire universe of quantitative reasoning. Thus our model of the construct for measurement purposes is presented in the form of a problem-solving process with three steps, each of which has multiple parts. This process is further delimited in any given measurement situation because most questions on tests are designed to assess only a portion of the complete problem-solving process.

A construct of quantitative reasoning is further described in terms of its relationship to categories of mathematics content. To assess quantitative reasoning for the purposes presented in this paper it is necessary to utilize mathematics content, almost exclusively in four areas well-defined by the NCTM's and related standards: Number and Operations; Algebra; Geometry and Measurement; and Data Analysis and Probability. To ensure test fairness and score interpretability solutions to test questions should not depend on a level of content knowledge beyond that which is assumed to be common to all of the test takers. Because this knowledge differs for different test-taking populations, tests of quantitative reasoning can be expected to vary with respect to the mathematical content of the questions.

Four elements that we do not view as part of the quantitative reasoning construct *per se* are discussed because they are commonly encountered in the measurement of quantitative reasoning. These are calculation load required by test questions, the inclusion of spatial and visual materials, reading load of test questions, and speed of responding. Each of these elements represents a potential obstacle to clear interpretation of test scores as they may affect test performance. Because they are not part of the intended quantitative reasoning construct, they need to be carefully evaluated in a particular testing context.

To illustrate different approaches to operationalizing the construct of quantitative reasoning major ETS tests that contain a quantitative reasoning measure were compared and contrasted with respect to their measurement of quantitative reasoning, their inclusion of mathematical content, and their published statements related to defining a construct of quantitative reasoning for their assessment purposes. The sources of information that we used for this comparison include published test preparation advice; explanations of the mathematics content in the test and how it is used; and test specifications, including descriptions of the question types used by the programs to assess quantitative reasoning. Although the target construct of quantitative reasoning appears to be highly similar among the tests we reviewed, specific variations with respect to test content and format exist and seem related to such factors as differences in program purposes and populations.

The publications and the research record for most quantitative reasoning assessments do not fully address the central question of what the test is intended to measure. A more comprehensive construct-centered approach is needed in order to link the assessment's core concept to the actual content of the test and the type of reasoning it assesses. This would lead us to answering the most important question of all, how an individual's scores on the test should be interpreted.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Memphis, TN: Author.Washington, DC: American Educational Research Association.

American Mathematical Association of Two-Year Colleges. (1995). C*rossroads in mathematics: Standards for introductory college mathematics before calculus.* Retrieved October 15, 2002, from http://www.imacc.org/standards/

Ball, D. L. (1991). Teaching mathematics for understanding: What do teachers need to know about subject matter? In M. M. Kennedy (Ed.), *Teaching academic subjects to diverse learners* (pp. 63-83). New York: Teachers College Press.

Braswell, J. S. (1991, April). *Overview of changes in the SAT mathematics test in 1994*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Braswell, J. S. (1992). Changes in the SAT in 1994. *The Mathematics Teacher*, *85,* 16-21.

Briel, J. B., O'Neill, K. A., & Scheuneman, J. D. (1993). *GRE technical manual: Test development, score interpretation, and research for the Graduate Record Examinations Board*. Princeton, NJ: Educational Testing Service.

Brigham, C. C. (1923). *A study of American intelligence.* Princeton, NJ: Princeton University Press.

Brigham, C. C. (1932). *A study of error: A summary and evaluation of methods used in six years of study of the scholastic aptitude test of the College Entrance Examination Board.* New York: College Entrance Examination Board.

Burt, C. (1922). *Mental and scholastic tests.* London: P. S. King and Son.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge, England: Cambridge University Press.

Casey, M. B., Nuttall, R. L., & Pezaris, E. (1997). Mediators of gender differences in mathematics college entrance. *Developmental Psychology, 33*(4), 669-680.

Casey, M. B., Nuttall, R., Pezaris, E., & Benbow, C. P. (1995). The influence of spatial ability on gender differences in mathematics college entrance test scores across diverse samples. *Developmental Psychology, 31*(4), 697-705.

College Entrance Examination Board. (1926). *The work of the College Entrance Examination Board 1901-1925.* Boston, MA: Ginn.

College Board. (2001). *Taking the SAT I Reasoning Test.* New York: Author.

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process*. Boston: Heath.

Donlon, T. F. (1984). *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests.* New York: College Entrance Examinations Board.

Educational Testing Service. (2001). *Praxis I Academic Skills Assessment test at a glance.* Princeton, NJ: Author.

