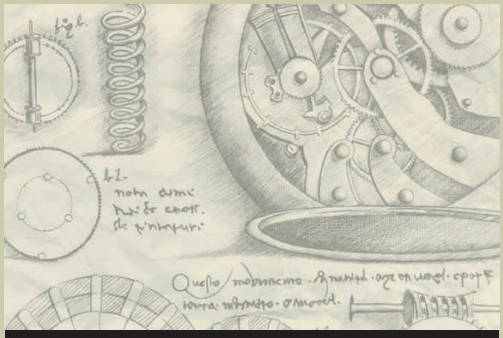


InView™

Technical Bulletin

CTB McGraw-Hill

54166

The McGraw-Hill Companies

NOTICE TO THOSE WISHING TO SCAN, SCORE, REPORT, AND ANALYZE THIS INSTRUMENT

Any school, agency, commercial entity or person wishing to scan, score, report, and/or analyze any CTB/McGraw-Hill test instrument, by machine or otherwise, is required to obtain a written license from the publisher by applying at the following address:

Contract Services
CTB/McGraw-Hill
20 Ryan Ranch Road
Monterey, California 93940-5703
Telephone (831) 393-7839

The policy includes processing carried out on a for-profit or non-profit basis, and applies whether or not the processing is limited to a school district's own students.

This notice supersedes all notices with a similar title published in earlier editions of CTB/McGraw-Hill test instruments.

Individual student data is received by CTB/McGraw-Hill with the understanding that it is CTB/McGraw-Hill's policy to retain scored data in electronic form for approximately three years. Such data is available only to the agency which ordered the scoring services, and no data or summary information will be released to any other party without the express permission of said party. CTB/McGraw-Hill may from time to time access and use retained data without individual or institutional identification for research purposes.

Published by CTB/McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc., 20 Ryan Ranch Road, Monterey, California 93940-5703.

FOREWORD

The *InView*™ Technical Bulletin is one of a series of technical publications for the *InView* assessment system. Subsequent publications will be produced as the data become available. The technical information in this bulletin is intended for those who evaluate tests, interpret scores, or use test results in making educational decisions. It is assumed that the reader has technical knowledge of test construction and measurement procedures, as described in Standard 11.1 of *Standards for Educational and Psychological Testing* (American Educational Research Association, 1999).

Table of Contents

TABLES

Part 1: Description

Overview

InView™, a cognitive abilities test, provides highly reliable cognitive ability and anticipated achievement information that reflects current research. *InView* is composed of the following five subtests: Sequences, Analogies, Quantitative Reasoning, Verbal Reasoning—Words, and Verbal Reasoning—Context. As with other CTB/McGraw-Hill cognitive abilities tests, such as the *Test of Cognitive Skills, Second Edition* (TCS/2), *InView* assesses the cognitive skills and abilities of students in Grades 2 through 12.

Most cognitive abilities cannot be measured directly but are inferred by assessing behaviors that reflect those abilities. The *InView* subtests are intended to sample many of the cognitive abilities that are reflected in academic performance and are important for success in educational programs. Items have been developed from a variety of important ability domains, such as understanding verbal concepts and analyzing and comprehending relationships among quantitative concepts. Each subtest presents students with novel problems to solve.

InView has six levels that correspond to the grades shown below:

Level 1: Grades 2–3	Level 3: Grades 6–7	Level 5: Grades 10–11
Level 2: Grades 4–5	Level 4: Grades 8–9	Level 6: Grades 11–12

Description of Subtests

The five *InView* subtests are described below.

Sequences

The Sequences subtest is a nonverbal measure of students' ability to comprehend a rule or principle implicit in a pattern or in a sequence of figures, letters, or numbers. Students must analyze the pattern presented and then select the answer choice that would continue or complete the pattern. Since students must distinguish between those elements that continue the pattern and those that do not, Sequences items measure the ability to make inferences. Items involve recognition of spatial relationships, ordered patterns, progressions, and combinations of parts to form a whole.

At Level 1, the items comprise patterns and sequences of figures. At Levels 2 through 6, the items consist of figures, letters, and numbers.

Analogies

The Analogies subtest is a nonverbal measure of students' ability to discern various types of relationships among picture pairs and then infer parallel relationships between incomplete picture pairs. The pictures comprise scenes, people, animals, objects, and abstract graphic symbols. In each item, students must recognize the nature of the relationship between two pictures and then, given a third picture, find an answer choice that will produce a relationship parallel to that of the first two pictures.

Quantitative Reasoning

The Quantitative Reasoning subtest is intended to measure students' ability to think with numbers. This test also measures students' ability to solve problems through complex quantitative reasoning processes and systematic logic, including attribute analysis and deductive and inductive reasoning.

Since both verbal and quantitative skills are related to success in school, Quantitative Reasoning receives the same prominence as Verbal Reasoning. In the Quantitative Reasoning subtest, innovative item formats focus assessment on critical quantitative processes, rather than learned mathematics skills. Students are presented with several item formats—they must identify arithmetic patterns; model complex concepts and relationships; classify according to common attributes; infer relationships among quantitative data, concepts, and processes;

apply deductive and inductive mathematical reasoning; and draw logical conclusions based on quantitative data. The Quantitative Reasoning subtest provides precise measures of students' skills in quantitative reasoning.

Verbal Reasoning—Words

The Verbal Reasoning—Words subtest is a measure of students' ability to solve verbal problems by reasoning deductively, analyzing category attributes, and discerning relationships and patterns. This measure of verbal ability contains several item formats, such as items that require students to identify essential elements of objects or concepts, classify according to common attributes, and infer relationships between separate but related sets of words. This subtest provides advanced measures of complex verbal reasoning processes, logic, and applied thought.

Verbal Reasoning—Context

The Verbal Reasoning—Context subtest is a measure of students' ability to solve verbal problems by reasoning deductively and inductively. These verbal problems require the student to identify essential elements of concepts presented in short passages and draw logical conclusions. This measure of verbal ability contains a single item format.

Item Formats

The five subtests in *InView* contain a variety of item formats that provide strong indicators of the cognitive skills being assessed. Table 1 presents the item formats that make up each subtest.

Table 1

***InView* Item Formats**

Item Types	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Sequences						
Figural	■	■	■	■	■	■
Geometric	■	■	■	■	■	■
Mixed	■	■	■	■	■	■
Numeric		■	■	■	■	■
Alphabetic		■	■	■	■	■
Analogies						
Action	■	■	■	■	■	■
Common Element	■	■	■	■	■	■
Opposites	■	■	■	■	■	■
Part/Whole	■	■	■	■	■	■
Phrase-Mediated	■	■	■	■	■	■
Visual Rule	■	■	■	■	■	■
Quantitative Reasoning						
Number Operations Puzzle	■					
Grid Comparison	■		■			
Algebraic Substitution—Equations	■			■		■
Grid Inference		■				
Number Equivalence		■				
Substitution		■				
Number Operations Flowchart			■	■		■
Grid Fraction				■	■	■
Function			■		■	
Algebraic Substitution—Balances					■	
Verbal Reasoning—Words						
Necessary Part	■	■	■	■	■	■
Category Inclusion	■	■	■	■	■	■
Category Exclusion	■	■	■	■	■	■
Analogous Sets	■	■	■	■	■	■
Verbal Reasoning—Context						
Deductive Reasoning	■	■	■	■	■	■
Inductive Reasoning	■	■	■	■	■	■

Time Requirements

Table 2 shows the number of items and time limits for administering *InView*. Item tryout studies were conducted to determine typical time limits, that is, the expected time that an average class takes to complete each section. From these studies, the specified time limits provide adequate testing time to enable most students to complete the test sections. Please note that the times shown for Level 1 are intended as estimates to allow for the variations that occur when items are read aloud. Of course, examiners may stop the test session as soon as all students are finished with the section; they may then administer the next test section.

Table 2

Number of Items and Time Limits for *InView*

	Level 1		Levels 2–6	
Test Section	**No. of Items**	**Time Limit**	**No. of Items**	**Time Limit**
Practice Test (Total)	**17**	**20* mins.**	**28**	**30 mins.**
Sequences	20	15 mins.	20	15 mins.
Analogies	20	20 mins.	20	10 mins.
Quantitative Reasoning	20	30 mins.	20	30 mins.
Verbal Reasoning—Words	20	20 mins.	20	20 mins.
Verbal Reasoning—Context	20	20 mins.	20	20 mins.
Total Test	**100**	**105 mins.**	**100**	**95 mins.**

Note: Time limits are for actual testing time and do not include instruction time for administration of sample items.
* This section is read by the examiner; therefore, times shown are approximate.

Highlights of *InView*

InView represents CTB's latest edition of cognitive skills assessment. As with TCS/2 (CTB's previous version of the cognitive skills test), *InView* uses selected-response items to measure students' cognitive abilities and thinking skills in both the nonverbal and verbal domains.

InView *Subtests*

Compared to its predecessor, TCS/2, *InView* incorporates several changes to the number and type of subtests (see below).

Subtests

InView	**TCS/2**
Sequences	Sequences
Analogies	Analogies
Quantitative Reasoning	Memory
Verbal Reasoning—Words	Verbal Reasoning
Verbal Reasoning—Context	

- A new Quantitative Reasoning subtest has been added. The Quantitative Reasoning subtest measures students' ability to think with numbers and quantitative processes and includes item formats that are unlike most mathematics achievement test formats.

- The Verbal Reasoning subtest was separated into Verbal Reasoning—Words and Verbal Reasoning—Context. The Verbal Reasoning—Words subtest measures students' ability to solve verbal problems using critical thinking and logic. The Verbal Reasoning—Context subtest measures students' ability to solve verbal problems using deductive and inductive reasoning.

Additional Highlights of InView:

- An *InView* Online component with an intuitive, user-friendly interface has been added. *InView* Online provides immediate feedback based on test performance. Technical information regarding *InView* Online will be available at a later date.
- *InView* was co-normed with *TerraNova, The Second Edition,* in 2000. The updated norms for *InView* are inclusive, all students who participated in regular testing under accommodations were included in the standardization sample. For further information, refer to Part 3 and Part 4.
- *InView* score report formats have been updated to report information in a way that is clear, informative, and easy to understand. Data are presented in user-friendly score reports that can be presented to a wide range of audiences including parents, students, and administrators.
- *InView* can be administered as a stand-alone cognitive skills test or in conjunction with *TerraNova, The Second Edition*. When administered with *TerraNova, The Second Edition,* anticipated achievement scores are available. Anticipated achievement scores compare students with others of similar age, grade, and cognitive ability. Anticipated achievement scores provide useful information to teachers in order to evaluate whether students are performing above or below expectations.

Test Components and Ancillary Materials

A full range of testing components and ancillary materials has been developed for *InView*, including test books, answer documents, examiner's manuals, practice tests, a teacher's guide, and a norms book. Please see the *InView Teacher's Guide* for a description of these products.

Part 2: Content and Test Construction

Development of the Test

Item Writing

Content blueprints provided guidelines for writers to develop items for *InView*. To provide a large pool of items for final test selection, twice the number of items necessary for operational forms were developed for four of the five subtests. Four times the number of items necessary were developed for Quantitative Reasoning. A staff of professional item writers, all with previous teaching experience, researched, selected, and wrote items following the specifications listed in the *InView* content blueprint. All new items were written for the Verbal Reasoning—Context and Quantitative Reasoning subtests. The writing process included considerable analysis of the reasoning inherent in the stimuli. Initial development of each stem and set of corresponding distractors was followed by a rigorous cycle of internal review.

All test items and test directions were carefully reviewed for content and editorial accuracy. Close adherence to vocabulary specifications ensured that all words used were appropriate for the target grades at each test level.

Bias Reviews

During the development of *InView*, test editors gave careful attention to issues of ethnic, racial, gender, and age bias. During all stages of development, materials were reviewed to conform to the editorial policies and guidelines of CTB/McGraw-Hill, as defined in *Guidelines for Bias-Free Publishing* (McGraw-Hill, 1983). Additionally, all tryout materials were reviewed by educators from various ethnic groups. CTB/McGraw-Hill conducts extensive research in an attempt to identify and minimize racial and ethnic bias in its products. Items that were identified through bias reviews as being potentially biased were eliminated or revised appropriately.

Item Tryout Studies

Typically, not all new items are appropriate for use in final test books. To gather empirical data on the newly written items, tryout studies were conducted in schools throughout the country during the 1999 school year. Items for the Sequences, Analogies, Verbal Reasoning—Words, and Verbal Reasoning—Context subtests were administered in the spring of 1999. Because Quantitative Reasoning is a new subtest, several newly developed item formats were first locally piloted in classrooms. Based on the results of the spring 1999 pilot study, the number of Quantitative Reasoning item formats was reduced and items in the final formats were administered in the fall of 1999.

New items were organized into tryout books for each level and subtest. In addition to the new items being tried out, anchor items were included in each tryout book. These anchor items were selected from the corresponding test and level of TCS/2 for subtests of Sequences, Analogies, Verbal Reasoning—Words, and Verbal Reasoning—Context. Anchor items for Quantitative Reasoning were selected from corresponding levels of *TerraNova* Mathematics. By using the anchor items to link the tryout items to a normed scale and making use of the normative scale score distributions, estimates of national performance on the new items were obtained.

In addition to the statistical data acquired from the item tryout, valuable information was also received from the teachers who administered the tryout edition of *InView*. All examiners were asked to review and provide comments concerning the content, illustrations, instructions, and time limits for the materials they administered. The comments and suggestions from the teachers resulted in numerous improvements in the final tests.

Tryout Data Collection and Analysis

Samples

Each tryout study included a heterogeneous sample of students enrolled in public and private schools across the country. Items were administered at several grades. For example, items thought to be appropriate for Grades 4 and 5 were tried out at Grades 3, 4, 5, 6, and 7. Special samples of African-American and Hispanic-American students were obtained from schools with predominantly African-American or Hispanic-American enrollments. These samples were assigned a lower *InView* test level relative to the standard sample. For example, items tried out at Grade 3 for the standard sample were tried out at Grade 4 for the African-American and Hispanic-American samples. This shift in the samples was made so that meaningful responses to the items could be obtained from the minority groups that historically have obtained lower scores than the standard group. The test design specified a sample size of at least 600 students in the standard sample and 200 students in each of the African-American and Hispanic-American samples for each level and content area. More than 17,000 students were involved in the *InView* tryout.

Answer Choice Information

Statistical information about student performance on each item was produced by grade. For each item, three answer choice statistics were examined: (1) the proportion of students choosing each answer, (2) the point-biserial correlation between the answer choice and the number-correct score on the rest of the test, and (3) the mean number-correct score of students choosing that answer. Following professional measurement principles, statistical data derived from the answer choice information was used by content editors to aid in selecting items that functioned well psychometrically. For example, content editors avoided using items that had a very low point-biserial correlation on the correct answer but relatively high point-biserial correlation for the distractors.

Item Response Theory Models

InView tryout items were scaled and calibrated using item response theory (IRT) procedures similar to those followed in the development of TCS/2.

The three-parameter logistic (3PL) model was used to scale the dichotomously scored items from *InView* (Lord & Novick, 1968; Lord, 1980). The 3PL model defines a selected-response item in terms of three item parameters: the item difficulty or location, the item discrimination, and the guessing parameter. In the 3PL model, the probability that a student with scale score θ responds correctly to item i is

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp\left[-1.7a_i\left(\theta - b_i\right)\right]},$$

where a_i is the item discrimination, b_i is the item difficulty, and c_i is the probability of a correct response by a very low-scoring student. Further discussions of IRT can be found in van der Linden and Hambleton (1996).

Parameter Estimation

Calibration was implemented using PARMATE (Burket, 1990). The Stocking and Lord (1983) procedure was used to place the parameter estimates on the scale from which the anchor items were drawn. The *InView* subtests, Sequences, Analogies, Verbal Reasoning—Words, and Verbal Reasoning—Context, were scaled to the corresponding TCS/2 scales, while Quantitative Reasoning was scaled to the *TerraNova* Mathematics scale. Extensive quality control procedures were used to ensure that calibration results were accurate and reasonable. Examples include checking data integrity, fit statistics, and the quality of calibration and equating results.

Estimated National Difficulties

The proportions of students in the national norm group who would have correctly answered each tryout item were estimated. This estimate made use of the item parameters and the national normative distribution of the scale scores from which the anchor items were chosen. Let θ_j be the scale score at the j-th percentile of the national normative distribution at the fall or spring of the grade of interest, where $j = 1 \ldots J$. The estimated national difficulty for item i is

$$P_i = \frac{1}{J}\sum_{j=1}^{J} P_i\left(\theta_j\right), \text{ where } J = 99.$$

Information on estimated national difficulties is used to select items with difficulties matched to the level of student performance for each grade.

Differential Item Functioning

During the tryout of *InView*, separate data were gathered for students from different ethnic backgrounds. Differential item functioning (DIF) was evaluated using the Linn and Harnisch (1981) procedure implemented by PARMATE (Burket, 1990). These analyses resulted in each item being given an ethnic DIF rating from 1 (no DIF) to 5 (DIF against African-Americans and Hispanic-Americans), and a gender DIF rating of 1 (no DIF) to 4 (DIF in favor of one group and against another).

For the ethnic DIF analyses, the African-American group included all students who identified themselves as such; similarly, the Hispanic-American group included students who identified themselves as Hispanic-American. DIF results were used in item selection in order to minimize the number of DIF items in the final tests.

For a more detailed discussion of the DIF analyses of the *InView* tryout study, refer to Part 7, "Test Fairness and Differential Item Functioning."

Discrimination Rating

The item discrimination parameter obtained from the IRT parameter estimation mentioned above indicates how effectively an item distinguishes between examinees of different abilities on the trait being measured. An item that does not discriminate well provides little information regarding an examinee's level of ability.

The discriminations of the *InView* tryout items were compared to the item discriminations of the complete TCS/2 or *TerraNova* test from which the tryout anchor items were obtained. Discrimination ratings were assigned to the *InView* tryout items using the mean and standard deviation of the discriminations of the TCS/2 or *TerraNova* test for each level. A rating of *good* was assigned if the tryout item discrimination was equal to or greater than the mean discrimination of the TCS/2 or *TerraNova* test. Tryout items with discriminations less than the TCS/2 or *TerraNova* mean discrimination minus the TCS/2 or *TerraNova* standard deviation were assigned *poor* ratings. All other items were assigned a rating of *fair.*

Fit Rating

Item fit is a statistic that is used to evaluate whether an item is performing as is expected given the model assumptions. A procedure described by Yen (1981) was used to measure how well *InView* tryout items fit the 3PL model. In this procedure, students are rank-ordered on the basis of $\hat{\theta}$ values and sorted into ten cells with ten percent of the sample in each cell. For each item, the number of students in cell k who answered item i, N_{ik}, and the number in that cell who answered correctly, R_{ik}, were determined. The observed proportion in cell k passing item i, O_{ik}, is R_{ik} / N_{ik}. The fit index (Q_1) for item i is

$$Q_{1i} = \sum_{k=1}^{10} \frac{N_{ik}\left(O_{ik} - E_{ik}\right)^2}{E_{ik}\left(1 - E_{ik}\right)},$$

where

$$E_{ik} = \frac{1}{N_{ik}} \sum_{j \varepsilon \text{ cell } k}^{N_{ik}} P_i\left(\theta_j\right).$$

Items with $Q_{li} \geq 14.1$ (the chi-square value with 7 degrees of freedom at the .05 significance level) are given a fit rating of *poor;* other items have a fit rating of *acceptable*. Developers avoid selecting items with poor fit ratings.

Item Selection Criteria

Item selection for *InView* was conducted by content experts and reviewed by the psychometric staff. Some of the *InView* subtests have varying item formats. Within the limits set by the format requirements, content editors selected items with the best statistical characteristics, including appropriate distractor performance, adequate discrimination, and acceptable fit. Editors also attempted to avoid selecting items that showed ethnic or gender DIF and to produce final tests that minimize DIF.

Content editors also picked items that minimized measurement error throughout the range of achievement expected for each test level, called "nominal range." The nominal range for each test level was the range between the scale score at the 5th percentile of the fall anchor test distribution for the lower target grade to the 95th percentile of the spring distribution for the higher target grade. For example, the nominal range for Level 2 Analogies spanned from the 5th percentile of the fall 1996 TCS/2 distribution for Grade 4 to the 95th percentile of the spring 1996 TCS/2 distribution for Grade 5. Measurement error was estimated using the reciprocal of the square root of the IRT information function (Lord, 1980, p. 71).

