CONTENTS

CHAPTER ONE:
GENERAL OVERVIEW

The MEZURE is an interactive computer-administered comprehensive measure of general intelligence which has been standardized on subjects between the ages of 6 through Adult. It is composed of 14 subtests that assess associative reasoning, word knowledge, visual-spatial competencies, auditory and visual memory, hypothetical-deductive reasoning, general knowledge, processing speed, and social awareness. There are two additional subtests for ages 18.0 and over which measure Stress Tolerance. It is suitable for use by professionals in schools, clinics, residential treatment centers, hospitals, and private practices. The MEZURE is also available in Spanish and Russian, in addition to standard American English, allowing examinees to be accurately assessed in their native language.

The MEZURE offers many advantages over the traditional paper-pencil model. First, by using the MEZURE as a diagnostic tool, the examiner is free to concentrate on clinical observations of the examinee's behavior and responses. Second, because administration is automated, the occasional sources of examiner error (i.e., scoring miscalculations, modifications in directions, unnecessary prompting, lack of close adherence to time limits) and examiner bias are eliminated (Anastasi, 1988). Finally, the multimedia administration of the MEZURE is interesting and enjoyable, promoting optimum attention, concentration, and effort on the part of the examinee.

The subtests are organized into a **Screening Battery** that consists of 4 subtests and requires approximately 15 minutes administration time, or a 7-subtest **Standard Battery**, which takes 30 minutes to complete. In addition, there are 5 supplemental subtests that facilitate enhanced clinical assessment of memory, processing speed, and social apperception.

The 7 core subtests are divided into **Fluid** and **Crystallized Scales**. The **Fluid Scale** measures a person's ability to evaluate new and unusual problems. The tasks require inductive and deductive reasoning and emphasize hypothesis testing and problem solving. None of the tasks reflect competencies gained from prior learning or schooling experiences. By comparison, the **Crystallized Scale** measures acquired knowledge and is heavily influenced by formal schooling, cultural experiences, and verbal conceptual development. The **Screening Battery** has 2 Crystallized subtests and 2 Fluid subtests, while the **Standard Battery** has 3 Crystallized and 4 Fluid subtests.

MEZURE BATTERY	FLUID	CRYSTALLIZED	ADMINISTRATION TIME
Screening	• Vis. Closure • Vis. Analogies	• Categorization • Information	15 – 20 Minutes.
Standard	• Vis. Closure • Vis. Analogies • Vis. Memory • Aud. Memory	• Categorization • Information • Vocabulary	25 - 30 Minutes.

The full MEZURE Battery also includes the following Supplemental Subtests:

SUBTEST	ADMINISTRATION TIME
• Processing Speed	3-4 minutes
• Social Apperception	5 minutes
• Auditory Memory with Visual Distractions	3 minutes
• Auditory Memory with Auditory Distractions	2-3 minutes
• Visual Memory with Auditory Distractions	3-4 minutes
	15-20 minutes total administration time

Theoretical Orientation

The MEZURE is based on the contemporary and empirically supported models of the extended Gf-Gc theory of Horn and Cattell (Cattell, 1941; Cattell & Horn, 1978; Horn, 1968; Horn & Cattell, 1966) and the Cattell-Horn-Carroll theory of cognitive abilities (Carroll 1989, 1993). It is designed to measure a broad range of cognitive abilities as represented in current theories of human intelligence. Such an approach allows for empirically driven interpretation of the results. (For a full discussion of the theoretical model of the MEZURE, please refer to Chapter 2).

IQ Scores, Subtest Scores, and Percentiles

The MEZURE yields standard scores (Mean = 100. Standard Deviation = 15) for 3 scales: Crystallized IQ, Fluid IQ, and Composite IQ. The 12 subtests are reported as scaled scores (Mean = 10, Standard Deviation = 3). Percentile ranks are provided for both the standard IQ scores and the subtest scaled scores as an additional aid to interpretation.