Educational Testing Service. (2002). *ETS standards for quality and fairness.* Princeton, NJ: Author.

Gallagher, Ann. (1992). *Strategy use on multiple-choice and free-response items: An analysis of sex differences among high scoring examinees on the SAT-M* (ETS RR-92-54). Princeton, NJ: Educational Testing Service.

Galton, F. (1978). *Hereditary genius.* New York: St. Martin's Press. (Original work published 1869)

Glaser, R. (1976). Cognitive psychology and instructional design. In D. Klahr (Ed.), *Cognition and instruction* (pp. 303-316). Hillsdale, NJ: Lawrence Erlbaum Associates.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist, 39*(2), 93-104.

Glaser, R., & Baxter, G. P. (2002). Cognition and construct validity: Evidence for the nature of cognitive performance in assessment situations. In H. Braun & D. Wiley (Eds.), *Under construction: The role of construct in psychological and educational measure. A Festschrift in honor of Sam Messick* (pp.179-192). Mahwah, NJ: Lawrence Erlbaum.

Gordon, R. A. (1997). Everyday life as an intelligence test: Effects of intelligence and intelligence context. *Intelligence, 24*(1), 203-320.

Graduate Management Admission Council. (2000). *Official guide for GMAT review.* McLean, VA: Author.

Graduate Record Examinations Board. (1998). *Practicing to take the General Test.* Princeton, NJ: Educational Testing Service.

Greeno, J. G. (1978). Understanding and procedural knowledge in mathematics instruction. *Educational Psychologist, 12,* 262-283.

Greeno, J. G., & Simon, H. A. (1988). Problem solving and reasoning. In R. Atkinson, R. Herrnstein, G. Lindzey, & R. Luce (Eds.), *Stevens' handbook of experimental psychology learning and cognition* (2nd ed., pp. 589-672). New York: Wiley.

Guilford, J. P. (1959). *Personality*. New York: McGraw Hill.

Howe, R. (1998). The AMS and mathematics education: The revision of the "NCTM standards." *Notices of the AMS, 45*(2), 243-247.

Linn, M. C., & Peterson, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development, 56*, 1479-1498.

Linn, M. C., & Peterson, A. C. (1986). A meta-analysis of gender differences in spatial ability: Implications for mathematics and science achievement. In J. Hyde & M. Linn (Eds.), *The psychology of gender: Advances through meta-analysis* (pp.67-101). Baltimore, MD: Johns Hopkins University Press.

Lohman, D. (in press). Reasoning abilities. In R. Sternberg, J. Davidson, & J. Pretz (Eds.), *Cognition and intelligence*. New York: Cambridge University Press.

Ma, L. P. (1999). *Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States.* Mahwah, NJ: Lawrence Erlbaum.

Markman, A. B., & Gentner, D. (2001). Thinking. *Annual Review of Psychology*, *52*, 223-247.

Masters, M. S., & Sanders, B. (1993). Is the gender difference in mental rotation disappearing? *Behavioral Genetics, 23*, 337-341.

Mathematical Association of America (MAA). (2003). *Guidelines for programs and departments in undergraduate mathematical sciences.* Washington, DC: Author.

McCall, W. A. (1923). *How to measure in education.* New York: The Macmillan Company.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*, 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 30*, 1012-1027.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & I. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 61-73). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Monroe, W. S., DeVoss, J. C., & Kelly, F. J. (1917). *Educational tests and measurement.* New York: Houghton Mifflin.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

National Education Association. (1938). *The purposes of education in American democracy.* Washington, DC: Author.

Newell, A. (1980). Reasoning, problem solving, and decision processes: The problem space as a fundamental category. In R. S. Nickerson (Ed.), *Attention and performance VIII* (pp. 693-718). Hillsdale, NJ: Lawrence Erlbaum.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Nickerson, R. S. (1988). On improving thinking through instruction. In E. Z. Rothkopf (Ed.), *Review of Research in Education, 15*, 3-57.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Orr, E. W. (1997). *Twice as less*. New York: W. W. Norton & Company.

Polya, G. (1957). *How to solve it* (2nd ed.). New York: Doubleday.

Polya, G. (1962). *Mathematical discovery: On understanding, learning, and teaching problem solving (Vol. 1).* Boston: Heath.

Polya, G. (1965). *Mathematical discovery: On understanding, learning, and teaching problem solving (Vol. 2).* Boston: Heath.

Presseisen, B. Z. (1986). *Critical thinking and thinking skills: State of the art definitions and practice in public schools.* Philadelphia: Research for Better Schools.