This criterion of minimizing measurement error throughout the expected range of performance resulted in items being chosen that have a range of difficulty appropriate for the target grades. This criterion also helps increase the likelihood that the resulting test does not exhibit large proportions of extreme scores, that is, floor or ceiling effects.

Item Selection Process

Item selection was facilitated by the ITEMSYS software (Burket, 1988). ITEMSYS allows the content editor to make informed decisions regarding an item selection. This software monitors the impact of each decision made during the item selection process and offers a variety of options for grouping, classifying, sorting, and ranking items to highlight key information as it is needed. (See also Green, Yen, & Burket, 1989.)

The ITEMSYS program has three parts. The first part is used to select a working item pool of manageable size from the larger tryout pool; items clearly inappropriate to the target grade range are eliminated. There is information about each item in the pool, including the item format to which the item is assigned, a descriptive phrase about the item, the association of the item with a stimulus, a bias rating indicating whether the item shows DIF to a particular population of students, the item parameters, and a fit rating indicating how well the item fits the expectations based on the IRT model used.

The second part of the ITEMSYS program uses the working item pool created in the first step to perform the actual test selection. Typically, the developer begins by specifying the number of items to be included in the test and a target number of items for each item format. The program can then be prompted to select automatically a test that represents the best possible statistical combination of items. These automatic selections can then be used as a reference set to which other selections are compared. Successive selections are plotted on a graphic display that shows the standard error of measurement for each set of selected items across the nominal range for the target grade. This standard error of measurement curve becomes the visual cue to adjust the selection. While maintaining the format requirements when applicable, the developer simultaneously attempts to minimize error. Lowering the error

curve results in a test with less measurement error. Smoothing bumps out of the curve creates a test that measures more consistently across the target range of difficulty.

At any time during the selection process, the developer can call up one of many information screens that rank items according to criteria such as difficulty, ability to discriminate between low- and high-achieving students, and contribution to lowering the standard error curve at one particular point in the range. Such screens help the developer pinpoint the exact item to add to the selection or indicate when an optimum choice has already been made.

In the third part of the program, a table shows both expected number correct and standard error of measurement as functions of scale score, as well as statistical and graphic summaries on bias, fit, and the average standard error of the test as selected. Any fault in the selection—whether the test is too easy or too difficult for the target grade, contains biased items, or does not adequately cover part of the range—becomes immediately apparent as the final statistics are generated. A content editor detecting any such problems can then return to the second stage of the program and revise the selection. The flexibility and utility of the program encourages multiple attempts at fine-tuning the selection.

Part 3: Validity

"Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed users of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests" (AERA, APA, & NCME, 1999). The purpose of test score validation is not to validate the test itself, but to validate interpretations of the test scores for particular purposes or uses. Test score validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the entire assessment process. Every aspect of an assessment provides evidence in support of its validity (or evidence to the contrary), including design, content specifications, item development, psychometric quality, and inferences made from the results. *InView* was designed and developed to provide fair and accurate cognitive ability scores that support appropriate, meaningful, and useful educational decisions.

Validation begins with the proposed scope and extent of the test. *InView* measures aspects of cognitive ability that have been found to be related to success in school. These include nonverbal abilities (Sequences, Analogies, and Quantitative Reasoning) and verbal abilities (Verbal Reasoning—Words and Verbal Reasoning—Context).

Content Validity

The evolution of *InView* reflects a rich history of scientific research in the assessment of reasoning skills. *InView* is the product of a long evolution which began in 1936 with the publication of CTB's original instrument for measuring cognitive abilities. Through the years, as new perceptions about cognitive skills emerged and new research findings were published, CTB/McGraw-Hill's cognitive tests have been revised and updated to reflect the advances and educational needs of the time.

InView distinguishes itself from other cognitive skills tests in its emphasis on reasoning abilities that are important for success in educational programs. *InView* is primarily intended to be used in educational settings and includes items that measure nonverbal and verbal reasoning skills that typically are not explicitly taught as part of formal subject-matter curricula.

A number of other issues were also carefully considered and addressed to increase content validity. For example, item formats from CTB/McGraw-Hill's previous cognitive skills tests were reconceptualized, and new formats were developed for *InView* with the focus on measuring cognitive abilities important to scholastic success. Studies were conducted so that the graphic design of any new item formats would have no negative effect on performance. New formats and items in Quantitative Reasoning were piloted several times. In the development of *InView,* careful consideration was given to the fact the students come from diverse linguistic and ethnic backgrounds. Finally, teacher and student questionnaires provided information regarding overall test effectiveness (i.e., speededness, clarity, and appropriateness for the target grade).

Nonverbal Ability

InView measures nonverbal cognitive ability with three subtests: Sequences, Analogies, and Quantitative Reasoning.

- The Sequences subtest presents nonverbal sequences of graphic symbols and alpha-numeric combinations that require students to complete the sequence. This subtest is a measure of sequential processing in a visual closure context.
- The Analogies subtest presents nonverbal pictorial analogies using symbols. Students are presented with three pictures. Students must first discern the relationship between the first two pictures. Students must then select a picture from among the answer choices that forms an analogy with the third picture that is parallel to the relationship between the first two pictures.
- The Quantitative Reasoning subtest is a measure of students' ability to think with numbers. Unlike most mathematics achievement tests, this subtest presents a combination of item formats that require quantitative reasoning abilities rather than learned skills.

Verbal Ability

InView measures verbal ability with two subtests: Verbal Reasoning—Words and Verbal Reasoning—Context.

- The Verbal Reasoning—Words subtest presents a combination of verbal items that require deductive and analytical reasoning.
- The Verbal Reasoning—Context subtest requires the student to read short passages and draw conclusions. This subtest is a measure of verbal knowledge, an aspect of cognitive ability.

Construct Validity

Construct validity—what test scores mean and what kinds of inferences they support—is a central concept underlying the *InView* test validation process. Evidence for construct validity is comprehensive and integrates evidence from many sources. For example, to demonstrate comprehensiveness, *InView* tests must contain items that assess a variety of cognitive skills that pertain to education. To establish meaningfulness, the test must correlate highly with independent measures of achievement and cognitive ability. Furthermore, the patterns of correlations among the *InView* subtests and with other assessments should demonstrate convergent and discriminant validity. That is, subtests designed to measure similar cognitive reasoning abilities should correlate more highly than subtests designed to measure different cognitive reasoning abilities. The correlations presented in this bulletin, which include correlations between *InView* subtests and correlations between *InView* and *TerraNova, The Second Edition* (CTB/McGraw-Hill, 2001) demonstrate convergent and discriminant validity.

Intercorrelations

Table 3 presents intercorrelations of scores between *InView* subtests as well as correlations of scores between *InView* and *TerraNova, The Second Edition* for Grades 2 through 12. The data in this table are derived from the samples of students who participated in the spring 2000 standardization and completed both *InView* and *TerraNova, The Second Edition.* See Part 4 for more details regarding the standardization samples.

In general, the correlations presented in Table 3 are consistent with expectations for convergent and discriminant validity. Among the five *InView* subtests, for example, Verbal Reasoning—Words consistently correlates more highly with Verbal Reasoning—Context than with Quantitative Reasoning. As a further example, the data show that Sequences typically correlates more highly with Analogies and Quantitative Reasoning than with either Verbal Reasoning—Words or Verbal Reasoning—Context.

The correlations between Verbal Reasoning—Words and Quantitative Reasoning may, in some cases, be higher than anticipated. In spite of the verbal/nonverbal distinction of the subtests, all five subtests of *InView* purport to measure reasoning abilities, and moderate correlations can be expected between any pair of the *InView* subtests. The similarities between reasoning abilities measured by different *InView* subtests are analogous to the similarities between two achievement content areas that are deemed to be the most dissimilar in instructional content, such as Reading and Mathematics Computation. The correlations between Reading and Mathematics Computation typically between .40 and .60 are not dissimilar to the correlations between Verbal Reasoning—Words and Quantitative Reasoning.

The convergent and discriminant validity of *InView* scores can also be evaluated by examining relationships of *InView* subtests with independent measures, in this case *TerraNova, The Second Edition*. For example, the *InView* Nonverbal Total scores typically correlate more highly with the Mathematics scores than with the Reading and Language scores on *TerraNova, The Second Edition*.

Factor Analysis

Factor analysis was conducted in order to confirm the dimensionality of the *InView* subtests.

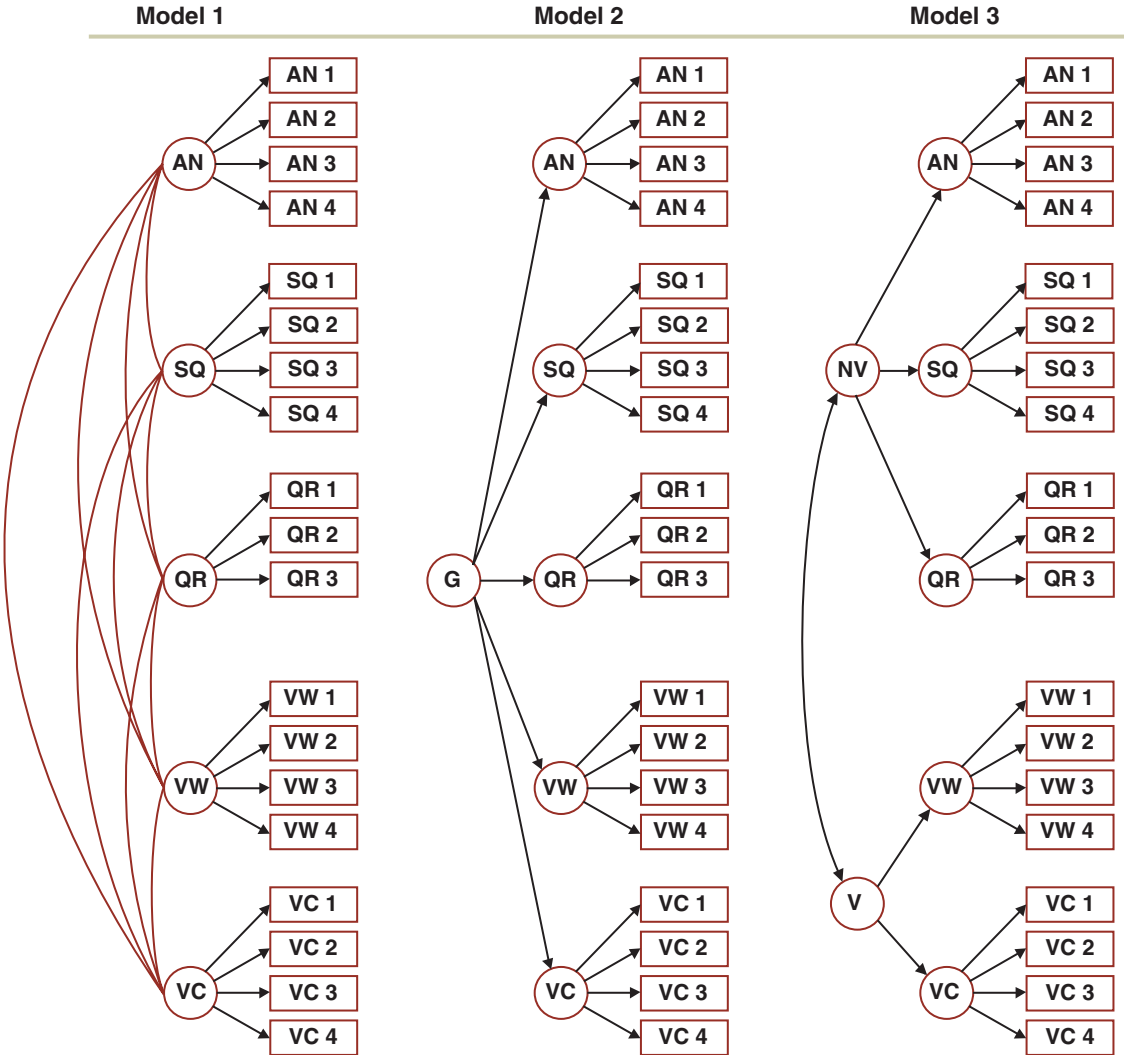
InView draws its conceptual framework for assessing cognitive abilities important in education from current theories of cognitive skills and abilities. For example, both the Gf-Gc models (e.g., Cattell, 1943; Horn, 1968; Horn & Noll, 1997; Carroll, 1993) and the single essence (*g*) factor model (Spearman, 1927) provide conceptual support. Cattell described fluid intelligence (Gf) as fundamental abilities in reasoning and higher mental processes, and crystallized

intelligence (Gc) as abilities learned by an individual from education and experience. With *InView*, it is reasonable to suppose that its nonverbal subtests measure Gf-related abilities, while the verbal subtests that involve acquired language skills measure Gc-related abilities. Additionally, all these abilities are subsumed under a general cognitive ability, *g*, although Spearman's *g* includes numerous other cognitive abilities than the ones assessed by *InView*.

Models Tested

A confirmatory factor analysis was conducted to test the existence of a single, general factor and two factors (nonverbal and verbal) underlying the five *InView* subtests. Figure 1 shows the three models that were tested in the analysis. Model 1 is the "baseline" model that relates observed variables to the five latent constructs (i.e., traits of Sequences, Analogies, Quantitative Reasoning, Verbal Reasoning—Words, and Verbal Reasoning—Context) without assuming any higher-order factors. Model 2, a second-order factor model, was used to test the hypothesis that there is a general cognitive ability (*g*) factor which underlies the five first-order factors. Model 3 is also a second-order model and was used to test the hypothesis that there are two unobservable variables (nonverbal and verbal) underlying the five subtests. The two second-order factor models were compared with the baseline model with regard to model fit. See the Appendix for further explanation of a second-order factor model.

Figure 1
Higher-order Factor Structures for Models 1, 2, and 3



AN = Analogies, SQ = Sequences, QR = Quantitative Reasoning,
VW = Verbal Reasoning–Words, VC = Verbal Reasoning–Context

Note that in Figure 1, regardless of the model, there are only three or four observed variables per subtest although each subtest contains 20 items. This is because each observed variable is a "bundle" of items (for increased score reliabilities). The items were grouped into bundles by item format for Sequences, Quantitative Reasoning, and Verbal Reasoning—Words. Sequences has two item formats for the Level 1 test and four item formats for all other test levels. The Verbal Reasoning—Words tests have four item formats; the Quantitative Reasoning tests have three item formats. Analogies and Verbal Reasoning—Context each have a single item format so items were bundled randomly. Random bundles were created for Analogies and Verbal Reasoning—Context by assigning a random number between 0 and 1 to each item and classifying the items into four groups based on the random numbers (0–.25, .26–.50, .51–.75, .76–1.0). The numbers of items in the bundles are given in Table 4 by subtest.

Data and Analyses

As noted in Part 4, *InView* was nationally standardized in the spring of 2000. For the confirmatory factor analysis presented here, approximately 2,500–3,500 cases were randomly chosen from the standardization data. The sample sizes used in the analysis are summarized by grade in Table 5.

The analyses were performed within the linear structural equation framework developed by Joreskog and Sorbom (Joreskog, 1969; Joreskog & Sorbom, 1983). The first- and second-order factor models were evaluated using EQS 5.7 (Bentler, 1998). EQS is a statistical tool for analyzing covariance matrices according to systems of structural equations.

The fit indices chosen for the evaluations were chi-square (χ^2), target coefficient (T), goodness of fit (GFI), adjusted goodness of fit (AGFI), comparative fit index (CFI), and root-mean-square error of approximation values (RMSEA; See the Appendix for more details regarding fit indices). The often used criterion of good model fit for the GFI, AGFI, and CFI measures is .90 or greater, and less than .05 for the RMSEA index. The target coefficient T has an upper limit of 1 (Schumaker & Lomax, 1996).

Results

The results from the analyses are presented in Table 6. All the χ^2s are significant at alpha = .01, but this is most likely because the sample sizes are relatively large (Marsh, Balla, & McDonald, 1988). Therefore, it would be wrong to conclude on the basis of the χ^2 results that none of the models fit the data well. According to the additional indices of fit, all three models fit the data well, indicating that the data support any of the models. When Models 2 and 3 are compared, the results in terms of the GFI, AGFI, and CFI fit indices are very comparable between the two models. However, the T ratios and the RMSEAs suggest that Model 3 with two second-order factors (Verbal and Nonverbal) fits the data better than Model 2 with one general second-order factor (G).

Conclusion

The results from the confirmatory factor analysis did not reject the postulation of a single, general trait and two traits (nonverbal and verbal) for the five *InView* subtests. Based on the results, the following three aggregate scores are reported in addition to the five subtest scores:

- Total Score—the average of the five individual subtest scores
- Nonverbal Total—the average of Sequences, Analogies, and Quantitative Reasoning
- Verbal Total—the average of Verbal Reasoning—Words and Verbal Reasoning—Context

Assessment Accommodations

CTB/McGraw-Hill understands the need to include all students in large-scale testing programs as part of the educational process. Federal, state, and local regulations or policies often stipulate that students be provided with appropriate accommodations during testing. The characteristics of the tested population of large-scale state and district testing programs are expected to change in response to the movement toward inclusiveness. Test publishers and those who desire to make valid interpretations of test results must consider how these changes affect the concept and practice of standardized assessment.

A General Framework for Reconciling Standardization and Accommodation in Support of Inclusive Testing Practices

Standardization is a fundamentally important characteristic of educational assessments that are designed to support comparisons among participants. Historically, the definition of standardization in educational assessment has focused on compliance with uniform administration conditions. Requiring uniform administration conditions has resulted in the systematic exclusion of students for whom those conditions are not appropriate. The purposes of modern educational assessment now extend beyond student-to-student comparisons under uniform conditions, and the inclusion of all students in educational assessment has become highly valued and required by law. Furthermore, the interpretations to be made from assessment results have broadened, and the validity of these interpretations may be compromised by a requirement of uniform conditions. A re-conceptualization of the principle of standardization is required to support the valid interpretation of results from inclusive test administrations.

CTB advocates and has adopted an approach to standardization that recognizes inclusiveness and accommodations as equally important, non-conflicting, characteristics of modern assessment practices. The approach consists of four main principles. First, publishers of standardized tests should clearly define default conditions under which tests are to be administered, and these conditions should allow for broad participation by the vast majority of students. This principle discourages default conditions that are needlessly restrictive (e.g., short time limits, very small print), but also recognizes the importance of defining default conditions rather than leaving them to the discretion of users.

Second, changes to publisher default conditions will be necessary for some students to meaningfully participate in the assessment. Decisions regarding the use of such accommodations should be made by appropriately trained individuals familiar with the students' disabilities. These decisions should be documented in writing, such as in an Individualized Education Program (IEP), a 504 Plan, or accommodation plans specific to LEP/ELL students. Under this principle it is generally not appropriate for test publishers or policy makers to decree some accommodations as universally valid or invalid. Policy makers may exercise judgment regarding the treatment of scores arising from various accommodations.

The third principle is to define as standard the inclusive administration of assessments to all students who may meaningfully participate under either default or IEP-specified conditions. As a result, students with documented needs for accommodations will participate in assessments under the accommodated conditions they experience in daily instruction. This inclusive definition of standardization guides the creation, publication, and norming of standardized tests at CTB/McGraw-Hill, including *InView.* CTB/McGraw-Hill believes that it allows for a more meaningful set of interpretations from the results.

Finally, interpretation of the results of inclusive administrations requires careful consideration of the targeted skills being measured, the nature and frequency of accommodations used, and the likely impact of the accommodations on reported performance.

Appropriate Interpretation of Test Results from Inclusive Administration

This section discusses appropriate interpretations of test results when some students take the test under accommodated conditions. It focuses on the validity of inferences to be made from the results of a test. When test administration conditions vary from the default conditions specified by the test publisher, the interpretation of test scores, both criterion- and norm-referenced, should take into account the actual administration conditions.

Criterion-referenced interpretations of test scores may be supported. They represent a fixed level of achievement that can be interpreted in terms of what students know and are able to do at a given score—raw score, scale score, or performance level defined by a range of scale scores. For example, a student who achieves a performance level designated as "Proficient" on a second-grade mathematics test may demonstrate the following knowledge, skills, and abilities:

Proficient students: know place value to the hundreds; count by twos; measure to nearest inch and centimeter; measure lengths in nonstandard units; read time to the half-hour; describe and classify common shapes; have three-dimensional spatial sense; combine pattern blocks to form given shapes; complete tally charts and bar graphs; use data to solve problems; extend numerical and geometric patterns.