General Administration Guidelines

The MEZURE should be administered in an environment that is quiet, comfortable, and free of distractions. Adequate lightening and ventilation, as well as lack of glare on the computer screen, are essential. Establish rapport before testing begins by explaining the purpose and interactive nature of the test. Specifically, examinees should be aware that all visual stimuli will be presented to them on the computer screen, that all instructions will be provided by the computer via headphones or speakers, and that they will be responding to all questions by using their mouse to click on items shown on the computer screen. *Please note:* The mouse was used as the standard input device during standardization, however, any peripheral input device may be used for all subtests - ***other than the Supplemental Processing Speed subtest*** **-** without affecting test results. Since the Processing Speed subtest measures the speed as well as the accuracy of responses, a standard mouse must be used to interpret test results based on normative data. It should be noted that although Visual Closure seemingly incorporates speed in addition to accuracy in the interpretive formula, use of a touch screen instead of a mouse should not affect the resulting score. This is since the scoring methodology of Visual Closure was designed to preclude variations in different computer speeds; a sophisticated interpretive formula tracks the portion of the total stimulus revealed rather than utilizing the simple response time measurement. In addition, since the subject is only required to "click" ***anywhere on the screen*** in order to register his closure of the stimulus, the variations in "pointing time" inherent in different input devices are rendered irrelevant.

The MEZURE is entirely computer administered with subtest entry points, basal points (i.e., the predetermined number of correct answers required for testing to continue), and ceiling points (i.e., the predetermined number of incorrect answers required for testing to be discontinued) calculated automatically. Thus, examinees do not become frustrated by receiving too many overly easy or overly difficult items.

Local and Remote Administration

The examiner should always be able to follow the testing and observe examinee behaviors while remaining unobtrusive. A suggested arrangement for local administration would be to sit alongside but slightly behind the examinee (slightly outside the examinee's field of vision unless he turns specifically to the examiner). For remote administration, position the subject's webcam view such that the examiner can always retain a full view of the subject. However, the examiner should turn off his/her video and mute his/her audio once testing begins, and only turn his/her video & audio back on when needed. Any prompts should preferably only be provided in between sub-tests so as not to distract the examinee during the assessment. The clinician's supervision is meant to ensure that the examinee is following the directions, is comfortable with the computer and the use of a mouse, that the subject's environment is appropriate, and that he is actively attending to test stimuli throughout test administration. In addition, the MEZURE screens the examinee prior to the commencement of each subtest to determine if he / she demonstrates the minimum functionality required to take that subtest. If the pre-subtest

screening indicates that the examinee is not able to attempt a specific subtest, the MEZURE will ask the examiner to decide whether to skip that subtest or to proceed with the administration. The examiner must be available, whether locally or remotely, to make that determination.

Since response times are recorded, allowing the examinee to take a break during an item or within a subtest is not recommended. If a break is required, do so between subtests only. ***However, once you have exited a subtest, you cannot reenter it during the same testing session.*** By refreshing your browser during a subtest, the MEZURE will restart the subtest automatically. This is not recommended for ideal testing situations, only in exceptional situations.

Because the clinician's presence may be distracting to the examinee or provoke anxiety during test administration, it is recommended that examiners, for both local and remote administration, direct the examinee's attention to the computer while explaining that they will be available - though not always easily visible - throughout the testing session; for example: "The computer is going to tell you everything you have to do. I will be available to answer any questions that you might have or to help you if there is any problem."

Please note that the MEZURE was *NOT* designed for use with visually impaired, color blind, or hearing-impaired individuals. *Assessment Technologies, Inc. assumes NO responsibility for any results obtained by administration of the MEZURE to these subject populations.*

The MEZURE provides two simple training items before actual testing begins, which allow the examinee to practice mouse skills and become familiar with the question-answer format. In addition, these initial items might provide a very basic screening to determine if the subject is able to understand basic instructions and take the test properly. All necessary instructions for the examinee are already incorporated within the MEZURE. It is recommended that examiners do not provide any extra prompts during the administration of the test unless absolutely necessary.

Subtest Entry Point Routing Procedure

The MEZURE uses an innovative procedure which incorporates *dynamic adaptive routing technology (DART ™)* to determine the appropriate entry level of each subtest. This allows for an in-depth assessment of an individual in a short amount of time. Such a routing procedure eliminates the less precise method of determining the entry point by chronological age.