Resnick, D. P., & Resnick, L. B. (1977). The nature of literacy: An historical exploration. *Harvard Educational Review*, *47*, 370-385.

Schoenfeld, A. H. (1983). Beyond the purely cognitive: Belief systems, social cognitions, and metacognitions as driving forces in intellectual performance. *Cognitive Science, 7,* 329-363.

Schoenfeld, A. H. (1985). *Mathematical problem solving.* Orlando, FL: Academic Press.

Schoenfeld, A. H. (2002). Making mathematics work for all children: Issues of standards, testing, and equity. *Educational Researcher, 31*(1), 13-25.

Shea, D. L., Lubinski, D., & Benbow, C. P. (2001). Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year longitudinal study. *Journal of Educational Psychology, 93*, 604-614.

Sons, L. (Ed.). (1996). *Quantitative reasoning for college graduates: A complement to the standards.* Retrieved October 6, 2003, from the Mathematical Association of America Web site: http://www.maa.org/past/ql/ql_toc.html

Stigler, J. W., & Baranes, R. (1988). Culture and mathematics learning. In E. Z. Rothkopf (Ed.), *Review of Research in Education, 15*, 253-306.

Stodolsky, S. S. (1988). *The subject matters.* Chicago, IL: University of Chicago Press.

Swineford, F. (1949). *Law School Admission Test—WLS* (ETS Research Bulletin 49-12). Princeton, NJ: Educational Testing Service.

Terman, L. M. (1916). *The measurement of intelligence.* Boston, MA: Houghton Mifflin.

Thorndike, E. L., Cobb, M. V., Orleans, J. S., Symonds, P. M., Wald, E., & Woodyard, E. (1924). *The psychology of algebra.* New York: The Macmillan Company.

Thurstone, L. L. (1976). *The nature of intelligence.* Westport, CT: Greenwood Press. (Original work published in 1924)

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 29-44). Hillsdale, NJ: Lawrence Erlbaum.

Wilson, J. W., Fernandez, M. L., & Hadaway, N. (2002). *Mathematical problem solving.* Retrieved October 6, 2003, from The University of Georgia Web site: http://jwilson.coe.uga.edu/emt725/Pssyn/Pssyn.html

Yerkes, R. M. (Ed.). (1921). Psychological examining in the United States Army. *Memoirs of the National Academy of Sciences, 15.* Washington, DC: Government Printing Office.

Notes

[1] In an early report of the development of the SAT (College Entrance Examination Board, 1926), the committee in the charge of its development said:

> The term "scholastic aptitude test" has reference to examination of a type now in current use and variously called "psychological tests", "intelligence tests", "mental ability tests", "mental alertness tests", et cetera. The committee uses the term "aptitude" to distinguish such tests from tests of training in school subjects. (p. 44)

The committee struck a more modern note, however, by cautioning against overinterpretation of and overreliance on test scores:

> The present status of all efforts of man to measure or in any way estimate the worth of other men, or to evaluate the results of their nurture, or to reckon their potential possibilities does not warrant any certainty of prediction…. To place too great emphasis on test scores is as dangerous as the failure properly to evaluate any score or mark in conjunction with other measures and estimates which it supplements.

[2] Stigler and Baranes (1988) point out that Wittgenstein saw mathematics as social in nature and inseparable from the social realm in which it is used, a point since reiterated by numerous scholars in the field of sociology of mathematics.

[3] For example, Howe (1998) makes the point that the standards developed by the American Mathematical Association extend the NCTM standards by defining college-level mathematical reasoning to include the concept of creating proofs. (p. 246)

[4] "The importance of domain-specific knowledge to thinking is not really debatable. To think effectively in any domain one must know something about the domain and, in general, the more one knows, the better." (Nickerson, 1988, p. 13)

[5] Customary rules of thumb used by many ETS programs for evaluating the aspects of speededness concerned with reaching the end of a test are 1) Virtually all of the test takers complete three-quarters of the test in the allotted time and 2) At least 80% of the test takers complete the entire test. The apparent first formulation of these rules of thumb can be found in Swineford, 1949. Note that these rules of thumb do not address various strategies that individuals might elect for completing the test (Nunnally, 1978). For example, individuals who reach the end of the test by answering all remaining questions with little or no attempt to actually work the problems are included on the same basis as those who have considered the

questions fully and answered to the best of their ability. The rules of thumb also do not consider individual students' optimal pacing or other preferences that might affect scores.