When a student achieves the "Proficient" performance level with the accommodation "extra time," for example, the testing conditions should be considered along with the knowledge and skills ascribed to the student. In this case, the interpretation may be that, given the particular raw score, scale score, or performance level, the student can demonstrate the knowledge, skills, and abilities cited above, with the accommodation "extra time."

Norm-referenced interpretations test scores may also be supported. National percentile rank (NP), normal curve equivalent (NCE), and grade equivalent (GE) scores are examples of norm-referenced scores. Norm-referenced scores are interpreted in terms of a student's performance compared with the performance of a specified (traditional or inclusive) norm group. When a student achieves a given norm-referenced score, say the 50th NP, on a mathematics test with the accommodation "teacher reads the test directions, stimulus material, and questions," the testing conditions should be considered along with the NP score. In this case, the valid interpretation is that the student who took the mathematics test that was read aloud performed as well as or better than 50% of the students in the norm group. In accordance with the principles set forth, CTB/McGraw-Hill's inclusive norm group includes those students able to participate in the test administration with or without accommodations.

Given these interpretive guidelines, CTB/McGraw-Hill recommends the following approaches for interpreting the results of inclusive test administrations.

Appropriate information for the interpretation of individual performance. A student who takes a test using accommodations should receive the same scale scores referenced to the same norms tables as students with the same test performance achieved under default conditions. However, individual student results obtained using testing accommodations should be interpreted in light of the accommodation(s) used. These summaries are composed of common accommodations categorized according to the effect on the appropriate interpretation of individual student results, as detailed in the following section.

Appropriate information for the interpretation of group performance. CTB/McGraw-Hill recommends that summaries of results that are used for accountability purposes be presented in both aggregated and disaggregated forms. *Aggregated* results are summaries of results that include all students tested. These should be presented with the number and percent of students who took the test(s) using accommodations so that the aggregated results can be interpreted with respect to changes in the use of accommodations across groups and years. Identifying the number and percent of students using accommodations provides valuable information.

Disaggregated result summaries include only students who meet a specified criterion, such as students who took the test(s) under the conditions defined as default by the test publisher. CTB/McGraw-Hill recommends disaggregating results for students who take tests under default conditions and presenting the results separately from those for students who take tests under accommodated conditions. However, reports should never be presented for any group for which the number is so small that the confidentiality of student information would be violated (FERPA, 1974). It is also important not to base inferences or important decisions on small numbers of students.

Additional information of assessment accommodations and a framework for developing accommodations may be found in *Guidelines for Inclusive Test Administration* (CTB/McGraw-Hill, 2002).

Additional Evidence for Validity

Please see Part 7 for more information regarding test fairness and differential item functioning. In addition, the scale score descriptive statistics presented in Table 13 also provide evidence of validity. The data in this table shows clear growth across the grades in performance on the *InView* subtests. This growth suggests that the *InView* subtests are measuring the change in cognitive skills that coincides with students' advancement through the primary and secondary school grades.

Table 3

InView and *TerraNova, The Second Edition* (TN2) Correlations

Grade 2 (Level 1)

SQ	AN	QR	VW	VC	Nonverb	Verb	*InView* Total	Subtest
								TN2
0.50	0.53	0.49	0.64	0.67	0.60	0.72	0.71	Reading
0.48	0.51	0.51	0.66	0.65	0.59	0.72	0.70	Vocabulary
0.53	0.56	0.54	0.70	0.71	0.64	0.77	0.75	Composite Reading
0.50	0.51	0.53	0.62	0.62	0.60	0.68	0.69	Language Arts
0.53	0.53	0.53	0.61	0.60	0.62	0.66	0.69	Language Mechanics
0.55	0.57	0.57	0.67	0.68	0.66	0.74	0.76	Composite Language
0.58	0.51	0.57	0.54	0.56	0.66	0.61	0.70	Mathematics
0.46	0.44	0.53	0.48	0.51	0.56	0.55	0.61	Mathematics Computation
0.58	0.53	0.62	0.57	0.61	0.69	0.65	0.73	Composite Mathematics
0.47	0.48	0.44	0.46	0.48	0.55	0.51	0.58	Science
0.48	0.47	0.46	0.54	0.53	0.56	0.59	0.62	Social Studies
0.33	0.35	0.36	0.54	0.51	0.41	0.58	0.53	Spelling
0.47	0.47	0.47	0.59	0.58	0.56	0.65	0.65	Word Analysis
0.62	0.60	0.65	0.70	0.74	0.73	0.79	0.82	Total
								InView
	0.64	0.53	0.52	0.51	0.88	0.57	0.81	Sequences (SQ)
		0.50	0.54	0.52	0.87	0.59	0.81	Analogies (AN)
			0.51	0.52	0.76	0.57	0.74	Quantitative Reasoning (QR)
				0.65	0.62	0.91	0.81	Verbal Reasoning—Words (VW)
					0.61	0.91	0.81	Verbal Reasoning—Context (VC)
						0.68	0.94	Nonverbal Total (Nonverb)
							0.89	Verbal Total (Verb)

Sample Size Range = 11,225–11,417

Table 3 (continued)

InView and *TerraNova, The Second Edition* (TN2) Correlations

Grade 3 (Level 1)

SQ	AN	QR	VW	VC	Nonverb	Verb	*InView* Total	Subtest
								TN2
0.50	0.54	0.50	0.62	0.68	0.61	0.72	0.71	Reading
0.46	0.47	0.47	0.59	0.61	0.54	0.66	0.64	Vocabulary
0.51	0.53	0.52	0.64	0.65	0.61	0.72	0.71	Composite Reading
0.51	0.52	0.51	0.60	0.66	0.60	0.70	0.70	Language Arts
0.49	0.52	0.51	0.53	0.58	0.59	0.62	0.66	Language Mechanics
0.57	0.58	0.56	0.61	0.68	0.67	0.73	0.75	Composite Language
0.62	0.57	0.60	0.59	0.68	0.71	0.71	0.78	Mathematics
0.47	0.43	0.53	0.46	0.47	0.56	0.52	0.59	Mathematics Computation
0.59	0.54	0.61	0.57	0.62	0.69	0.66	0.74	Composite Mathematics
0.52	0.53	0.51	0.58	0.60	0.61	0.66	0.68	Science
0.49	0.52	0.49	0.58	0.61	0.58	0.66	0.67	Social Studies
0.34	0.35	0.40	0.47	0.46	0.43	0.52	0.50	Spelling
0.50	0.53	0.53	0.58	0.58	0.61	0.65	0.68	Word Analysis
0.60	0.55	0.59	0.63	0.68	0.68	0.74	0.77	Total
								InView
	0.65	0.55	0.51	0.52	0.88	0.57	0.81	Sequences (SQ)
		0.52	0.55	0.54	0.88	0.60	0.83	Analogies (AN)
			0.49	0.54	0.78	0.57	0.75	Quantitative Reasoning (QR)
				0.62	0.61	0.90	0.79	Verbal Reasoning—Words (VW)
					0.62	0.90	0.80	Verbal Reasoning—Context (VC)
						0.69	0.94	Nonverbal Total (Nonverb)
							0.89	Verbal (Verb)

Sample Size Range = 9,934–10,095

Table 3 (continued)

InView and *TerraNova, The Second Edition* (TN2) Correlations

Grade 4 (Level 2)

SQ	AN	QR	VW	VC	Nonverb	Verb	*InView* Total	Subtest
								TN2
0.52	0.52	0.52	0.63	0.66	0.62	0.72	0.72	Reading
0.45	0.47	0.47	0.60	0.59	0.55	0.66	0.66	Vocabulary
0.52	0.53	0.52	0.67	0.66	0.62	0.75	0.74	Composite Reading
0.54	0.51	0.53	0.62	0.65	0.63	0.71	0.73	Language Arts
0.50	0.47	0.51	0.53	0.55	0.59	0.60	0.65	Language Mechanics
0.57	0.52	0.56	0.61	0.65	0.66	0.71	0.74	Composite Language
0.61	0.52	0.62	0.58	0.64	0.70	0.69	0.77	Mathematics
0.43	0.38	0.53	0.40	0.48	0.54	0.49	0.57	Mathematics Computation
0.58	0.51	0.64	0.55	0.62	0.69	0.66	0.75	Composite Mathematics
0.56	0.59	0.52	0.65	0.62	0.67	0.71	0.75	Science
0.53	0.53	0.51	0.64	0.63	0.63	0.72	0.73	Social Studies
0.36	0.32	0.35	0.48	0.46	0.42	0.53	0.51	Spelling
0.57	0.66	0.52	0.70	0.62	0.70	0.72	0.77	Total
								InView
	0.61	0.57	0.49	0.51	0.88	0.56	0.81	Sequences (SQ)
		0.49	0.52	0.52	0.84	0.58	0.80	Analogies (AN)
			0.46	0.52	0.80	0.55	0.76	Quantitative Reasoning (QR)
				0.60	0.58	0.89	0.78	Verbal Reasoning—Words (VW)
					0.61	0.90	0.80	Verbal Reasoning—Context (VC)
						0.67	0.94	Nonverbal Total (Nonverb)
							0.88	Verbal (Verb)

Sample Size Range = 10,177–10,293

Table 3 (continued)

InView and *TerraNova, The Second Edition* (TN2) Correlations

Grade 5 (Level 2)

SQ	AN	QR	VW	VC	Nonverb	Verb	*InView* Total	Subtest
								TN2
0.48	0.48	0.50	0.59	0.62	0.58	0.69	0.68	Reading
0.42	0.45	0.43	0.60	0.54	0.52	0.65	0.62	Vocabulary
0.51	0.52	0.50	0.67	0.64	0.61	0.74	0.73	Composite Reading
0.52	0.48	0.53	0.58	0.61	0.61	0.67	0.69	Language Arts
0.47	0.42	0.44	0.47	0.51	0.53	0.56	0.59	Language Mechanics
0.56	0.50	0.53	0.58	0.61	0.63	0.67	0.71	Composite Language
0.65	0.58	0.62	0.62	0.63	0.73	0.70	0.78	Mathematics
0.48	0.39	0.48	0.42	0.43	0.53	0.48	0.55	Mathematics Computation
0.63	0.55	0.61	0.58	0.60	0.71	0.66	0.75	Composite Mathematics
0.52	0.54	0.50	0.61	0.59	0.63	0.68	0.70	Science
0.53	0.54	0.48	0.65	0.61	0.62	0.71	0.71	Social Studies
0.38	0.37	0.36	0.49	0.45	0.44	0.53	0.52	Spelling
0.64	0.55	0.66	0.61	0.74	0.73	0.76	0.80	Total
								InView
	0.60	0.57	0.51	0.52	0.88	0.58	0.82	Sequences (SQ)
		0.48	0.54	0.51	0.84	0.59	0.80	Analogies (AN)
			0.46	0.51	0.80	0.54	0.75	Quantitative Reasoning (QR)
				0.59	0.61	0.89	0.79	Verbal Reasoning—Words (VW)
					0.61	0.89	0.80	Verbal Reasoning—Context (VC)
						0.68	0.94	Nonverbal Total (Nonverb)
							0.89	Verbal (Verb)

Sample Size Range = 9,593–9,659

Table 3 (continued)

InView and *TerraNova, The Second Edition* (TN2) Correlations

Grade 6 (Level 3)

SQ	AN	QR	VW	VC	Nonverb	Verb	*InView* Total	Subtest
								TN2
0.54	0.56	0.60	0.70	0.67	0.66	0.75	0.74	Reading
0.49	0.51	0.56	0.66	0.62	0.60	0.70	0.68	Vocabulary
0.55	0.60	0.62	0.73	0.69	0.68	0.78	0.77	Composite Reading
0.57	0.56	0.62	0.68	0.67	0.67	0.74	0.74	Language Arts
0.53	0.49	0.57	0.57	0.59	0.61	0.63	0.66	Language Mechanics
0.61	0.56	0.64	0.66	0.68	0.70	0.73	0.76	Composite Language
0.65	0.56	0.69	0.64	0.63	0.73	0.69	0.76	Mathematics
0.51	0.41	0.60	0.49	0.52	0.58	0.55	0.60	Mathematics Computation
0.64	0.54	0.72	0.62	0.63	0.73	0.69	0.75	Composite Mathematics
0.57	0.63	0.59	0.68	0.62	0.69	0.71	0.74	Science
0.57	0.57	0.61	0.70	0.66	0.68	0.75	0.75	Social Studies
0.39	0.33	0.42	0.47	0.45	0.44	0.51	0.50	Spelling
0.70	0.67	0.71	0.76	0.74	0.80	0.82	0.86	Total
								InView
	0.63	0.65	0.61	0.59	0.88	0.65	0.83	Sequences (SQ)
		0.57	0.63	0.56	0.86	0.65	0.82	Analogies (AN)
			0.63	0.64	0.85	0.69	0.83	Quantitative Reasoning (QR)
				0.68	0.72	0.92	0.86	Verbal Reasoning—Words (VW)
					0.69	0.91	0.83	Verbal Reasoning—Context (VC)
						0.77	0.96	Nonverbal Total (Nonverb)
							0.92	Verbal (Verb)

Sample Size Range = 10,052–10,207

Table 3 (continued)

InView and *TerraNova, The Second Edition* (TN2) Correlations

Grade 7 (Level 3)

SQ	AN	QR	VW	VC	Nonverb	Verb	*InView* Total	Subtest
								TN2
0.51	0.53	0.60	0.64	0.65	0.63	0.71	0.71	Reading
0.50	0.52	0.54	0.65	0.59	0.61	0.68	0.68	Vocabulary
0.56	0.60	0.62	0.71	0.67	0.69	0.77	0.77	Composite Reading
0.51	0.48	0.60	0.59	0.62	0.61	0.67	0.67	Language Arts
0.53	0.47	0.56	0.53	0.57	0.60	0.61	0.64	Language Mechanics
0.58	0.54	0.64	0.61	0.67	0.68	0.71	0.74	Composite Language
0.66	0.57	0.70	0.65	0.66	0.75	0.72	0.78	Mathematics
0.52	0.44	0.59	0.48	0.52	0.60	0.55	0.61	Mathematics Computation
0.65	0.55	0.71	0.62	0.65	0.74	0.69	0.77	Composite Mathematics
0.58	0.58	0.60	0.67	0.63	0.68	0.71	0.74	Science
0.55	0.57	0.59	0.69	0.65	0.66	0.73	0.74	Social Studies
0.42	0.39	0.48	0.52	0.50	0.50	0.56	0.55	Spelling
0.70	0.65	0.74	0.74	0.79	0.79	0.83	0.85	Total
								InView
	0.63	0.65	0.60	0.60	0.88	0.66	0.83	Sequences (SQ)
		0.55	0.63	0.57	0.85	0.66	0.81	Analogies (AN)
			0.62	0.64	0.85	0.69	0.83	Quantitative Reasoning (QR)
				0.67	0.72	0.92	0.85	Verbal Reasoning—Words (VW)
					0.70	0.91	0.84	Verbal Reasoning—Context (VC)
						0.78	0.96	Nonverbal Total (Nonverb)
							0.93	Verbal (Verb)

Sample Size Range = 10,063–10,239

Table 3 (continued)

InView and *TerraNova, The Second Edition* (TN2) Correlations

Grade 8 (Level 4)

SQ	AN	QR	VW	VC	Nonverb	Verb	*InView* Total	Subtest
								TN2
0.50	0.50	0.57	0.62	0.60	0.60	0.66	0.67	Reading
0.49	0.50	0.53	0.61	0.57	0.58	0.64	0.65	Vocabulary
0.56	0.57	0.62	0.70	0.65	0.67	0.73	0.74	Composite Reading
0.54	0.50	0.57	0.63	0.61	0.62	0.67	0.68	Language Arts
0.49	0.43	0.52	0.54	0.55	0.56	0.59	0.61	Language Mechanics
0.59	0.55	0.62	0.66	0.64	0.68	0.71	0.74	Composite Language
0.63	0.58	0.63	0.58	0.58	0.72	0.63	0.73	Mathematics
0.55	0.46	0.60	0.49	0.53	0.62	0.56	0.63	Mathematics Computation
0.67	0.59	0.70	0.61	0.63	0.76	0.68	0.77	Composite Mathematics
0.54	0.59	0.60	0.64	0.62	0.66	0.69	0.71	Science
0.52	0.53	0.55	0.62	0.58	0.62	0.66	0.68	Social Studies
0.45	0.41	0.44	0.52	0.44	0.50	0.53	0.54	Spelling
0.65	0.66	0.74	0.72	0.71	0.78	0.78	0.83	Total
								InView
	0.60	0.67	0.59	0.59	0.87	0.64	0.83	Sequences (SQ)
		0.61	0.64	0.58	0.85	0.66	0.82	Analogies (AN)
			0.62	0.63	0.88	0.68	0.84	Quantitative Reasoning (QR)
				0.69	0.72	0.92	0.85	Verbal Reasoning—Words (VW)
					0.69	0.92	0.83	Verbal Reasoning—Context (VC)
						0.76	0.96	Nonverbal Total (Nonverb)
							0.92	Verbal (Verb)

Sample Size Range = 9,370–9,616

Table 3 (continued)

InView and *TerraNova, The Second Edition* (TN2) Correlations

Grade 9 (Level 4)

SQ	AN	QR	VW	VC	Nonverb	Verb	*InView* Total	Subtest
								TN2
0.52	0.51	0.57	0.64	0.61	0.60	0.67	0.67	Reading
0.45	0.50	0.51	0.63	0.58	0.56	0.65	0.64	Vocabulary
0.55	0.56	0.60	0.71	0.65	0.65	0.74	0.73	Composite Reading
0.52	0.48	0.55	0.60	0.58	0.59	0.63	0.65	Language Arts
0.44	0.37	0.50	0.51	0.49	0.51	0.54	0.56	Language Mechanics
0.56	0.48	0.59	0.61	0.58	0.63	0.65	0.67	Composite Language
0.67	0.63	0.75	0.64	0.63	0.78	0.68	0.79	Mathematics
0.58	0.50	0.66	0.56	0.54	0.67	0.59	0.68	Mathematics Computation
0.68	0.60	0.76	0.64	0.63	0.78	0.69	0.79	Composite Mathematics
0.51	0.54	0.58	0.62	0.59	0.62	0.65	0.68	Science
0.53	0.56	0.59	0.64	0.60	0.63	0.67	0.69	Social Studies
0.47	0.41	0.49	0.54	0.49	0.52	0.55	0.57	Spelling
0.67	0.61	0.73	0.73	0.70	0.76	0.77	0.80	Total
								InView
	0.63	0.69	0.60	0.61	0.89	0.65	0.84	Sequences (SQ)
		0.63	0.64	0.59	0.85	0.66	0.82	Analogies (AN)
			0.63	0.65	0.89	0.69	0.86	Quantitative Reasoning (QR)
				0.71	0.71	0.92	0.85	Verbal Reasoning—Words (VW)
					0.70	0.93	0.85	Verbal Reasoning—Context (VC)
						0.76	0.96	Nonverbal Total (Nonverb)
							0.92	Verbal (Verb)

Sample Size Range = 10,700–11,040

Table 3 (continued)

InView and *TerraNova, The Second Edition* (TN2) Correlations

Grade 10 (Level 5)

SQ	AN	QR	VW	VC	Nonverb	Verb	*InView* Total	Subtest
								TN2
0.43	0.46	0.48	0.57	0.57	0.53	0.62	0.60	Reading
0.42	0.44	0.48	0.59	0.54	0.51	0.61	0.59	Vocabulary
0.46	0.49	0.54	0.62	0.62	0.57	0.68	0.65	Composite Reading
0.46	0.46	0.50	0.57	0.58	0.55	0.63	0.62	Language Arts
0.42	0.40	0.44	0.47	0.49	0.48	0.52	0.53	Language Mechanics
0.48	0.46	0.54	0.55	0.60	0.56	0.63	0.63	Composite Language
0.64	0.58	0.69	0.65	0.63	0.73	0.69	0.76	Mathematics
0.54	0.49	0.60	0.55	0.54	0.62	0.59	0.64	Mathematics Computation
0.64	0.59	0.71	0.66	0.64	0.75	0.70	0.77	Composite Mathematics
0.48	0.51	0.54	0.58	0.56	0.59	0.62	0.63	Science
0.49	0.52	0.57	0.65	0.61	0.60	0.68	0.68	Social Studies
0.39	0.41	0.43	0.51	0.48	0.47	0.54	0.53	Spelling
0.49	0.46	0.57	0.68	0.71	0.59	0.77	0.73	Total
								InView
	0.64	0.66	0.61	0.59	0.88	0.65	0.82	Sequences (SQ)
		0.62	0.68	0.60	0.86	0.69	0.83	Analogies (AN)
			0.65	0.67	0.88	0.71	0.85	Quantitative Reasoning (QR)
				0.70	0.74	0.92	0.86	Verbal Reasoning—Words (VW)
					0.71	0.92	0.85	Verbal Reasoning—Context (VC)
						0.79	0.96	Nonverbal Total (Nonverb)
							0.93	Verbal (Verb)

Sample Size Range = 9,514–9,729

Table 3 (continued)

InView and *TerraNova, The Second Edition* (TN2) Correlations

Grade 11 (Level 5, 6)

SQ	AN	QR	VW	VC	Nonverb	Verb	*InView* Total	Subtest
								TN2
0.42	0.42	0.52	0.58	0.57	0.53	0.63	0.61	Reading
0.43	0.46	0.48	0.63	0.56	0.53	0.65	0.62	Vocabulary
0.43	0.44	0.53	0.68	0.63	0.55	0.71	0.66	Composite Reading
0.47	0.45	0.54	0.60	0.60	0.57	0.65	0.64	Language Arts
0.45	0.40	0.47	0.53	0.53	0.52	0.58	0.58	Language Mechanics
0.51	0.47	0.53	0.64	0.62	0.59	0.68	0.67	Composite Language
0.66	0.61	0.70	0.66	0.69	0.76	0.74	0.79	Mathematics
0.60	0.54	0.66	0.56	0.60	0.69	0.64	0.71	Mathematics Computation
0.68	0.61	0.73	0.66	0.69	0.78	0.74	0.80	Composite Mathematics
0.47	0.51	0.54	0.61	0.59	0.59	0.65	0.65	Science
0.45	0.49	0.55	0.64	0.61	0.58	0.68	0.66	Social Studies
0.39	0.37	0.44	0.53	0.48	0.47	0.55	0.53	Spelling
0.66	0.65	0.75	0.80	0.79	0.80	0.85	0.85	Total
								InView
	0.63	0.65	0.59	0.60	0.87	0.64	0.82	Sequences (SQ)
		0.61	0.66	0.59	0.85	0.68	0.82	Analogies (AN)
			0.63	0.66	0.88	0.70	0.85	Quantitative Reasoning (QR)
				0.69	0.72	0.92	0.86	Verbal Reasoning—Words (VW)
					0.71	0.92	0.85	Verbal Reasoning—Context (VC)
						0.78	0.96	Nonverbal Total (Nonverb)
							0.92	Verbal (Verb)

Sample Size Range = 5,680–5,916

Table 3 (concluded)

InView and *TerraNova, The Second Edition* (TN2) Correlations

Grade 12 (Level 6)

SQ	AN	QR	VW	VC	Nonverb	Verb	*InView* Total	Subtest
								TN2
0.43	0.43	0.47	0.55	0.56	0.51	0.60	0.58	Reading
0.43	0.45	0.50	0.63	0.58	0.53	0.66	0.63	Vocabulary
0.50	0.51	0.54	0.69	0.65	0.60	0.73	0.70	Composite Reading
0.48	0.44	0.54	0.58	0.60	0.57	0.64	0.63	Language Arts
0.47	0.35	0.50	0.49	0.55	0.51	0.57	0.57	Language Mechanics
0.55	0.46	0.58	0.61	0.66	0.62	0.69	0.69	Composite Language
0.62	0.51	0.71	0.65	0.62	0.72	0.69	0.75	Mathematics
0.57	0.45	0.66	0.56	0.55	0.66	0.60	0.67	Mathematics Computation
0.64	0.50	0.73	0.65	0.62	0.73	0.69	0.76	Composite Mathematics
0.44	0.46	0.52	0.58	0.55	0.56	0.61	0.62	Science
0.46	0.49	0.57	0.63	0.58	0.60	0.66	0.66	Social Studies
0.38	0.34	0.42	0.48	0.41	0.45	0.48	0.49	Spelling
0.55	0.50	0.71	0.67	0.65	0.68	0.72	0.74	Total
								InView
	0.58	0.64	0.56	0.59	0.87	0.63	0.81	Sequences (SQ)
		0.56	0.64	0.57	0.83	0.66	0.80	Analogies (AN)
			0.61	0.65	0.86	0.69	0.83	Quantitative Reasoning (QR)
				0.70	0.71	0.92	0.85	Verbal Reasoning—Words (VW)
					0.71	0.92	0.85	Verbal Reasoning—Context (VC)
						0.77	0.96	Nonverbal Total (Nonverb)
							0.92	Verbal (Verb)

Sample Size Range = 6,149–6,351

Table 4

Number of Items in Item Bundles

Level	Bundle 1	Bundle 2	Bundle 3	Bundle 4
Sequences				
1	12	8		
2	6	5	6	3
3	5	5	5	5
4	3	6	3	8
5	5	6	6	3
6	7	5	3	5
Analogies				
1	5	4	6	5
2	5	4	6	5
3	5	4	6	5
4	5	4	6	5
5	5	4	6	5
6	5	4	6	5
Quantitative Reasoning				
1	8	6	6	
2	6	6	8	
3	6	6	8	
4	6	6	8	
5	6	6	8	
6	6	7	7	
Verbal Reasoning—Words				
1	4	6	6	4
2	6	5	6	3
3	7	3	3	7
4	6	3	3	8
5	5	5	2	8
6	4	3	5	8
Verbal Reasoning—Context				
1	4	7	6	3
2	4	7	6	3
3	4	7	6	3
4	4	7	6	3
5	4	7	6	3
6	4	7	6	3

Table 5

Factor Analysis Sample Size by Grade

Level	Grade	Sample Size
1	2	1,110
	3	1,347
2	4	1,707
	5	1,812
3	6	1,733
	7	1,713
4	8	1,610
	9	1,637
5	10	1,676
	11	1,621
6	11	1,227
	12	1,722

Table 6

Factor Analysis Results

Test Level	χ^2 (df)[1]	T[2]	GFI[3]	AGFI[4]	CFI[5]	RMSEA[6]
Model 1: Five-factor basic model						
Level 1	167.12* (109)	1	0.98	0.98	0.99	0.021
Level 2	314.13* (142)	1	0.98	0.98	0.98	0.026
Level 3	440.28* (142)	1	0.97	0.96	0.98	0.035
Level 4	440.83* (142)	1	0.97	0.96	0.98	0.036
Level 5	253.02* (142)	1	0.98	0.98	0.99	0.022
Level 6	369.23* (142)	1	0.97	0.96	0.98	0.033
Model 2: One general second-order (G) factor						
Level 1	264.12* (113)	0.63	0.98	0.97	0.98	0.033
Level 2	448.39* (146)	0.70	0.97	0.97	0.97	0.034
Level 3	567.36* (146)	0.78	0.97	0.96	0.97	0.041
Level 4	549.80* (146)	0.80	0.96	0.95	0.97	0.041
Level 5	356.26* (146)	0.71	0.98	0.97	0.98	0.030
Level 6	482.37* (146)	0.77	0.97	0.95	0.97	0.040
Model 3: Two second-order (V and NV) factors						
Level 1	197.56* (111)	0.85	0.98	0.97	0.99	0.025
Level 2	361.06* (144)	0.87	0.98	0.97	0.98	0.029
Level 3	531.94* (144)	0.83	0.97	0.96	0.97	0,040
Level 4	489.63* (144)	0.90	0.97	0.96	0.97	0.038
Level 5	328.07* (144)	0.77	0.98	0.97	0.99	0.028
Level 6	463.88* (144)	0.80	0.97	0.96	0.97	0.039

1) χ^2 = chi-square, df = degree of freedom
2) T = target coefficient, ratio of the chi-square of the first-order model to the chi-square of the second-order model
3) GFI = goodness of fit
4) AGFI = adjusted goodness of fit
5) CFI = comparative fit index
6) RMSEA = root-mean-square error of approximation values
* significant at the .01 level

Part 4: National Standardization

Purpose and Testing Schedule

InView was nationally standardized in the spring of 2000. The national standardization was conducted to allow students' test scores to be compared with the performance of students across the nation under the standardized test administration. The test administration period was April 10–May 19, 2000. Using the standardization data, the *InView* vertical scale was created, the item parameters were estimated, and the norms were developed. See Part 5 for more information on *InView* scaling and norming.

More Inclusive Standardization

More inclusive standardization was implemented in the sampling design to ensure that the norms reflect the requirements of current testing programs. A standard school administration is defined as one that involves all students according to their Individual Education Plans (IEPs), including participation with testing accommodations as indicated. This definition—which complements the familiar definition of a standard student administration—allowed for the collection of meaningful, inclusive norming data in the standardization study. More precise and inclusive norms will better meet the need for valid interpretations of the results of inclusive administrations of the new assessments. More details are provided in the following sections.

Sampling Procedures

In developing norms, it is not feasible to administer tests to all the students in the nation at all times of the year, thus the tests are administered as part of a standardization study to a representative sample of students—the norming sample—at a particular time, the norming date. More than 100,000 students, from Grade 2 through Grade 12, participated in the standardization study for *InView*.

The norms for *InView* reflect a major redesign of the norming procedures. This redesign ensured collection of national norming information that was both more precise and more inclusive.

With the *InView* standardization, greater precision was achieved by sampling individual schools rather than districts, and by stratifying the nation's schools according to more detailed demographic information. For example, under the previous norming design we could assure that we selected an appropriate number of students from districts classified as urban, suburban, or rural. We are now able to select an appropriate number of students from schools located in the urban fringe of large central cities and at six other levels of urbanization.

Students in the spring norm groups were identified using stratified random sampling procedures designed to ensure that the norm group constituted a sample of students that accurately represented the nation's school population, thereby fairly representing all minority and socioeconomic groups. To achieve this goal, two sampling designs were used. Public schools were stratified by region, community type, and socioeconomic status (Orshansky percentile). Catholic schools and private, non-Catholic schools were stratified by region and community type.

To ensure continuity in the norms from grade to grade, a special sampling procedure was used for public schools and Catholic schools. High schools were randomly sampled and asked to identify the major elementary and/or middle schools that sent their students to these selected high schools. The elementary and middle schools thus identified, were asked to be included in the testing. All students were asked to be tested in the private, non-Catholic schools involved in the standardization study.

Stratification Variables

Quality Education Data, Inc. (QED) and their National Education Database™ (a comprehensive census database of K–12 educational institutions in the United States) provide accurate and timely information about the nation's educational institutions. This database was used to obtain sampling data on geographic region, community type, socioeconomic status, and special needs. Use of these data helps to ensure that our norms accurately reflect achievement levels of the nation's student population as a whole.

Geographic Region

Four geographic regions constitute the first variable used for stratification during the standardization study. These regions, and the states included in each, are defined as follows:

- Eastern: Connecticut, Delaware, District of Columbia, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont
- Southern: Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, West Virginia
- Mid-continent: Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin
- Western: Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oklahoma, Oregon, Texas, Utah, Washington, Wyoming

Community Type

The second stratification variable, community type, was based on seven-levels of metropolitan classification as designated by the U.S. Bureau of the Census. The levels identified included: large central city, mid-size central city, urban fringes of large central city, urban fringe of mid-size central city, large town, small town, and rural.

Socioeconomic Status

Public schools were further stratified into two categories of socioeconomic status (high and low) using the percentage of students eligible for Chapter 1 funding. The percentage of eligible Chapter 1 students indicates the percentage of children living in households with incomes below the federal poverty level. To assure further sample diversity, schools within cells were implicitly stratified by state and school size.

Special Needs

Schools selected to participate in the standardization study were asked to test all students who would ordinarily be included in their regular testing program, including those who would participate in regular testing under accommodated conditions. For an accurate description of the standardization samples, each standardization examiner was asked to describe the category of special needs that any student may have, as well as any modified conditions for testing specified by the student's Individualized Education Program (IEP) that were applied during the standardized testing. This information was identified on the answer sheet during each test administration. The available categories included learning disabled, physically disabled, emotionally disabled, and mentally disabled.

Standardization Sample

Nearly 108,000 students in Kindergarten through Grade 12 participated in the spring 2000 standardization. Data from these students were used in samples for vertical scaling, item calibration, TCS/2 to *InView* equating, and factor analysis. The Kindergarten and Grade 1 standardization data resulted in the updated norms for the *Primary Test of Cognitive Skills* (PTCS). PTCS is a cognitive skills test intended for Kindergartners and first graders that measures verbal, spatial, memory, and conceptual abilities. For more details of PTCS, refer to *Primary Test of Cognitive Skills, Technical Bulletin.*

Table 7 shows the sample sizes for each grade of the *InView* standardization study. Tables 8 and 9 show the characteristics of the sample of the *InView* standardization study. Statistical sampling weights are used to adjust the percentage of each cell in standardization sample so that it more closely represents the proportion in the nation. Weighted percentages in Tables 8 and 9 show the percentages for standardization sample after the statistical sampling weights have been applied.

Co-normed Assessment

InView and PTCS were normed concurrently with *TerraNova, The Second Edition.* PTCS was previously normed with *TerraNova* in 1996. The co-standardization of *InView* and *TerraNova, The Second Edition* allowed for the derivation of anticipated achievement scores. See Part 5 for more details on anticipated achievement.

Table 7

Sample Size by Grade for the Standardization Study

Grade	Sample Size
2	11,563
3	10,259
4	10,455
5	9,789
6	10,518
7	10,465
8	9,819
9	11,341
10	9,920
11	7,310
12	6,473
Total	**107,912**

Table 8

Characteristics of Students Participating in the Standardization Study

	Unweighted Percentage	Weighted Percentage	National Percentage
School Type			
Public	78.6	89.3	90.7[1]
Private Non-Catholic	10.5	5.9	4.4
Catholic	10.9	4.8	5.0
Community Type			
Large Central City	16.7	15.7	17.7
Mid-size Central City	12.2	17.6	16.4
Total Urban	**28.9**	**34.3**	**34.2**
Urban Fringe of Large City	15.3	23.0	28.6
Urban Fringe of Mid-size City	12.6	11.5	9.9
Large Town	1.4	1.8	1.8
Small Town	16.2	12.9	11.0
Total Suburban	**45.5**	**49.2**	**51.2**
Rural	24.6	16.2	13.2
Unclassified	1.1	1.4	1.4
Region			
East	10.5	13.1	20.1[1]
Middle	29.6	25.6	23.2
South	34.1	24.9	23.9
West	25.9	36.5	32.8
Ethnicity			
African American	16.2	15.9	15.1[2]
Asian	3.0	3.4	4.5
Hispanic	9.4	8.7	16.2
Other	71.4	72.0	64.2

[1] National percentages based on information provided by QED, 2001

[2] National percentages based on 1999 data provided by the U.S. Bureau of Census, Current Population Survey, October 1999

Table 9

Characteristics of Public School Students Participating in the Census, Current Population Survey, October 1999

	Unweighted Percentage	Weighted Percentage	National Percentage
Community Type			
Large Central City	12.0	15.6	17.5
Mid-size Central City	11.1	16.9	16.0
Total Urban	**23.1**	**32.5**	**33.6**
Urban Fringe of Large City	14.6	23.1	29.2
Urban Fringe of Mid-size City	12.4	11.2	9.5
Large Town	1.3	1.5	1.6
Small Town	17.9	13.0	11.0
Total Suburban	**46.2**	**48.8**	**51.4**
Rural	29.8	17.5	13.9
Unclassified	1.0	1.2	1.2
Region			
East	8.2	11.1	19.1[1]
Middle	24.6	24.8	22.8
South	38.0	25.6	24.3
West	29.2	38.5	33.9
SES			
Low	16.2	13.2	15.6[1]
Below Average	29.0	27.1	19.9
Average	32.3	33.5	30.3
Above Average	15.3	15.2	19.4
High	7.2	11.0	14.9
Ethnicity			
African American	17.0	16.6	17.2[2]
Asian	2.9	3.4	4.0
Hispanic	8.6	8.7	15.6
Other	71.5	71.3	63.2
%Free/Reduced Lunch	35.2	32.5	34.6[3]

[1] National percentages based on information provided by QED, 2001

[2] National percentages based on school year 1999–2000 data by the Department of Education, National Center for Education Statistics

[3] National percentages based on school year 1997–1998 data by the Department of Education, National Center for Education Statistics

PART 5: SCALING AND SCORE TYPES

Scaling

Using the empirical data collected from the standardization, the *InView* items across all six test levels were calibrated to create a single continuum. The process for creating such a single scale across test levels is called "vertical scaling." After the vertical scaling, the items were again calibrated using different data sets to obtain the item parameter estimates, and then scale-score scales were created. IRT was used throughout the scaling to place the items and the students on a common scale. The item parameter estimates on the scales will be used to estimate and report student scale scores. More details of these steps follow. For details regarding the IRT model that was used to calibrate the *InView* items, please refer to Part 2.

Vertical Scaling

A common item equating design was used to create a scale that can be used to measure student performance across all grade levels. Because students in the same grade took two adjacent test levels, IRT procedures can be used to create a vertical scale that links the tests across all levels. Linkages among the levels of the *InView* subtests were established in the following manner. During the spring standardization, Grade 3 students took both Level 1 and Level 2 of *InView*. In addition, Grade 5 students took Levels 2 and 3; Grade 7 students took Levels 3 and 4; Grade 9 students took Levels 4 and 5; and Grade 11 students took Levels 5 and 6. Using this data, 120 items at all 6 test levels were simultaneously calibrated to create a single scale for the subtest. The table below specifies the sample sizes included in the vertical scaling.

Grades	Level	Sample Size
3	1 and 2	1,897
5	2 and 3	1,269
7	3 and 4	1,545
9	4 and 5	1,844
11	5 and 6	1,228
		Total = 7,783

This vertical scale of *InView* permits inferences to be made regarding student performance from the early grades of elementary school to the end of high school. This vertical scale can be viewed as a developmental continuum. As students develop new skills, they move along the continuum.

Item Calibration

After a vertical scale was created for a given subtest, the items were calibrated separately at each of the six test levels using a larger sample and linked back to the vertical scale. The data used for the vertical scaling were not part of the data used for item calibration. The sample sizes for item calibration are shown below.

Grades	Level	Sample Size
2 and 3	1	16,121
4 and 5	2	15,373
6 and 7	3	15,468
8 and 9	4	15,825
10 and 11	5	11,620
11 and 12	6	6,835
		Total = 81,242

Creation of Scale Scores and Setting Lowest and Highest Obtainable Scale Scores

After the item calibration, scale scores were created. Scale scores are units of a single, equal interval vertical scale applied across all levels of *InView*, regardless of grade, test level, or time of testing. These scale scores are expressed in numbers that range theoretically from 0 to 999. As stated above, the equal interval property of scale scores permits arithmetic functions to be performed using the scale scores. Because each subtest was calibrated

separately, the scale for each subtest is unique. For *InView*, a subtest scale-score scale was created so that the scale score mean and standard deviation were approximately 500 and 60, respectively, at Level 4, Grade 8.

The maximum-likelihood procedure cannot produce scale-score estimates for students with perfect scores or scores with zero correct responses. Furthermore, maximum-likelihood estimates are available for students with extreme scores other than perfect or zero, but these estimates occasionally have standard errors of measurement that are very large, and differences between these extreme values have little meaning. Scores for these students, therefore, are established based on a rational but necessarily non-maximum-likelihood procedure. The scale scores for these extreme scores, called the lowest obtainable scale score (LOSS) and the highest obtainable scale score (HOSS), were set separately by levels during the standardization analysis. The final LOSS and HOSS values were chosen to increase as test level increases and to be more extreme than the maximum-likelihood scale scores that are obtainable from number-correct scores for that level (Yen, 1984). The final LOSS and HOSS values also were chosen to have predicted standard errors of measurement not more than 15 times larger than the smallest standard error of measurement for scale scores at that level. The *InView* LOSS and HOSS values are reported in the *InView Norms Book*.

Score Types

Scale Scores

For each subtest, scale scores are based on students' performance on all the items in the test. To obtain this scale score, either of two procedures can be used: item response-pattern scoring or number-correct (raw) scoring.

The IRT, or item response-pattern, scoring procedure produces maximum-likelihood trait estimates (scale scores) based on patterns of item responses, as described by Lord (1974; 1980, pp. 179–181). The likelihood equation can have multiple maxima; therefore, a scoring algorithm was developed that finds the scale score at the global maximum in the likelihood function. This algorithm was used in obtaining item response-pattern scale scores for normative distributions and is used for operational scoring.