[6] Guilford addressed this issue nearly 50 years ago (1959), saying

> The present status of the question of speed factors is that it is doubtful whether they are of much consequence and whether they are aspects of intellect. …certain speed tests have correlated with leadership behavior, a result that is consistent with the hypothesis that speed tests measure to some extent a motivational component. (p. 398)

[7] Gordon (1997) quotes a distinguished physicist who, when asked how he came to outshine much brighter friends, replied, "It's true. Back in school, they could do whatever I did in half the time. But now I've got the time." (p. 213)

Appendix

Sample Questions

Quantitative Comparisons

Directions: A quantitative comparison question consists of two quantities, one in Column A and one in Column B. You are to compare the two quantities and select answer:

A if the quantity in Column A is greater
B if the quantity in Column B is greater
C if the two quantities are equal
D if the relationship cannot be determined from the information given.

The average (arithmetic mean) of 10 numbers is 52. When one of the numbers is discarded, the average of the remaining numbers becomes 53.

Column A	Column B
The discarded number	51

Correct Answer: B

Explanation: Because the average of the 10 numbers is 52, the sum of the ten numbers must be 10 * 52 = 520. Because the average of the remaining 9 numbers is 53, the sum of the 9 numbers must be 9 * 53 = 477, so the discarded number must be 520 – 477 = 42, which is less than 51.

Note. From *GRE: Practicing to Take the General Test: Big Book*. (ETS, 1996, p. 83).

Standard Multiple-choice

Directions: Solve each problem. Then decide which is the best of the choices given and fill in the corresponding oval on the answer sheet.

If a and b are integers and $a - b = 6$, then $a + b$ CANNOT be

(A) 0
(B) less than 6
(C) greater than 6
(D) an even integer
(E) an odd integer

Correct Answer: E

Explanation: Suppose $a - b = 6$ and b is even integer. Then $a = 6 + b$ is an even integer, and $a + b$ must be even, since the sum of two even integers is even. If b is not an even integer, then it is an odd integer and $a = 6 + b$ is an odd integer, because the sum of an even integer (6) and an odd integer (b) must be odd. But then $a + b$ is the sum of an odd integer (a) and an odd integer (b) and the sum of two odd integers must be even. The sum of $a + b$ must be even, so it CANNOT be an odd integer, and E is the correct answer.

Note. Taken from *GRE: Practicing to Take the General Test: Big Book* (ETS, 1996, p. 833).

Student-produced Responses (Grid-ins)

Directions: Grid-ins (student-produced response questions) require you to solve the problem and enter your answer by marking the ovals in the special grid.

One out of residents of Central Village was born in that village. If its population is 12,000, what is the total number of its residents who were not born in Central Village?

(from the October 1996 test)

Correct Answer: 9600

Explanation: One of the first things you should notice about this problem is that you are asked to find the number of residents who were NOT born in Central Village. If out of residents, or 1/5 of the total population, was born in the village, this means that 4 out of 5 residents, or 4/5 of the total population, were not born in the village.

Therefore,

$$\frac{4}{5} \text{ of } 12{,}000 = \frac{4}{5} \times 12{,}000 = 9600$$

Note. Taken from *About PSAT/NMSQT* (College Board, 2003).

Data Sufficiency

Pam and Ed are in line to purchase tickets. How many people are in line?

(1) There are 20 people behind Pam and 20 people in front of Ed.

(2) There are 5 people between Pam and Ed.

(A) Statement (1) ALONE is sufficient, but statement (2) alone is not sufficient.

(B) Statement (2) ALONE is sufficient, but statement (1) alone is not sufficient.

(C) BOTH statements TOGETHER are sufficient, but NEITHER statement ALONE is sufficient.

(D) EACH statement ALONE is sufficient.

(E) Statements (1) and (2) TOGETHER are NOT sufficient.

Correct Answer: E

Explanation: There is not enough information to determine how many people are in line, even if both (1) and (2) are assumed. If Ed is in front of Pam, then by (1) and (2) Ed is the 21st (because there are 20 people in front of Ed) person in line, Pam is the 27th (since Ed is 21st and there are 5 people between them, and Ed is ahead of Pam.) So if Ed is in front of Pam, there are 47 people in line, because there are 20 people behind Pam.

Suppose, however, that Pam is in front of Ed. Ed is still the 21st person in line because there are 20 people in front of Ed. Now Pam is the 15th person in line, however, because there are 5 people between them and Pam is in front of Ed the last person in line would be the 35th person, since there are 20 people behind Pam. So if Pam is in front of Ed, there are 35 people in line.

Note. From *The Official Guide for GMAT Review* (Graduate Management Admissions Council [GMAC], 1992, p.165).