In addition to pattern scoring, a second scoring procedure is available based on the number of correct items (Yen, 1984). This procedure produces trait estimates based on number-correct scores. A number-correct to scale score (that is, trait estimate) conversion table is produced for each subtest and is included in the *InView Norms Book.*

Tau-equivalence of Item Response-Pattern and Number-Correct Scale Scores

The scale scores based on item response-pattern scoring and those based on number-correct scoring are tau-equivalent (Yen, 1984). Tau-equivalence is met when students are expected to receive the same scale scores from the two scoring methods allowing for differences in error variances in these scores. Tau-equivalence has been verified empirically for groups of African-American students as well as for heterogeneous groups (Green & Yen, 1983; Yen, 1984). The item response-pattern method has lower standard errors of measurement than the number-correct method because the item response-pattern contains more information than does the number-correct score (Yen & Candell, 1991).

Total Scores

Three total scores are available and each is the average of the scale scores that contribute to it. The Nonverbal Total is the average of Sequences, Analogies, and Quantitative Reasoning; the Verbal Total is the average of Verbal Reasoning—Words and Verbal Reasoning—Context. The Total Test score is the average of the scale scores for the five subtests.

Derived Scores

Derived scores, such as percentile ranks, normal curve equivalents, and grade equivalents, are normative. These scores are derived from a scale score distribution of a norming sample. Norm-referenced scores allow comparisons

of students' performance with the performance of students in a specified group. National norms are developed so that students' scores can be compared with the performance of other students across the country.

The following procedures were followed for each subtest and the three total scores to derive the grade and age percentile norms. The tests administered to the norming sample were first scored using item response-pattern scoring to produce scale scores. Scale-score frequency distributions were then tabulated by grade. The representative national frequency distributions were obtained by first calculating a weight for each student and then using the student's weight instead of unity to tabulate frequency. The student's weight was calculated as a function of the total number of students tested at that grade in the given sampling cell and the population of that cell.

Specifically, the weight for all students in cell *i* at Grade *j* equals the proportion of the national Grade *j* population divided by the proportion of the sample in cell *i* at Grade *j*. Thus, if the obtained number of students in the sample in cell *i* at Grade *j* is smaller than a strictly proportionate sample, then each student in that cell and grade has a weight larger than unity; in that case, the aggregate contribution of those students is proportionally representative.

Each weighted distribution was smoothed using a procedure described by Holland, King, and Thayer (1988). Next, cumulative midpoint percentile ranks were computed for each scale score. This yielded scale score-to-percentile rank conversions for the spring (first week of April) empirical norm date. The percentile rank norms for other test dates were determined by linear or straight-line interpolation. The same general procedures were used to generate age-percentile ranks. Weighted scale-score frequency distributions were tabulated and smoothed for each subtest and the total test by chronological age, and cumulative midpoint percentile ranks were computed. Age and grade stanines, which are normalized scores, were obtained by transforming the age and grade percentile ranks to the stanine scale of nine equal units.

The Cognitive Skills Index (CSI) is an age-based normalized score used to describe a student's overall performance on *InView*. The CSI is based on the normalized age-percentile norms for Total score. CSI has a mean of 100 and a standard deviation of 16 at any given chronological age, measured in monthly increments.

Anticipated Achievement Scores

When administered in combination with *TerraNova, The Second Edition, InView* yields anticipated achievement scores. Anticipated achievement scores may be viewed as a special kind of norm that enables users to compare an individual's level of performance on *TerraNova, The Second Edition* with the performance of students of similar age, grade, and cognitive ability. The primary purpose of anticipated achievement scores is to provide an academic performance evaluation standard that is more fair and reasonable than an undifferentiated national norm. The following four anticipated achievement scores are available for each component and the total test of *TerraNova, The Second Edition*: the anticipated achievement scale score (AASS), the anticipated achievement national percentile (AANP), the anticipated achievement grade equivalent (AAGE), and the anticipated achievement normal curve equivalent (AANCE).

These anticipated achievement scores are not a simple, grade-based national norm, but are an estimated norm derived through multiple regression analyses that included as predictors chronological age in months, grade, test month, and the scores on the five *InView* subtests. Grade-based national norms indicate performance relative only to grade placement; these types of grade-based norms are especially useful as descriptive statistics and as an empirical basis for group comparisons. Anticipated achievement scores, however, provide a more realistic expectation for individual performance because the reference group is more specific. A student's anticipated achievement score represents an estimate of the average score for students with the same age, grade and month in school, and *InView* scores. The difference between a student's obtained score and anticipated achievement score is an estimate of the student's achievement *above* or *below* the average of students with similar attributes. A discrepancy between a student's potential and actual performance can be used to help screen students for further diagnostic testing.

Anticipated achievement is based on the assumption that academic achievement (as measured by *TerraNova, The Second Edition*) is conditioned by cognitive skills (as measured by *InView*). In other words, the cognitive skills

test reflects skills that contribute to academic performance but are typically not explicitly taught as part of school curricula. The interpretation of anticipated achievement, therefore, does not require addressing such issues as whether the cognitive skills test measures "innate" ability or merely a level of ability that reflects an out-of-school environment that is advantageous for learning.

Anticipated achievement does not capture all relevant sources of out-of-school variability. The statistical procedures used to develop the anticipated achievement norms were such that the average anticipated achievement score in the national sample on which the norms were based would exactly equal the average score on the test. This procedure does not ensure that demographically advantaged schools will not receive better than average anticipated achievement norms; it does, however, reduce the gap that would be found if the regular national norms were used.

Table 10 shows the levels of *InView* and *TerraNova, The Second Edition* that may be administered in combinations to obtain anticipated achievement scores. The *InView* level in the table marked with a (◊) is outside the range for which data were collected during the standardization, but may be used with the understanding that resulting anticipated achievement scores are less precise than those scores derived from the test combinations for which data were collected. As stated above, anticipated achievement takes a student's age into account. If a student's age is outside the age range used in the formulas to compute anticipated achievement, his or her age is reset to the minimum or maximum value of the range, whichever is appropriate. The anticipated achievement scores for students whose ages have been reset may not be as precise as the anticipated achievement scores for students whose ages are within the range specified for each grade.

Table 10

InView Combinations with *TerraNova, The Second Edition*

Grade	*TerraNova, The Second Edition*	*InView*
2	Level 11*, 12, or 13	Level 1
3	Level 12 or 13	Level 1
3 (Spring)	Level 14*	Level 2◊
4	Level 14 or 15	Level 2
5	Level 14, 15, or 16*	Level 2
6	Level 15*, 16, or 17	Level 3
7	Level 16, 17, or 18*	Level 3
8	Level 17*, 18, or 19	Level 4
9	Level 18, 19, or 20*	Level 4
10	Level 19*, 20, or 21/22	Level 5
11	Level 20 or 21/22	Level 5
11	Level 21/22	Level 6
12	Level 20* or 21/22	Level 6

* Off-level testing

Part 6: Descriptive Statistics and Reliability

Estimated Raw Score Descriptive Statistics

Item response theory was used to obtain descriptive statistics for model derived estimated raw scores for the *InView* subtests. The methods used to obtain these values are described in the Appendix of this report. These data are presented for descriptive purposes and should not be used as normative information. Normative information for *InView* is available in the *InView Norms Book*.

Table 11 presents descriptive statistics for the five *InView* subtests and the Total test. The Total test comprises all items from the five subtests. Listed by grade and level are the subtests, number of items, number of students in the spring standardization sample (N), the mean raw scores (RS Means), the raw score standard deviations (RS SD), average p-values, standard error of measurement (SEM) for raw scores, and internal consistency (KR-20). The average p-value equals the mean raw score divided by the maximum possible raw score for the test.

The data in Table 11 suggest that the tests are well targeted to the students' ability across the different grade levels. The raw score means of the various subtests range between 12 to 15 out of 20 points possible. The standard deviations suggest that the distribution of test scores falls mainly in the middle score range such that a very small proportion of students are receiving either very low or very high raw scores. In other words, the tests exhibit little or no floor or ceiling effects. Average p-values, typically between .60 to .75, further support the claim that, on average, the test difficulties are well targeted to student ability.

Reliability is an index of the consistency of test score results. A reliable test produces stable scores, that is, very similar score distributions would result if the test were administered repeatedly under similar conditions to the same students without memory or fatigue affecting scores. Test development procedures are designed to ensure that *InView* maintains a high degree of reliability.

As a measure of test reliability, the Kuder-Richardson Formula 20 (KR-20) is a frequently used measure of internal consistency for tests consisting of dichotomously scored items (Allen & Yen, 1979). Based on a single administration of a test, this statistic provides a reliability estimate that equals the average of all split-half coefficients that would be obtained on all possible divisions of the test into halves. Such a split-half coefficient would be obtained by correlating scores on one half of the test with the scores on the other half, and then adjusting the correlation with the Spearman-Brown formula so that it applies to the whole test. The high degree of internal consistency among the items composing the subtests are demonsrated by the KR-20s, predominantly in the range of .80 to .95.

Standard Error of Measurement

It is important to note that a test score is a single observation of behavior, and that measurement error is associated with any test score. From these observations, inferences about the cognitive skills of an individual or group of students may be made. The fact that an observed test score does not perfectly reflect an individual's true score gives rise to a statistic called the standard error of measurement (SEM).

A student's exact true score cannot be known. It is the average test score that would result if the test could be administered repeatedly without the effects of practice or fatigue. The standard error of measurement is an estimate of the standard deviation of an individual's observed scores from the repeated administrations. For practical purposes, this statistic can be used to obtain a range within which a student's true score is likely to occur.

It is expected that 68 percent of the time a student's true score would fall within one standard error of measurement of that student's obtained score from a single administration, and that 95 percent of the time the true score would fall within two standard errors of the observed score. An observed score, therefore, should be regarded not as a student's true score but as a point within a range that probably includes a student's true score.

Expected Item Difficulty

Table 12 presents the expected *InView* item difficulties, also known as p-values, in the form of proportion correct response. The derivation of the expected item difficulties is given in the Appendix. In this table, "Item" indicates the sequential order of the item as it appears in the test book.

The item difficulties presented in Table 12 ranged from approximately .95 for the easiest items to approximately .30 for the most difficult items. Items at the extreme ends of difficulty are necessary to account for and discriminate between the wide range of ability among examinees. The majority of items had difficulties that typically ranged from .50 to .70. Item difficulties in this range are most useful because they supply a large amount of information with which to discriminate the level of ability among examinees. The data in Table 12 show that the item difficulties for the *InView* subtests are appropriate for measuring the range of ability found at the various grade levels.

In general, the items near the beginning of each test are easier and the items near the end of each test are more difficult. This variance in difficulty allows for the tests to account for the differences in students' ability within each grade. Item difficulties also tend to vary somewhat by item format. As stated in Part 1, the *InView* subtests contain several different item formats. Items in the same format are grouped together into sections. Within these sections, easier items generally are presented first and more difficult items are presented later. As such, difficulties do not necessarily progress exactly from the easiest item to the most difficult item across the entire length of the test.

Scale Score Descriptive Statistics

Table 13 presents scale score descriptive statistics for *InView* by grade and level. Included in the table are number of students in the sample (N), the mean scale score (Mean), the scale score standard deviation (SD), and the median scale score (Median). These statistics were computed using normative distributions from the spring standardization data and are presented for descriptive purposes only. Information regarding norms for *InView* are available in the *InView Norms Book*.

The scales scores show steady growth in average performance over the entire grade range. This growth in performance is greatest in the early grades and least in the later grades. Given the decrease in growth at the later grades, score distributions show more overlap for high school students than for elementary schools students. Scale-score standard deviations tend to be mostly in the 50s and 60s and are relatively similar over the grades.

Table 11

Raw Score Statistics for *InView* *

Grade 2 (Level 1)

Scale	Number of Items	N	RS Means	RS SD	Average p-values	SEM	KR-20
Sequences	20	11,407	13.2	3.8	0.66	1.7	0.80
Analogies	20	11,410	12.9	4.2	0.64	1.7	0.84
Quantitative Reasoning	20	11,411	13.2	3.8	0.66	1.8	0.77
Verbal Reasoning—Words	20	11,424	13.3	3.8	0.67	1.7	0.80
Verbal Reasoning—Context	20	11,388	12.4	4.4	0.62	1.8	0.84
Total	**100**	**11,237**	**64.7**	**16.1**	**0.65**	**3.9**	**0.94**

Grade 3 (Level 1)

Scale	Number of Items	N	RS Means	RS SD	Average p-values	SEM	KR-20
Sequences	20	10,083	14.9	3.5	0.75	1.6	0.79
Analogies	20	10,079	14.8	4.0	0.74	1.6	0.84
Quantitative Reasoning	20	10,050	14.8	3.6	0.74	1.7	0.78
Verbal Reasoning—Words	20	10,098	15.2	3.4	0.76	1.6	0.79
Verbal Reasoning—Context	20	10,064	14.6	4.0	0.73	1.6	0.85
Total	**100**	**9,942**	**74.9**	**14.9**	**0.74**	**3.6**	**0.94**

Grade 4 (Level 2)

Scale	Number of Items	N	RS Means	RS SD	Average p-values	SEM	KR-20
Sequences	20	10,301	12.8	4.0	0.64	1.8	0.80
Analogies	20	10,301	13.7	4.0	0.69	1.7	0.82
Quantitative Reasoning	20	10,250	12.1	4.2	0.61	1.8	0.80
Verbal Reasoning—Words	20	10,294	12.8	3.9	0.64	1.8	0.79
Verbal Reasoning—Context	20	10,278	13.4	4.1	0.67	1.7	0.82
Total	**100**	**10,187**	**64.7**	**16.1**	**0.65**	**3.9**	**0.94**

Grade 5 (Level 2)

Scale	Number of Items	N	RS Means	RS SD	Average p-values	SEM	KR-20
Sequences	20	9,654	14.0	3.8	0.70	1.7	0.81
Analogies	20	9,653	14.8	3.7	0.74	1.6	0.81
Quantitative Reasoning	20	9,654	13.2	4.2	0.66	1.8	0.82
Verbal Reasoning—Words	20	9,648	13.8	4.0	0.69	1.7	0.82
Verbal Reasoning—Context	20	9,651	14.4	3.9	0.72	1.6	0.83
Total	**100**	**9,591**	**70.3**	**16.0**	**0.70**	**3.8**	**0.95**

* These statistics are based on the spring 2001 normative data.

Table 11 (continued)

Raw Score Statistics for *InView* *

Grade 6 (Level 3)

Scale	Number of Items	N	RS Means	RS SD	Average p-values	SEM	KR-20
Sequences	20	10,223	13.1	4.3	0.65	1.8	0.82
Analogies	20	10,202	13.2	4.2	0.66	1.7	0.83
Quantitative Reasoning	20	10,149	12.1	4.0	0.60	1.8	0.80
Verbal Reasoning—Words	20	10,200	13.3	4.1	0.67	1.7	0.82
Verbal Reasoning—Context	20	10,201	13.2	3.9	0.66	1.7	0.81
Total	**100**	**10,062**	**64.7**	**17.3**	**0.65**	**3.9**	**0.94**

Grade 7 (Level 3)

Scale	Number of Items	N	RS Means	RS SD	Average p-values	SEM	KR-20
Sequences	20	10,247	13.6	4.3	0.68	1.8	0.83
Analogies	20	10,232	13.8	4.1	0.69	1.7	0.83
Quantitative Reasoning	20	10,179	12.7	4.2	0.63	1.8	0.82
Verbal Reasoning—Words	20	10,229	14.0	4.0	0.70	1.7	0.83
Verbal Reasoning—Context	20	10,218	13.6	3.9	0.68	1.7	0.82
Total	**100**	**10,078**	**67.7**	**17.6**	**0.68**	**3.8**	**0.94**

Grade 8 (Level 4)

Scale	Number of Items	N	RS Means	RS SD	Average p-values	SEM	KR-20
Sequences	20	9,624	13.1	4.3	0.66	1.8	0.82
Analogies	20	9,611	12.9	4.4	0.64	1.8	0.83
Quantitative Reasoning	20	9,583	11.6	4.6	0.58	1.8	0.84
Verbal Reasoning—Words	20	9,607	12.8	4.4	0.64	1.8	0.83
Verbal Reasoning—Context	20	9,598	13.4	4.1	0.67	1.7	0.84
Total	**100**	**9,366**	**63.4**	**19.4**	**0.64**	**4.0**	**0.96**

Grade 9 (Level 4)

Scale	Number of Items	N	RS Means	RS SD	Average p-values	SEM	KR-20
Sequences	20	11,051	13.4	4.3	0.67	1.8	0.83
Analogies	20	11,034	13.3	4.4	0.66	1.8	0.83
Quantitative Reasoning	20	10,832	12.0	4.7	0.60	1.8	0.85
Verbal Reasoning—Words	20	11,012	13.2	4.4	0.66	1.8	0.84
Verbal Reasoning—Context	20	10,987	13.7	4.1	0.68	1.6	0.84
Total	**100**	**10,709**	**65.3**	**19.5**	**0.66**	**3.9**	**0.95**

* These statistics are based on the spring 2001 normative data.

Table 11 (concluded)

Raw Score Statistics for *InView* *

Grade 10 (Level 5)

Scale	Number of Items	N	RS Means	RS SD	Average p-values	SEM	KR-20
Sequences	20	9,735	12.7	4.3	0.63	1.8	0.82
Analogies	20	9,732	12.6	4.2	0.63	1.9	0.80
Quantitative Reasoning	20	9,665	12.0	4.2	0.60	1.8	0.82
Verbal Reasoning—Words	20	9,724	12.9	4.1	0.64	1.7	0.83
Verbal Reasoning—Context	20	9,707	13.1	4.2	0.65	1.7	0.83
Total	**100**	**9,519**	**62.4**	**18.7**	**0.63**	**4.0**	**0.95**

Grade 11 (Level 5)

Scale	Number of Items	N	RS Means	RS SD	Average p-values	SEM	KR-20
Sequences	20	5,921	12.8	4.3	0.64	1.8	0.82
Analogies	20	5,897	12.9	4.2	0.65	1.8	0.81
Quantitative Reasoning	20	5,779	12.2	4.3	0.61	1.8	0.83
Verbal Reasoning—Words	20	5,863	13.4	4.1	0.67	1.7	0.83
Verbal Reasoning—Context	20	5,825	13.5	4.2	0.67	1.7	0.84
Total	**100**	**5,683**	**64.3**	**18.6**	**0.65**	**4.0**	**0.95**

Grade 11 (Level 6)

Scale	Number of Items	N	RS Means	RS SD	Average p-values	SEM	KR-20
Sequences	20	5,921	12.5	4.2	0.62	1.8	0.81
Analogies	20	5,897	11.9	4.2	0.59	1.8	0.81
Quantitative Reasoning	20	5,779	10.8	4.6	0.54	1.8	0.85
Verbal Reasoning—Words	20	5,863	11.5	4.6	0.57	1.9	0.84
Verbal Reasoning—Context	20	5,825	12.3	4.4	0.61	1.8	0.84
Total	**100**	**5,683**	**58.0**	**19.2**	**0.59**	**4.0**	**0.95**

Grade 12 (Level 6)

Scale	Number of Items	N	RS Means	RS SD	Average p-values	SEM	KR-20
Sequences	20	6,354	12.8	4.3	0.64	1.8	0.82
Analogies	20	6,359	12.0	4.2	0.60	1.8	0.81
Quantitative Reasoning	20	6,294	11.1	4.6	0.56	1.8	0.85
Verbal Reasoning—Words	20	6,334	12.0	4.4	0.60	1.8	0.83
Verbal Reasoning—Context	20	6,311	12.5	4.3	0.62	1.7	0.84
Total	**100**	**6,160**	**58.8**	**19.3**	**0.60**	**4.1**	**0.96**

* These statistics are based on the spring 2001 normative data.

Table 12

Item Difficulties (p-values)*

Grade 2 (Level 1)

Item	SQ	AN	QR	VW	VC
1	0.90	0.94	0.92	0.82	0.91
2	0.94	0.95	0.89	0.75	0.85
3	0.94	0.81	0.77	0.38	0.78
4	0.90	0.92	0.68	0.39	0.83
5	0.85	0.83	0.62	0.87	0.83
6	0.88	0.81	0.49	0.81	0.77
7	0.84	0.71	0.48	0.83	0.79
8	0.72	0.87	0.30	0.64	0.73
9	0.80	0.72	0.91	0.46	0.70
10	0.66	0.63	0.87	0.41	0.69
11	0.64	0.57	0.60	0.89	0.68
12	0.73	0.74	0.70	0.76	0.66
13	0.63	0.51	0.51	0.87	0.49
14	0.44	0.50	0.42	0.84	0.60
15	0.45	0.48	0.88	0.75	0.46
16	0.42	0.21	0.70	0.40	0.33
17	0.43	0.43	0.60	0.87	0.45
18	0.44	0.51	0.70	0.87	0.28
19	0.25	0.25	0.61	0.45	0.32
20	0.30	0.48	0.52	0.25	0.25

Grade 3 (Level 1)

Item	SQ	AN	QR	VW	VC
1	0.94	0.97	0.95	0.88	0.95
2	0.97	0.97	0.93	0.82	0.93
3	0.97	0.88	0.83	0.50	0.87
4	0.95	0.96	0.74	0.52	0.91
5	0.92	0.90	0.72	0.93	0.92
6	0.94	0.89	0.56	0.91	0.89
7	0.92	0.83	0.52	0.91	0.89
8	0.82	0.91	0.35	0.79	0.85
9	0.88	0.83	0.95	0.61	0.83
10	0.78	0.74	0.93	0.56	0.81
11	0.77	0.71	0.71	0.95	0.81
12	0.84	0.81	0.80	0.86	0.80
13	0.75	0.69	0.62	0.93	0.65
14	0.58	0.60	0.53	0.91	0.73
15	0.55	0.64	0.94	0.87	0.59
16	0.56	0.31	0.81	0.54	0.48
17	0.55	0.55	0.69	0.92	0.55
18	0.56	0.64	0.81	0.94	0.41
19	0.31	0.35	0.72	0.56	0.43
20	0.38	0.58	0.66	0.31	0.32

* These p-values are based on spring standardization data.

Table 12 (continued)

Item Difficulties (p-values)*

Grade 4 (Level 2)

Item	SQ	AN	QR	VW	VC
1	0.89	0.92	0.76	0.81	0.96
2	0.95	0.85	0.43	0.52	0.86
3	0.89	0.94	0.49	0.57	0.82
4	0.85	0.80	0.37	0.39	0.82
5	0.87	0.92	0.36	0.70	0.87
6	0.86	0.93	0.31	0.78	0.91
7	0.86	0.87	0.77	0.59	0.85
8	0.77	0.85	0.77	0.65	0.85
9	0.66	0.60	0.69	0.70	0.84
10	0.43	0.86	0.52	0.66	0.82
11	0.60	0.70	0.57	0.72	0.59
12	0.64	0.64	0.27	0.89	0.65
13	0.59	0.73	0.88	0.83	0.57
14	0.54	0.37	0.85	0.81	0.50
15	0.58	0.51	0.82	0.85	0.56
16	0.50	0.48	0.79	0.76	0.47
17	0.38	0.49	0.73	0.26	0.32
18	0.27	0.45	0.65	0.76	0.52
19	0.42	0.33	0.55	0.25	0.41
20	0.21	0.47	0.54	0.27	0.26

Grade 5 (Level 2)

Item	SQ	AN	QR	VW	VC
1	0.92	0.93	0.80	0.85	0.97
2	0.97	0.89	0.49	0.59	0.89
3	0.93	0.96	0.55	0.65	0.86
4	0.90	0.85	0.43	0.46	0.86
5	0.91	0.95	0.41	0.75	0.91
6	0.90	0.96	0.36	0.82	0.94
7	0.91	0.91	0.82	0.67	0.89
8	0.84	0.90	0.83	0.72	0.89
9	0.74	0.70	0.76	0.77	0.88
10	0.52	0.91	0.59	0.72	0.87
11	0.70	0.76	0.64	0.78	0.67
12	0.71	0.72	0.30	0.92	0.73
13	0.69	0.77	0.91	0.87	0.63
14	0.61	0.45	0.88	0.85	0.55
15	0.66	0.58	0.86	0.88	0.62
16	0.57	0.56	0.83	0.81	0.54
17	0.45	0.56	0.79	0.31	0.37
18	0.35	0.49	0.71	0.80	0.57
19	0.48	0.39	0.62	0.29	0.47
20	0.26	0.54	0.59	0.32	0.29

* These p-values are based on spring standardization data.

Table 12 (continued)

Item Difficulties (p-values)*

Grade 6 (Level 3)

Item	SQ	AN	QR	VW	VC
1	0.92	0.94	0.67	0.91	0.85
2	0.86	0.94	0.76	0.87	0.87
3	0.75	0.88	0.47	0.75	0.73
4	0.89	0.89	0.57	0.75	0.91
5	0.82	0.85	0.32	0.73	0.92
6	0.76	0.74	0.25	0.70	0.88
7	0.78	0.79	0.88	0.26	0.93
8	0.78	0.72	0.83	0.89	0.86
9	0.64	0.60	0.85	0.92	0.86
10	0.67	0.82	0.85	0.81	0.87
11	0.68	0.71	0.52	0.66	0.64
12	0.64	0.36	0.30	0.63	0.58
13	0.52	0.71	0.73	0.46	0.44
14	0.66	0.66	0.69	0.80	0.40
15	0.57	0.45	0.87	0.75	0.59
16	0.43	0.54	0.64	0.73	0.33
17	0.39	0.36	0.83	0.64	0.53
18	0.40	0.43	0.37	0.44	0.36
19	0.46	0.55	0.38	0.37	0.27
20	0.46	0.27	0.29	0.28	0.35

Grade 7 (Level 3)

Item	SQ	AN	QR	VW	VC
1	0.93	0.95	0.70	0.93	0.87
2	0.88	0.95	0.78	0.89	0.88
3	0.77	0.90	0.51	0.78	0.75
4	0.91	0.91	0.60	0.79	0.91
5	0.84	0.88	0.35	0.78	0.93
6	0.78	0.77	0.27	0.76	0.89
7	0.80	0.82	0.89	0.28	0.93
8	0.80	0.75	0.85	0.91	0.87
9	0.67	0.64	0.87	0.93	0.87
10	0.70	0.85	0.87	0.84	0.88
11	0.71	0.73	0.56	0.71	0.67
12	0.67	0.38	0.34	0.66	0.62
13	0.56	0.74	0.76	0.48	0.48
14	0.69	0.69	0.72	0.83	0.44
15	0.61	0.50	0.88	0.79	0.62
16	0.46	0.57	0.68	0.76	0.37
17	0.42	0.39	0.85	0.69	0.54
18	0.41	0.47	0.42	0.50	0.40
19	0.50	0.58	0.42	0.42	0.29
20	0.49	0.30	0.33	0.32	0.37

* These p-values are based on spring standardization data.

Table 12 (continued)

Item Difficulties (p-values)*

Grade 8 (Level 4)

Item	SQ	AN	QR	VW	VC
1	0.88	0.91	0.83	0.71	0.91
2	0.84	0.77	0.85	0.76	0.92
3	0.84	0.90	0.71	0.67	0.94
4	0.78	0.65	0.66	0.44	0.93
5	0.81	0.83	0.49	0.46	0.80
6	0.76	0.82	0.49	0.18	0.72
7	0.82	0.84	0.87	0.82	0.91
8	0.73	0.40	0.38	0.79	0.85
9	0.78	0.63	0.67	0.52	0.73
10	0.75	0.71	0.52	0.74	0.89
11	0.70	0.74	0.69	0.66	0.54
12	0.53	0.50	0.29	0.49	0.69
13	0.66	0.58	0.88	0.85	0.72
14	0.41	0.53	0.52	0.83	0.47
15	0.44	0.67	0.57	0.82	0.50
16	0.53	0.52	0.62	0.71	0.42
17	0.48	0.47	0.42	0.62	0.40
18	0.53	0.50	0.42	0.69	0.33
19	0.36	0.43	0.43	0.52	0.46
20	0.46	0.50	0.33	0.47	0.24

Grade 9 (Level 4)

Item	SQ	AN	QR	VW	VC
1	0.89	0.92	0.84	0.73	0.92
2	0.85	0.79	0.86	0.77	0.93
3	0.85	0.91	0.73	0.70	0.95
4	0.79	0.67	0.68	0.47	0.94
5	0.82	0.85	0.51	0.49	0.81
6	0.77	0.84	0.52	0.19	0.73
7	0.83	0.85	0.87	0.84	0.92
8	0.75	0.41	0.40	0.81	0.87
9	0.79	0.65	0.69	0.55	0.75
10	0.77	0.72	0.54	0.76	0.91
11	0.72	0.76	0.71	0.69	0.56
12	0.56	0.52	0.31	0.51	0.71
13	0.68	0.61	0.89	0.86	0.73
14	0.43	0.56	0.55	0.84	0.49
15	0.47	0.69	0.59	0.84	0.52
16	0.55	0.55	0.64	0.73	0.44
17	0.50	0.49	0.44	0.64	0.42
18	0.55	0.52	0.44	0.72	0.35
19	0.38	0.45	0.45	0.55	0.48
20	0.48	0.51	0.36	0.50	0.25

* These p-values are based on spring standardization data.

Table 12 (continued)

Item Difficulties (p-values)*

Grade 10 (Level 5)

Item	SQ	AN	QR	VW	VC
1	0.88	0.92	0.86	0.93	0.93
2	0.89	0.75	0.92	0.90	0.86
3	0.79	0.79	0.86	0.65	0.87
4	0.82	0.85	0.81	0.67	0.85
5	0.78	0.82	0.67	0.37	0.86
6	0.79	0.65	0.31	0.85	0.86
7	0.68	0.64	0.85	0.82	0.79
8	0.58	0.70	0.60	0.88	0.81
9	0.73	0.72	0.79	0.59	0.73
10	0.72	0.68	0.54	0.13	0.65
11	0.59	0.54	0.58	0.70	0.64
12	0.62	0.61	0.36	0.38	0.65
13	0.58	0.55	0.82	0.76	0.66
14	0.51	0.60	0.73	0.84	0.56
15	0.48	0.63	0.42	0.77	0.53
16	0.55	0.49	0.32	0.35	0.41
17	0.40	0.43	0.40	0.68	0.44
18	0.53	0.28	0.45	0.61	0.40
19	0.41	0.57	0.34	0.61	0.26
20	0.38	0.37	0.38	0.36	0.30

Grade 11 (Level 5)

Item	SQ	AN	QR	VW	VC
1	0.88	0.93	0.87	0.94	0.93
2	0.90	0.77	0.92	0.92	0.87
3	0.80	0.80	0.87	0.68	0.88
4	0.83	0.86	0.82	0.70	0.86
5	0.79	0.83	0.68	0.40	0.87
6	0.80	0.67	0.32	0.87	0.88
7	0.68	0.66	0.86	0.83	0.81
8	0.58	0.71	0.61	0.90	0.83
9	0.73	0.73	0.80	0.63	0.76
10	0.72	0.70	0.55	0.14	0.68
11	0.59	0.56	0.59	0.73	0.66
12	0.63	0.63	0.37	0.41	0.68
13	0.59	0.58	0.83	0.79	0.69
14	0.52	0.61	0.74	0.86	0.59
15	0.49	0.65	0.43	0.79	0.56
16	0.56	0.51	0.34	0.39	0.43
17	0.40	0.45	0.41	0.72	0.46
18	0.54	0.30	0.46	0.64	0.43
19	0.41	0.59	0.35	0.65	0.27
20	0.39	0.39	0.39	0.40	0.31

* These p-values are based on spring standardization data.

Table 12 (concluded)

Item Difficulties (p-values)*

Grade 11 (Level 6)

Item	SQ	AN	QR	VW	VC
1	0.98	0.90	0.88	0.55	0.86
2	0.88	0.88	0.82	0.54	0.90
3	0.82	0.84	0.60	0.51	0.86
4	0.78	0.72	0.28	0.63	0.88
5	0.78	0.80	0.41	0.79	0.78
6	0.64	0.40	0.14	0.76	0.78
7	0.67	0.77	0.81	0.62	0.81
8	0.73	0.85	0.77	0.83	0.74
9	0.61	0.76	0.59	0.72	0.82
10	0.72	0.56	0.51	0.69	0.61
11	0.67	0.66	0.45	0.44	0.52
12	0.53	0.40	0.42	0.28	0.33
13	0.62	0.52	0.33	0.91	0.47
14	0.49	0.53	0.22	0.59	0.48
15	0.58	0.34	0.70	0.57	0.51
16	0.37	0.43	0.68	0.54	0.40
17	0.58	0.41	0.60	0.50	0.52
18	0.36	0.39	0.60	0.27	0.43
19	0.37	0.28	0.36	0.39	0.27
20	0.29	0.43	0.64	0.36	0.29

Grade 12 (Level 6)

Item	SQ	AN	QR	VW	VC
1	0.98	0.90	0.89	0.57	0.87
2	0.89	0.89	0.84	0.57	0.91
3	0.83	0.85	0.62	0.53	0.87
4	0.79	0.73	0.29	0.66	0.89
5	0.79	0.81	0.43	0.83	0.79
6	0.66	0.41	0.15	0.80	0.79
7	0.69	0.77	0.83	0.66	0.82
8	0.75	0.86	0.79	0.87	0.75
9	0.63	0.77	0.61	0.76	0.83
10	0.74	0.57	0.52	0.73	0.63
11	0.68	0.67	0.47	0.46	0.53
12	0.55	0.41	0.43	0.30	0.33
13	0.64	0.53	0.34	0.93	0.48
14	0.51	0.53	0.23	0.62	0.50
15	0.60	0.35	0.71	0.60	0.52
16	0.40	0.44	0.70	0.56	0.41
17	0.59	0.42	0.61	0.52	0.53
18	0.38	0.40	0.61	0.29	0.43
19	0.38	0.28	0.37	0.40	0.28
20	0.31	0.44	0.66	0.38	0.29

* These p-values are based on spring standardization data.

Table 13

InView Scale Score Descriptive Statistics*

	Grade 2 (Level 1)				Grade 3 (Level1)			
Test	N	Mean	SD	Median	N	Mean	SD	Median
Sequences	11,407	367	69.2	372	10,083	403	66.0	404
Analogies	11,410	390	69.3	392	10,079	425	68.5	428
Quantitative Reasoning	11,411	402	51.4	405	10,050	427	51.5	427
Verbal Reasoning—Words	11,388	372	66.8	376	10,064	411	65.0	413
Verbal Reasoning—Context	11,424	397	66.8	401	10,098	434	64.9	438
	Grade 4 (Level 2)				**Grade 5 (Level 2)**			
Test	N	Mean	SD	Median	N	Mean	SD	Median
Sequences	10,301	436	64.9	437	9,654	459	65.2	460
Analogies	10,301	441	65.2	444	9,653	460	61.9	462
Quantitative Reasoning	10,250	453	51.9	457	9,654	467	53.1	472
Verbal Reasoning—Words	10,278	434	59.5	434	9,651	453	63.8	452
Verbal Reasoning—Context	10,294	451	63.1	455	9,648	467	62.6	471
	Grade 6 (Level 3)				**Grade 7 (Level 3)**			
Test	N	Mean	SD	Median	N	Mean	SD	Median
Sequences	10,223	479	62.8	480	10,247	488	64.5	489
Analogies	10,202	479	64.4	484	10,232	488	63.7	493
Quantitative Reasoning	10,149	483	58.8	485	10,179	493	61.7	495
Verbal Reasoning—Words	10,201	477	69.7	477	10,218	489	69.8	493
Verbal Reasoning—Context	10,200	481	63.5	485	10,229	489	66.1	493

* These statistics are based on the spring 2001 normative data.

Table 13 (continued)

InView Scale Score Descriptive Statistics*

	Grade 8 (Level 4)				Grade 9 (Level 4)			
Test	N	Mean	SD	Median	N	Mean	SD	Median
Sequences	9,624	505	63.8	508	11,051	510	65.3	513
Analogies	9,611	502	58.5	506	11,034	507	59.1	512
Quantitative Reasoning	9,583	506	62.4	509	10,832	512	64.6	515
Verbal Reasoning—Words	9,598	504	63.9	504	10,987	510	65.0	512
Verbal Reasoning—Context	9,607	503	67.3	508	11,012	509	67.3	511
	Grade 10 (Level 5)				**Grade 11 (Levels 5 & 6)**			
Test	N	Mean	SD	Median	N	Mean	SD	Median
Sequences	9,735	518	64.6	521	5,921	521	65.4	523
Analogies	9,732	517	60.4	522	5,897	522	60.9	528
Quantitative Reasoning	9,665	521	66.5	525	5,779	526	68.3	528
Verbal Reasoning—Words	9,707	529	69.2	532	5,825	539	71.4	543
Verbal Reasoning—Context	9,724	524	67.6	527	5,863	531	67.5	536
	Grade 12 (Level 6)							
Test	N	Mean	SD	Median				
Sequences	6,354	526	66.5	529				
Analogies	6,359	525	60.2	530				
Quantitative Reasoning	6,294	532	66.3	532				
Verbal Reasoning—Words	6,311	549	66.3	550				
Verbal Reasoning—Context	6,334	535	66.0	539				

* These statistics are based on the spring 2001 normative data.

Part 7: Test Fairness and Differential Item Functioning

General Principles

The position of CTB/McGraw-Hill concerning test fairness is based on several general propositions. First, students may differ in their background knowledge, educational experiences, cognitive and academic skills, language, attitudes, and values. To the degree that these differences are large, no single test will be equally appropriate for all students. Furthermore, it is difficult to specify what amount of difference can be called *large,* and to determine how these differences will affect the outcome of a particular test.

Second, because test scores can have many sources of variation, the test publishers' task is to develop assessments that measure the intended abilities and skills without introducing extraneous elements or constructing irrelevant elements. When tests measure something other than what they are intended to measure, test scores will reflect these unintended skills and knowledge, as well as what is purportedly assessed by the test. If this occurs, these tests can be called biased (Angoff, 1993; Camilli & Shepard, 1994; Green, 1975). Factors that may render test scores to be biased include a lack of opportunities to learn and develop the skills and knowledge required, unfamiliarity with test formats, and differing cultural and socioeconomic experiences.

The third general proposition is that for some groups, notably those from families whose language and culture differ sharply from that of "Middle America," no test designed to be used nationally can be completely unbiased. The best alternative is to minimize the role of the extraneous elements, thereby increasing the number of students for whom the test is appropriate. Careful attention is taken in the test construction process to lessen the influence of these elements for large numbers of students. Unfortunately, in some cases these elements may continue to play a substantial role. For example, one cannot conclude that Spanish-instructed students have lower verbal reasoning skills if they perform poorly on the Verbal Reasoning test sections that are written in English, since they may be literate only in Spanish. Only when such students have acquired a substantial degree of facility with the English language can standard tests elicit performances adequately representative of their academic or cognitive level.

Procedures for Minimizing Bias

Four procedures are used at CTB/McGraw-Hill to minimize bias. The first is based on the premise that careful editorial attention to validity is an essential step in keeping bias to a minimum. Bias occurs when a test is measuring different constructs for different groups. If a test entails irrelevant skills or knowledge (however common), the possibility of bias is increased. Thus, careful attention was paid to content validity during the item writing and selection processes of *InView*.

The second procedure is to follow the CTB/McGraw-Hill guidelines designed to reduce or eliminate bias. Item writers are directed to the following published guidelines: *Guidelines for Bias-Free Publishing* (McGraw-Hill, 1983) and *Reflecting Diversity: Multicultural Guidelines for Educational Publishing Professionals* (Macmillan/McGraw-Hill, 1993). These guidelines were used when *InView* materials were reviewed. Internal editorial reviews were done by at least four separate people: the content editor, who directly supervises the item writers; the project director; a style editor; and a proofreader. The final tests built from the tryout materials were again reviewed by at least five people including quality assurance staff.

In the third procedure, educational professionals from various ethnic groups review all tryout materials. They consider and comment on the appropriateness of language, subject matter, and representation of groups of people. Editorial professionals provided many useful suggestions for reducing the bias of *InView.*

The first three procedures are believed to improve the quality of *InView* and reduce item and test bias. However, current evidence suggests that expertise in this area is no substitute for statistical analyses. Reviewers may not know which items perform differently between specific subgroups of students, apparently because some of their ideas about how students will react to items may be inaccurate (Camilli & Shepard, 1994; Jensen, 1980; Sandoval & Mille, 1979; Scheuneman, 1984).

Thus, an empirical differential item functioning (DIF) approach was also used as the fourth procedure to identify potential sources of item bias. DIF studies include a systematic item analysis to determine if examinees with the same underlying level of ability have the same probability of getting the item correct. Items identified with DIF are then examined to determine if item performance differences between identifiable subgroups of the population are due to extraneous or construct irrelevant information, making the items unfairly difficult for one of the subgroups. The inclusion of these items are minimized in the test development process. DIF studies have been routinely done for all major test batteries published by CTB/McGraw-Hill since 1970.

Analysis of Tryout Data

DIF of the *InView* tryout items was assessed for students identified as African-Americans or Hispanic-Americans. Analyses of gender DIF were also conducted.

Because *InView* was built using item response theory, DIF analyses that capitalized on the information and item statistics provided by this theory were implemented. There are several IRT-based DIF procedures including those that assess the equality of item parameters across groups (Lord, 1980), and those that assess area differences between item characteristic curves (Linn, Levine, Hastings, & Wardrop, 1981; Camilli & Shepard, 1994). However, these procedures require a minimum of 800 to 1,000 cases in each group to produce reliable and consistent results. In contrast, the Linn-Harnisch procedure (Linn & Harnisch, 1981) utilizes the information provided by the three-parameter IRT model but requires fewer cases. This was the procedure used to complete the ethnic and gender DIF studies for the tryout data.

An example of this procedure for ethnic DIF analyses follows. The parameters for each item (a_i, b_i, c_i) and the trait or scale score (θ) for each examinee are estimated for the three-parameter logistic model:

$$P_{ij} = \hat{c}_i + \frac{1 - \hat{c}_i}{1 + \exp\left[-1.7\hat{a}_i\left(\hat{\theta}_j - \hat{b}_i\right)\right]},$$

where P_{ij} is the probability that examinee j will answer item i correctly. Note that the item parameters are based on data from the total sample of examinees, which includes all categories (African-American, Hispanic-American, and Standard) in the tryout sample. The sample is then divided into ethnic groups, and the members in each group are sorted into ten equal score categories (deciles) based upon their location on the scale-score (θ) scale. The expected proportion correct for each group based on the model prediction is compared to the observed (actual) proportion correct obtained by the group.

The proportion of examinees in decile g who are expected to answer item i correctly is

$$P_{ig} = \frac{1}{n_g}\sum_{j\varepsilon g} P_{ij},$$

where n_g is the number of examinees in decile g. To compute the proportion of examinees expected to answer item i correctly (over all deciles) for a group (e.g., African-American) the formula is given by:

$$P_{i\cdot} = \frac{\sum_{g=1}^{10} n_g P_{ig}}{\sum_{g=1}^{10} n_g}.$$

The corresponding observed proportion correct for examinees in a decile (O_{ig}) is the number of examinees in decile g who answered item i correctly divided by the number of people in the decile (n_g). That is,

$$O_{ig} = \frac{\sum_{j \varepsilon g} u_{ij}}{n_g},$$

where u_{ij} is the dichotomous score for item i for examinee j.

The corresponding formula to compute the observed proportion answering each item correctly (over all deciles) for a complete ethnic group is given by:

$$O_i = \frac{\sum_{g=1}^{10} n_g O_{ig}}{\sum_{g=1}^{10} n_g}.$$

After the values are calculated for these variables, the difference between the observed proportion correct (for an ethnic group) and expected proportion correct can be computed. The decile group difference (D_{ig}) for observed and expected proportion correctly answering item i in decile g is

$$D_{ig} = O_{ig} - P_{ig},$$

and the overall group difference (D_i) between observed and expected proportion correct for item i in the complete group (over all deciles) is

$$D_i = O_i - P_i.$$

These indices are indicators of the degree to which members of an ethnic group perform better or worse than expected on each item, based on the parameter estimates from all subsamples. Differences for decile groups provide an index for each of the ten regions on the scale-score (θ) scale. The decile group difference (D_{ig}) can be either positive or negative. Use of the decile group differences as well as the overall group difference allows one to detect items that give a large positive difference in one range of θ and a large negative difference in another range of θ, yet have a small overall difference.

DIF Ratings

DIF is defined in terms of the decile group and total target subsample differences, the D_{i-} (sum of the negative group differences) and D_{i+} (sum of the positive group difference) values, and the corresponding standardized differences (Z_i) for the sample. (See Linn & Harnisch, 1981, p. 112.)

Items for which $|D_i| \geq 0.10$ and $|Z_i| \geq 2.58$ are flagged as DIF items. If D_i is positive, the item favors the target subsample. If D_i is negative, the item favors the standard sample.

The ethnic DIF ratings are on a scale from 1 (no DIF) to 5 (DIF against both African-Americans and Hispanic-Americans). The gender DIF ratings are on a scale from 1 (no DIF) to 4 (DIF in favor of one group and against the other).

The ethnic and gender rating schemes are as follows:

	African-American Subsample		
Hispanic-American Subsample	Against	In Favor	No DIF
Against	5	4	3
In Favor	4	3	2
No DIF	3	2	1

	Male Subsample		
Female Subsample	Against	In Favor	No DIF
Against	—	4	3
In Favor	4	—	2
No DIF	3	2	1

These ethnic and gender ratings were included in the item information used for test selection. The average pool DIF and the average DIF for the tests were available to the developers as they compared various versions of each test. Developers strove to produce tests with minimal DIF.

Standardization DIF Study

Students in the standardization study identified themselves as either Asian-American, African-American, Hispanic-American, or White/Caucasian (the standard sample). The performance of students so identified was used for the bias analysis of the standardization edition. The analysis procedures and criteria were the same as those used for the tryout.

Following the procedure used for the tryouts, the difference between expected and actual performance on an item was obtained for the various groups of students. As before, the expected performance was based on the scores of the students in the subgroup and the item parameters derived from the total sample of students taking the item. The results of this procedure are found in Tables 14–17. Tables 14 and 15 present the results for the ethnic groups. Tables 16 and 17 present the results for gender.

In general, the data presented in these tables show minimal gender and ethnic bias in these tests. In evaluating these results, it is important to consider that items in favor of a particular group in combination with items against the same group can result in DIF cancellation, which reduces the overall bias at the test level (Shealy & Stout, 1993).

Scale Score Descriptive Statistics for Subgroups

For information related to test fairness, Table 18 shows by grade and test level the number of students in the sample (N), mean scale scores (Mean), and scale score standard deviations (SD) for African-American, Hispanic-American, and standard sample. In considering these ethnic data, one should note that ethnicity was not a consideration in the sampling design, and no special weighting of the data was done. Therefore, these data are a picture of the sample, and are not subgroup norms. Table 19 shows, by grade and test level, the number of students in the sample (N), mean scale scores (Mean), and scale score standard deviations (SD) for males and females.

In general, the African-American and Hispanic-American samples had slightly lower average scores than did the students in the standard sample. The standard deviations presented in these tables, however, suggest that the distributions of scores for the three ethnic groups overlap. In other words, there are considerable variations in performance within each group, as well as some differences in performance between the groups. In terms of gender differences, the males and females had similar scores on the *InView* subtests. Females tended to score slightly higher than did males on the verbal subtests, which is not unusual. As with the ethnic data, the gender data shows as much or more variation within gender groups as between gender groups.

Table 14

Number of Standardization Items Flagged for DIF in Favor of (+) and Against (–) Each Ethnic Group by Level and Subtest

Test	Number of Items	Asian-American +	Asian-American –	Asian-American Total	African-American +	African-American –	African-American Total	Hispanic-American +	Hispanic-American –	Hispanic-American Total	Standard +	Standard –	Standard Total
Level 1													
Sequences	20	0	0	0	0	0	0	0	0	0	0	0	0
Analogies	20	1	1	2	0	0	0	0	0	0	0	0	0
Quantitative Reasoning	20	1	0	1	0	0	0	0	0	0	0	0	0
Verbal Reasoning—Words	20	1	1	2	0	1	1	0	1	1	0	0	0
Verbal Reasoning—Context	20	0	1	1	0	0	0	0	0	0	0	0	0
All Subtests	100	3	3	6	0	1	1	0	1	1	0	0	0
Level 2													
Sequences	20	0	0	0	0	0	0	0	0	0	0	0	0
Analogies	20	0	0	0	0	0	0	0	0	0	0	0	0
Quantitative Reasoning	20	0	0	0	0	0	0	0	0	0	0	0	0
Verbal Reasoning—Words	20	0	0	0	0	0	0	0	0	0	0	0	0
Verbal Reasoning—Context	20	0	0	0	0	0	0	0	0	0	0	0	0
All Subtests	100	0	0	0	0	0	0	0	0	0	0	0	0
Level 3													
Sequences	20	0	0	0	0	0	0	0	0	0	0	0	0
Analogies	20	0	0	0	0	0	0	0	0	0	0	0	0
Quantitative Reasoning	20	0	0	0	0	0	0	0	0	0	0	0	0
Verbal Reasoning—Words	20	0	0	0	0	0	0	0	0	0	0	0	0
Verbal Reasoning—Context	20	0	0	0	0	0	0	0	0	0	0	0	0
All Subtests	100	0	0	0	0	0	0	0	0	0	0	0	0
Level 4													
Sequences	20	0	0	0	0	0	0	0	0	0	0	0	0
Analogies	20	1	0	1	0	0	0	0	0	0	0	0	0
Quantitative Reasoning	20	0	0	0	0	0	0	0	0	0	0	0	0
Verbal Reasoning—Words	20	1	0	1	0	0	0	0	0	0	0	0	0
Verbal Reasoning—Context	20	0	0	0	0	0	0	0	0	0	0	0	0
All Subtests	100	2	0	2	0	0	0	0	0	0	0	0	0
Level 5													
Sequences	20	0	0	0	0	0	0	0	0	0	0	0	0
Analogies	20	0	0	0	0	0	0	0	0	0	0	0	0
Quantitative Reasoning	20	0	0	0	0	0	0	0	0	0	0	0	0
Verbal Reasoning—Words	20	0	0	0	0	0	0	0	0	0	0	0	0
Verbal Reasoning—Context	20	0	0	0	0	0	0	0	0	0	0	0	0
All Subtests	100	0	0	0	0	0	0	0	0	0	0	0	0
Level 6													
Sequences	20	0	0	0	0	0	0	0	0	0	0	0	0
Analogies	20	0	0	0	0	0	0	0	0	0	0	0	0
Quantitative Reasoning	20	0	0	0	0	0	0	0	0	0	0	0	0
Verbal Reasoning—Words	20	0	0	0	0	0	0	0	0	0	0	0	0
Verbal Reasoning—Context	20	0	0	0	0	0	0	0	0	0	0	0	0
All Subtests	100	0	0	0	0	0	0	0	0	0	0	0	0

Table 15

Number of Standardization Items Flagged for DIF in Favor of (+) and Against (−) Each Ethnic Group by Subtest and Level

Level	Number of Items	Asian-American +	Asian-American −	Asian-American Total	African-American +	African-American −	African-American Total	Hispanic-American +	Hispanic-American −	Hispanic-American Total	Standard +	Standard −	Standard Total
Sequences													
1	20	0	0	0	0	0	0	0	0	0	0	0	0
2	20	0	0	0	0	0	0	0	0	0	0	0	0
3	20	0	0	0	0	0	0	0	0	0	0	0	0
4	20	0	0	0	0	0	0	0	0	0	0	0	0
5	20	0	0	0	0	0	0	0	0	0	0	0	0
6	20	0	0	0	0	0	0	0	0	0	0	0	0
All Levels	120	0	0	0	0	0	0	0	0	0	0	0	0
Analogies													
1	20	1	1	2	0	0	0	0	0	0	0	0	0
2	20	0	0	0	0	0	0	0	0	0	0	0	0
3	20	0	0	0	0	0	0	0	0	0	0	0	0
4	20	1	0	1	0	0	0	0	0	0	0	0	0
5	20	0	0	0	0	0	0	0	0	0	0	0	0
6	20	0	0	0	0	0	0	0	0	0	0	0	0
All Levels	120	2	1	3	0	0	0	0	0	0	0	0	0
Quantitative Reasoning													
1	20	1	0	1	0	0	0	0	0	0	0	0	0
2	20	0	0	0	0	0	0	0	0	0	0	0	0
3	20	0	0	0	0	0	0	0	0	0	0	0	0
4	20	0	0	0	0	0	0	0	0	0	0	0	0
5	20	0	0	0	0	0	0	0	0	0	0	0	0
6	20	0	0	0	0	0	0	0	0	0	0	0	0
All Levels	120	1	0	1	0	0	0	0	0	0	0	0	0
Verbal Reasoning—Words													
1	20	1	1	2	0	1	1	0	1	1	0	0	0
2	20	0	0	0	0	0	0	0	0	0	0	0	0
3	20	0	0	0	0	0	0	0	0	0	0	0	0
4	20	1	0	1	0	0	0	0	0	0	0	0	0
5	20	0	0	0	0	0	0	0	0	0	0	0	0
6	20	0	0	0	0	0	0	0	0	0	0	0	0
All Levels	120	2	1	3	0	1	1	0	1	1	0	0	0
Verbal Reasoning—Context													
1	20	0	1	1	0	0	0	0	0	0	0	0	0
2	20	0	0	0	0	0	0	0	0	0	0	0	0
3	20	0	0	0	0	0	0	0	0	0	0	0	0
4	20	0	0	0	0	0	0	0	0	0	0	0	0
5	20	0	0	0	0	0	0	0	0	0	0	0	0
6	20	0	0	0	0	0	0	0	0	0	0	0	0
All Levels	120	0	1	1	0	0	0	0	0	0	0	0	0

Table 16

Number of Standardization Items Flagged for DIF in Favor of (+) and Against (−) Each Gender Group by Level and Subtest

Test	Number of Items	Male +	Male −	Male Total	Female +	Female −	Female Total
Level 1							
Sequences	20	0	0	0	0	0	0
Analogies	20	0	0	0	0	0	0
Quantitative Reasoning	20	0	0	0	0	0	0
Verbal Reasoning—Words	20	0	0	0	0	0	0
Verbal Reasoning—Context	20	0	0	0	0	0	0
All Subtests	100	0	0	0	0	0	0
Level 2							
Sequences	20	0	0	0	0	0	0
Analogies	20	0	0	0	0	0	0
Quantitative Reasoning	20	0	0	0	0	0	0
Verbal Reasoning—Words	20	0	0	0	0	0	0
Verbal Reasoning—Context	20	0	0	0	0	0	0
All Subtests	100	0	0	0	0	0	0
Level 3							
Sequences	20	0	0	0	0	0	0
Analogies	20	0	0	0	0	0	0
Quantitative Reasoning	20	0	0	0	0	0	0
Verbal Reasoning—Words	20	0	0	0	0	0	0
Verbal Reasoning—Context	20	0	0	0	0	0	0
All Subtests	100	0	0	0	0	0	0
Level 4							
Sequences	20	0	0	0	0	0	0
Analogies	20	0	0	0	0	0	0
Quantitative Reasoning	20	0	0	0	0	0	0
Verbal Reasoning—Words	20	0	0	0	0	0	0
Verbal Reasoning—Context	20	0	0	0	0	0	0
All Subtests	100	0	0	0	0	0	0
Level 5							
Sequences	20	0	0	0	0	0	0
Analogies	20	0	0	0	0	0	0
Quantitative Reasoning	20	0	0	0	0	0	0
Verbal Reasoning—Words	20	0	0	0	0	0	0
Verbal Reasoning—Context	20	0	0	0	0	0	0
All Subtests	100	0	0	0	0	0	0
Level 6							
Sequences	20	0	0	0	0	0	0
Analogies	20	1	0	1	0	1	1
Quantitative Reasoning	20	0	0	0	0	0	0
Verbal Reasoning—Words	20	0	0	0	0	0	0
Verbal Reasoning—Context	20	0	0	0	0	0	0
All Subtests	100	1	0	1	0	1	1

Table 17

Number of Standardization Items Flagged for DIF in Favor of (+) and Against (−) Each Gender Group by Subtest and Level

Level	Number of Items	Male +	Male −	Male Total	Female +	Female −	Female Total
Sequences							
1	20	0	0	0	0	0	0
2	20	0	0	0	0	0	0
3	20	0	0	0	0	0	0
4	20	0	0	0	0	0	0
5	20	0	0	0	0	0	0
6	20	0	0	0	0	0	0
All Levels	120	0	0	0	0	0	0
Analogies							
1	20	0	0	0	0	0	0
2	20	0	0	0	0	0	0
3	20	0	0	0	0	0	0
4	20	0	0	0	0	0	0
5	20	0	0	0	0	0	0
6	20	1	0	1	0	1	1
All Levels	120	1	0	1	0	1	1
Quantitative Reasoning							
1	20	0	0	0	0	0	0
2	20	0	0	0	0	0	0
3	20	0	0	0	0	0	0
4	20	0	0	0	0	0	0
5	20	0	0	0	0	0	0
6	20	0	0	0	0	0	0
All Levels	120	0	0	0	0	0	0
Verbal Reasoning—Words							
1	20	0	0	0	0	0	0
2	20	0	0	0	0	0	0
3	20	0	0	0	0	0	0
4	20	0	0	0	0	0	0
5	20	0	0	0	0	0	0
6	20	0	0	0	0	0	0
All Levels	120	0	0	0	0	0	0
Verbal Reasoning—Context							
1	20	0	0	0	0	0	0
2	20	0	0	0	0	0	0
3	20	0	0	0	0	0	0
4	20	0	0	0	0	0	0
5	20	0	0	0	0	0	0
6	20	0	0	0	0	0	0
All Levels	120	0	0	0	0	0	0

Table 18

Means and Standard Deviations of Each Ethnic Group for Each Subtest by Grade and Level

Level	African-American			Hispanic-American			Standard		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
Sequences									
Grade 2, Level 1	1,725	321.2	78.2	765	356.2	68.6	6,749	374.6	65.0
Grade 3, Level 1	1,606	360.1	73.0	618	388.7	66.0	5,034	407.0	63.5
Grade 4, Level 2	1,492	406.6	61.5	805	423.6	59.3	4,645	438.8	64.1
Grade 5, Level 2	1,162	425.0	64.1	906	454.6	62.2	5,080	464.5	63.6
Grade 6, Level 3	1,455	446.7	62.4	1,077	460.7	64.3	5,922	483.0	63.1
Grade 7, Level 3	1,561	455.4	61.0	979	467.4	62.5	4,761	496.5	62.8
Grade 8, Level 4	1,358	484.7	61.2	1,221	489.3	60.6	3,942	511.3	63.9
Grade 9, Level 4	1,746	486.4	63.4	1,062	489.0	64.1	4,815	520.9	66.1
Grade 10, Level 5	1,484	491.6	66.0	723	495.5	66.1	3,954	524.5	64.7
Grade 11, Level 5, 6	843	505.3	68.5	418	501.7	66.1	2,933	529.5	66.4
Grade 12, Level 6	966	501.6	67.2	395	503.9	66.9	3,438	521.5	67.3
Analogies									
Grade 2, Level 1	1,725	354.0	67.9	765	379.3	68.4	6,749	396.1	66.8
Grade 3, Level 1	1,606	379.5	68.4	618	417.5	72.3	5,034	431.4	69.2
Grade 4, Level 2	1,492	409.4	60.3	805	434.2	56.7	4,645	445.8	61.4
Grade 5, Level 2	1,162	422.6	59.5	906	458.1	58.4	5,080	467.1	60.4
Grade 6, Level 3	1,455	433.9	70.6	1,077	460.9	65.8	5,922	485.5	63.4
Grade 7, Level 3	1,561	449.1	71.6	979	472.7	68.6	4,761	499.6	61.3
Grade 8, Level 4	1,358	467.5	63.0	1,221	483.2	65.0	3,942	510.7	57.6
Grade 9, Level 4	1,746	470.4	62.8	1,062	487.8	59.4	4,815	518.7	58.2
Grade 10, Level 5	1,484	483.8	64.4	723	497.2	63.9	3,954	525.2	61.9
Grade 11, Level 5, 6	843	492.5	66.8	418	508.1	58.6	2,933	529.6	60.5
Grade 12, Level 6	966	492.6	61.5	395	503.2	65.2	3,438	526.9	59.9
Quantitative Reasoning									
Grade 2, Level 1	1,725	378.9	51.7	765	389.8	49.4	6,749	407.3	48.2
Grade 3, Level 1	1,606	401.2	51.8	618	413.1	49.2	5,034	430.3	50.9
Grade 4, Level 2	1,492	425.1	53.5	805	439.3	49.6	4,645	454.6	51.9
Grade 5, Level 2	1,162	435.0	58.3	906	460.0	51.8	5,080	472.9	51.5
Grade 6, Level 3	1,455	456.3	54.5	1,077	458.8	58.9	5,922	490.6	57.3
Grade 7, Level 3	1,561	458.8	57.8	979	463.8	60.6	4,761	500.3	60.8
Grade 8, Level 4	1,358	471.6	58.8	1,221	482.0	64.3	3,942	512.8	60.7
Grade 9, Level 4	1,746	474.1	64.2	1,062	484.5	64.8	4,815	523.7	63.2
Grade 10, Level 5	1,484	484.5	68.4	723	500.6	66.4	3,954	531.7	68.3
Grade 11, Level 5, 6	843	496.4	70.8	418	509.8	67.3	2,933	538.2	65.9
Grade 12, Level 6	966	496.4	61.8	395	510.4	60.9	3,438	530.7	65.6

Table 18 (concluded)

Means and Standard Deviations of Each Ethnic Group for Each Subtest by Grade and Level

Level	African-American			Hispanic-American			Standard		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
Verbal Reasoning—Words									
Grade 2, Level 1	1,725	344.0	63.4	765	350.1	71.3	6,749	381.7	63.3
Grade 3, Level 1	1,606	377.5	59.6	618	381.4	64.8	5,034	418.4	63.4
Grade 4, Level 2	1,492	411.7	54.2	805	411.7	63.5	4,645	440.7	59.8
Grade 5, Level 2	1,162	426.7	53.5	906	428.7	57.9	5,080	462.4	62.3
Grade 6, Level 3	1,455	435.3	62.8	1,077	447.7	67.6	5,922	487.7	67.9
Grade 7, Level 3	1,561	445.9	65.1	979	458.4	70.5	4,761	506.0	69.2
Grade 8, Level 4	1,358	476.3	55.9	1,221	477.2	59.4	3,942	514.4	62.7
Grade 9, Level 4	1,746	483.3	60.4	1,062	482.4	59.4	4,815	525.5	64.9
Grade 10, Level 5	1,484	494.7	65.4	723	498.0	71.2	3,954	542.7	72.3
Grade 11, Level 5, 6	843	510.0	70.0	418	508.0	67.0	2,933	552.8	69.8
Grade 12, Level 6	966	513.4	66.8	395	528.6	66.9	3,438	554.1	70.0
Verbal Reasoning—Context									
Grade 2, Level 1	1,725	365.5	61.3	765	367.7	63.2	6,749	406.9	65.4
Grade 3, Level 1	1,606	401.6	61.1	618	402.1	62.3	5,034	442.3	65.2
Grade 4, Level 2	1,492	426.6	60.0	805	421.1	58.1	4,645	455.5	61.8
Grade 5, Level 2	1,162	437.6	59.1	906	447.2	60.6	5,080	474.9	60.7
Grade 6, Level 3	1,455	450.6	60.9	1,077	452.5	64.3	5,922	488.8	61.1
Grade 7, Level 3	1,561	456.9	63.4	979	457.0	69.0	4,761	498.5	65.6
Grade 8, Level 4	1,358	483.8	52.5	1,221	476.1	59.8	3,942	515.1	64.4
Grade 9, Level 4	1,746	482.9	64.3	1,062	481.9	59.1	4,815	522.8	70.4
Grade 10, Level 5	1,484	495.1	66.4	723	492.5	67.2	3,954	533.0	69.7
Grade 11, Level 5, 6	843	503.3	64.2	418	506.5	66.2	2,933	541.0	68.7
Grade 12, Level 6	966	513.3	61.1	395	514.8	61.3	3,438	540.3	68.8

Table 19

Means and Standard Deviations of Each Gender Group for Each Subtest by Grade and Level

Level	Female N	Female Mean	Female SD	Male N	Male Mean	Male SD
Sequences						
Grade 2, Level 1	4,542	363.9	67.9	4,611	362.4	74.1
Grade 3, Level 1	3,579	395.2	66.5	3,625	395.5	70.6
Grade 4, Level 2	3,433	427.6	61.7	3,466	432.8	66.6
Grade 5, Level 2	3,606	456.1	63.5	3,491	457.6	66.6
Grade 6, Level 3	4,251	473.8	61.8	4,084	474.3	67.5
Grade 7, Level 3	3,510	483.5	62.4	3,660	484.4	67.0
Grade 8, Level 4	3,261	502.0	60.8	3,068	501.7	66.3
Grade 9, Level 4	3,925	509.5	64.1	3,502	508.4	70.3
Grade 10, Level 5	3,155	515.1	65.5	2,792	511.3	68.6
Grade 11, Level 5, 6	2,188	523.0	65.0	1,848	521.8	70.2
Grade 12, Level 6	2,482	518.7	64.0	2,141	513.0	71.6
Analogies						
Grade 2, Level 1	4,542	389.1	67.1	4,611	384.9	70.9
Grade 3, Level 1	3,579	420.6	69.7	3,625	417.4	74.9
Grade 4, Level 2	3,433	435.2	60.3	3,466	438.3	64.3
Grade 5, Level 2	3,606	458.9	60.5	3,491	458.7	64.0
Grade 6, Level 3	4,251	472.3	65.4	4,084	475.1	70.4
Grade 7, Level 3	3,510	483.8	66.2	3,660	486.8	69.7
Grade 8, Level 4	3,261	492.8	62.0	3,068	501.1	63.1
Grade 9, Level 4	3,925	499.9	60.7	3,502	508.0	65.2
Grade 10, Level 5	3,155	512.5	63.1	2,792	511.7	67.2
Grade 11, Level 5, 6	2,188	518.2	61.1	1,848	523.0	66.0
Grade 12, Level 6	2,482	514.5	58.2	2,141	522.0	66.2
Quantitative Reasoning						
Grade 2, Level 1	4,542	399.8	48.8	4,611	401.6	51.4
Grade 3, Level 1	3,579	422.6	50.8	3,625	423.1	53.5
Grade 4, Level 2	3,433	446.0	51.4	3,466	447.3	55.2
Grade 5, Level 2	3,606	465.9	53.0	3,491	464.4	55.9
Grade 6, Level 3	4,251	481.3	55.6	4,084	480.4	62.4
Grade 7, Level 3	3,510	487.6	59.7	3,660	486.0	66.1
Grade 8, Level 4	3,261	498.6	63.3	3,068	499.1	63.9
Grade 9, Level 4	3,925	508.5	67.4	3,502	506.4	67.5
Grade 10, Level 5	3,155	513.0	68.0	2,792	521.6	74.1
Grade 11, Level 5, 6	2,188	524.8	66.0	1,848	530.7	72.0
Grade 12, Level 6	2,482	524.3	63.4	2,141	520.1	68.6

Table 19 (concluded)

Means and Standard Deviations of Each Gender Group for Each Subtest by Grade and Level

Level	Female N	Female Mean	Female SD	Male N	Male Mean	Male SD
Verbal Reasoning—Words						
Grade 2, Level 1	4,542	377.0	63.3	4,611	367.3	68.2
Grade 3, Level 1	3,579	411.0	61.8	3,625	402.2	67.8
Grade 4, Level 2	3,433	434.3	57.3	3,466	428.2	63.6
Grade 5, Level 2	3,606	457.0	60.6	3,491	447.3	64.1
Grade 6, Level 3	4,251	474.1	68.2	4,084	473.4	72.4
Grade 7, Level 3	3,510	488.5	71.6	3,660	485.5	75.4
Grade 8, Level 4	3,261	502.2	59.9	3,068	497.3	66.6
Grade 9, Level 4	3,925	514.0	62.0	3,502	506.2	70.7
Grade 10, Level 5	3,155	530.4	72.3	2,792	521.5	75.2
Grade 11, Level 5, 6	2,188	542.5	69.1	1,848	537.8	75.2
Grade 12, Level 6	2,482	545.5	66.1	2,141	541.8	76.5
Verbal Resoning—Context						
Grade 2, Level 1	4,542	402.3	62.6	4,611	389.9	70.5
Grade 3, Level 1	3,579	435.9	63.1	3,625	424.8	69.2
Grade 4, Level 2	3,433	451.3	57.3	3,466	439.6	67.2
Grade 5, Level 2	3,606	472.3	57.8	3,491	458.0	65.8
Grade 6, Level 3	4,251	483.1	59.5	4,084	472.3	67.6
Grade 7, Level 3	3,510	491.8	62.8	3,660	476.8	73.1
Grade 8, Level 4	3,261	504.3	57.4	3,068	499.1	69.3
Grade 9, Level 4	3,925	511.7	62.6	3,502	504.6	77.9
Grade 10, Level 5	3,155	526.8	64.9	2,792	511.3	76.0
Grade 11, Level 5, 6	2,188	535.9	62.3	1,848	524.7	76.4
Grade 12, Level 6	2,482	539.8	59.5	2,141	524.8	75.4

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, D. C.: AERA.

Angoff, W. (1993). Perspectives on differential item functioning methodology. In P. W. Hotland and H. Warner (Eds.), *Differential Item Functioning* (pp. 3–24). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bentler, P. M. (1998). *EQS for Windows (Version 5.7)* [Computer software]. Encino, CA: Multivariate Software, Inc.

Burket, G. R. (1988). ITEMSYS [Computer program]. Unpublished.

Burket, G. R. (1990). PARMATE [Computer program]. Unpublished.

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items.* Newbury Park, CA: Sage.

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

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40*, 153–193.

CTB Macmillan/McGraw-Hill. (1992). *Primary Test of Cognitive Skills, Technical Bulletin*. Monterey, CA: Author.

CTB Macmillan/McGraw-Hill. (1992). *Test of Cognitive Skills, Second Edition*. Monterey, CA: Author.

CTB Macmillan/McGraw-Hill. (1996). *TerraNova*. Monterey, CA: Author.

CTB/McGraw-Hill. (2002). *Guidelines for inclusive test administration*. Monterey, CA: Author.

CTB/McGraw-Hill. (2001). *TerraNova, The Second Edition*. Monterey, CA: Author.

Family Education Rights and Privacy Act (FERPA) of 1974, 20 U.S.C. 1232 (g).

Green, D. R. (1975, December). *Procedures for assessing bias in achievement tests.* Presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.

Green, D. R., & Yen, W. M. (1983, April). *Number correct versus pattern scoring: Results for ethnic groups.* Presented at the annual meeting of the National Council on Measurement in Education, Montreal. (ERIC Document Reproduction Services No. ED 236 175).

Green, D. R., Yen, W. M., & Burket, G. R. (1989). Experiences in the application of item response theory in test construction. *Applied Measurement in Education, 2,* 297–312.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Hingham, MA: Kluwer-Nijhoff.

Holland, P. W., King, B. F., & Thayer, P. T. (1988). The standard error of equating for the kernal method of equating score distributions. Unpublished manuscript.

Horn, J. L. (1968). Organization of abilities and the development of intelligence. *Psychological Review, 75,* 242-259.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc Theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 53–91). New York: Guilford Press.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Joreskog, K. G. (1969). A general approach to confirmatory factor analysis. *Psychometrika, 34,* 183–202.

Joreskog, K. G., & Sorbom, D. (1983). LISREL V: *Analysis of linear structural relationships by the method of maximum likelihood.* Chicago: International Educational Services.

Linden W. J., van der, & Hambelton, R. K. (Eds.). (1996). Handbook of modern item response theory. Amherst, MA: University of Massachusetts.

Linn, R. L., & Harnisch, D. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18,* 109–118.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5,* 159–173.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39,* 247–264.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (pp. 71, 179–181). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Macmillan/McGraw-Hill. (1993). *Reflecting diversity: Multicultural guidelines for educational publishing professionals.* New York: Author.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*, 391–410.

McGraw-Hill. (1983). *Guidelines for bias-free publishing.* Monterey, CA: Author.

Sandoval, J. H., & Mille, M. P. (1979, August). *Accuracy of judgments of WISC-R item difficulty for minority groups.* Paper presented at the annual meeting of the American Psychological Association, New York.

Scheuneman, J. D. (1984). A theoretical framework for the exploration of causes and effects of bias in testing. *Educational Psychologist, 19,* 219–225.

Schumaker, R. E., & Lomax, R. G. (1996). A beginner's guide to structural equation modeling. Mahwah, NJ: Lawrence Erlbaum Associates.

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–239). Hillsdale, NJ: Lawrence Erlbaum Associates.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* New York: Macmillan. [Reprinted: New York: AMS Publishers, 1981].

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245–262.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21,* 93–111.

Yen, W. M., & Candell, G. L. (1991). Increasing score reliability with item-pattern scoring: An empirical study in five score metrics. *Applied Measurement in Education, 4,* 209–228.

APPENDIX: *IN VIEW* STATISTICS

Estimated Raw Score Descriptive Statistics

Item response theory was used to obtain descriptive statistics for each level of *InView*. The procedures for obtaining descriptive statistics involved using each test's item parameters and the normative distribution of scale scores at the target grade.

For selected-response items, let $\hat{a}_i$, $\hat{b}_i$ and $\hat{c}_i$ be the item discrimination, item difficulty, and item lower asymptote, respectively, for the *i*-th item in a given *n*-item test. The probability that an examinee with scale score $\hat{\theta}$ will answer item *i* correctly is

$$\hat{P}_i(\hat{\theta}) = \hat{c}_i + \frac{1-\hat{c}_i}{1+\exp\left[-1.7\hat{a}_i\left(\hat{\theta}-\hat{b}_i\right)\right]} .$$

Estimated Difficulty

Let $f(\hat{\theta})$ be the relative frequency of $\hat{\theta}$ in the normative distribution of interest for a test and level; it should be noted that $\hat{\theta}$ is a maximum likelihood estimate based on item-pattern scoring. The LOSS and the HOSS are defined to be the lowest and highest obtainable scale scores, respectively, and they cover the range of scale scores obtainable for any given level of a test. The estimated proportion-correct score (p-value) for the *i*-th selected-response item in a test is

$$\hat{P}_i = \sum_{\hat{\theta}=LOSS}^{HOSS} \hat{P}_i\ (\hat{\theta})\, f\,(\hat{\theta}) .$$

The average p-value for an *n*-item is

$$\overline{P} = \frac{1}{n}\sum_{i=1}^{n} \hat{P}_i ,$$

where *n* is the number of items.

Estimated Reliability

Reliability values were produced using the following procedures:

The expected raw score for an examinee with scale score $\hat{\theta}$ is

$$X(\hat{\theta}) = \sum_{i=1}^{n} P_i(\hat{\theta}) .$$

The raw score mean is obtained from

$$\hat{\mu}_x = \sum_{\hat{\theta}=LOSS}^{HOSS} X(\hat{\theta})\, f\,(\hat{\theta}) .$$

An estimate of the variance of the true scores over examinees can be obtained from the following:

$$\sigma_T^2 = \sum_{\hat{\theta}=LOSS}^{HOSS} X^2(\hat{\theta})\, f\,(\hat{\theta}) - \hat{\mu}_x^2 \,.$$

The conditional item score variance is

$$\sigma^2\left(X_i \middle| \hat{\theta}\right) = P_i(\hat{\theta})\, Q_i(\hat{\theta}) \,.$$

Note that the variance of the observed scores conditioned on $\hat{\theta}$ is the error variance. Given the assumption of local item independence, the raw score error variance for an examinee with scale score $\hat{\theta}$ is

$$\sigma_E^2(\hat{\theta}) = \sum_{i=1}^{n} \sigma^2\left(X_i \middle| \hat{\theta}\right).$$

The raw score error variance across all examinees can be expressed as

$$\sigma_E^2 = \sum_{\hat{\theta}=LOSS}^{HOSS} \sigma_E^2\,(\hat{\theta})\, f\,(\hat{\theta}) \,.$$

The item score variance for item (not conditioned) can be obtained from

$$\sigma_i^2 = P_i Q_i \,.$$

The coefficient alpha value is obtained by the standard formula,

$$C_\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} \sigma_i^2}{\sigma_x^2}\right],$$

where

$$\sigma_x^2 = \sigma_T^2 + \sigma_E^2 \,.$$

For the Total score, an approximation method is used to obtain reliability estimates based on the assumptions of classical true-score theory and the numbers of items in the tests comprising the composite (e.g., Sequences, Analogies, Quantitative Reasoning). Using the Spearman-Brown formula, an average item reliability is estimated for each test in the composite. Then, these average item reliabilities are used to produce three composite reliability estimates. In producing each composite reliability estimate, true scores from different tests were assumed to have a correlation of 0.8, rather than 1.0, which is the traditional assumption within tests. The 0.8 value was determined by reviewing a large number of achievement tests. Use of the 0.8 value rather than a uniform 1.0 value lowers the reliability estimates because the tests measure different but correlated skills. Lastly, the two composite reliability estimates are averaged to yield the final composite reliability estimate for the Total score.

The specific procedure is as follows. For the *k*-th test entering composite *T,* with $k = 1 \ldots K$, let

$$n_T = \sum_{k=1}^{K} n_k \, .$$

Define

$$n_c = \sum_{k=1}^{K} \sum_{k'=1}^{K} C_{kk'} \, n_k \, n_{k'} \ ,$$

where $C_{kk'} = 1.0$ for $k = k'$; $C_{kk'} = 0.8$ for $k \neq k'$. The estimate of the composite score reliability based on the reliability of the *k*-th test is

$$\hat{\alpha}_{T_k} = \frac{n_c}{n_c + \ n_T n_k \left(1 - \hat{\alpha}_k\right) / \hat{\alpha}_k} \ .$$

Fisher's r-to-z transformation is used to average these estimates and convert them back to the correlation metric:

$$Z_T = \sum_{k=1}^{K} \left(n_k \ / \ n_T\right) \ ln \left[\left(1 + \hat{\alpha}_{Tk}\right) / \left(1 - \hat{\alpha}_{Tk}\right)\right] ,$$

and

$$\hat{\alpha}_T = \left[\exp\left(Z_T\right) - 1\right] / \left[\exp\left(Z_T\right) + 1\right] .$$

Second-Order Factor Model

The equations for second-order factor models are

$$\eta = \Gamma\xi + \zeta \qquad (1)$$

$$y = \Lambda_y\eta + \varepsilon \qquad (2)$$

where,

y is a vector of input variables,

η is a vector of first-order factors (latent variables),

Λ is a matrix of coefficients of the regression of *y* on η,

ε is a vector of measurement errors in *y*,

Γ is a matrix of second-order factor loadings,

ζ is a vector of unique variables for η,

ξ is a vector of second-order factors (latent variables).

The relations between the first- and second-order factors are given by the first equation. The second equation represents the relations between the indicators (observed variables) and first-order factors.

Formulas for Fit Indices Used in the EQS Analyses

Chi-square (χ^2): The chi-square index is a measure of difference between the observed and estimated matrices. The estimation method used is maximum-likelihood (ML) procedure. A non-significant χ^2 value indicates that the two matrices are not statistically different.

Target coefficient (T): The index T (Marsh & Hocevar, 1985) is the ratio of the chi-square of the first-order model to the chi-square of the more restrictive model.

Goodness of fit index (GFI): The GFI is a measure of the amount of variance and covariance in the observed matrix that is predicted by the reproduced matrix.

$$GFI = 1 - \frac{\left(s - \hat{\partial}\right)' W^{-1}\left(s - \hat{\partial}\right)}{s' W^{-1} s},$$

where s is the original correlation matrix, $\hat{\partial}$ is the estimated correlation matrix, and W is the weight matrix consisting of the asymptotic covariance in s. The GFI here is based on weighted least squares approach.

Adjusted goodness of fit (AGFI): The AGFI is the GFI adjusted for the degrees of freedom of a model relative to the number of variables.

$$AGFI = 1 - \frac{(p+q)(p+q+1)}{2d}(1 - GFI),$$

where $p + q$ = number of observed variables analyzed, and d = degrees of freedom of the model.

Comparative fit index (CFI): The CFI is a measure of improvement in noncentrality when going from one model to another.

$$CFI = \frac{\hat{\tau}_i - \hat{\tau}_k}{\hat{\tau}_i},$$

where $\hat{\tau}_k = \max\left[(n\hat{Q}_k - d_k), 0\right]$, $\hat{Q}_k$ is the value of the fitting functions obtained at $\hat{\theta}_k$, and
$\hat{\tau}_k = \max\left[(n\hat{Q}_i - d_i), 0\right]$, $\hat{Q}_i$ is the value of the fitting functions obtained at $\hat{\theta}_i$.

Root-mean-square error of approximation (RMSEA): The RMSEA is an index of fit of two different models with the same data.

$$RMSEA = \sqrt{\frac{1}{k}\sum_{ij}\left(s_{ij} - \partial_{ij}\right)^2},$$

where s is the original correlation matrix, $\hat{\partial}$ is the estimated correlation matrix, and $k = (p + q)(p + q + 1)/2$.