After the two introductory training items, all examinees are presented with the first item of the first subtest, **Visual Closure**. The scoring of this subtest is based upon how many of the 600 total "units" (squares of the picture which randomly appear) need to be revealed before the examinee can successfully identify the item. The raw score is later transformed utilizing a scale of 8 thresholds which were empirically derived from the normative sample. Every examinee can successfully identify every item if he waits long enough, and he/she has no way of determining the transparent scale utilized by the MEZURE to score each item response. This subtest is administered first, since it can be administered with no starting points or ceilings without resulting in examinee frustration. The **Visual Closure** Subtest allows examinees to become further accustomed to the computer, the use of a mouse, and the question-answer format while allowing the MEZURE to make a preliminary determination of the examinee's functional level. The second subtest, **Visual Analogies**, also begins at the first item for all ages since empirical studies of the normative sample dictated that no starting points were indicated. The examinee exits **Visual Analogies** when he reaches a ceiling of 3 consecutive incorrect responses. The combined results of **Visual Closure** and **Visual Analogies** allow the MEZURE to determine the proper starting point for the third subtest, **Information**. The MEZURE then proceeds to select all subsequent entry levels based on the ***cumulative performance of all previously administered subtests***. For example, the entry point for the third subtest, **Information**, is determined by the scores obtained on the first two subtests (**Visual Closure** and **Visual Analogies**), while the entry point for **Categorization** is determined by the scores obtained on **Visual Closure**, **Visual Analogies**, and **Information**. Thus, with each subsequent subtest, the entry point is more refined. ***This advanced methodology significantly reduces administration time, as well as the frustration which might occur when an examinee is presented with extraneous items.*** Some individuals, however, exhibit such significant subtest scatter, that even the fine-tuned entry points determined by the DART™ methodology might be inappropriate for a specific subtest. For example, an examinee might be weak in vocabulary although his performance on all prior subtests could be quite high. The MEZURE will automatically accommodate for this possibility as well. If the examinee demonstrates that the automatic entry point is at a level that is too difficult for him / her, the MEZURE will then begin item presentation for that subtest again with the easiest possible item and continue upward until a ceiling is attained. Each subtest has a ceiling point wherein testing is discontinued. The entry and exit points for each subtest were determined by empirical studies of the standardization sample which are automatically incorporated into the test administration.

Screening Battery

The Screening Battery consists of 2 fluid subtests (**Visual Closure** and **Analogies**) and 2 crystallized subtests (**Information** and **Categorization**). The screening is intended to provide a brief measure of cognitive functioning. Utilization of the Screening Battery is useful in clinical settings as an expeditious measure of general functioning or in school settings where a full-length triennial evaluation is not required.

Standard Battery

The Standard Battery is intended to provide a comprehensive measure of an individual's current intellectual functioning in both fluid and crystallized domains. The Fluid and Crystallized scales as well as the individual subtests yield rich information regarding cognitive strengths and weakness. The Fluid Scale has 4 subtests: **Visual Closure, Analogies, Visual Memory,** and **Auditory Memory**. The Crystallized Scale consists of 3 subtests: **Information, Categorization,** and **Vocabulary**. The **Composite IQ** is a total of the 7 subtests and is viewed as a summative index of general intellectual functioning. The Standard Battery would be appropriate for clinical and psychoeducational purposes such as identification of learning disability, verification of learning styles, or diagnosis of Mental Retardation or Attention Deficit Hyperactivity Disorder (ADHD), and any environment which requires the measures of these abilities.

Supplemental Subtests

To further enhance the clinical evaluation of memory, learning, and social functioning, five additional subtests are available. These subtests are not part of the Screening or Standard Battery. Rather, each subtest reflects distinct processing modalities that may prove helpful in in-depth psychoeducational, neuropsychological, or clinical assessments. For example, 3 subtests measure visual and auditory short-term memory acquisition under various distracting stimuli (**Visual Memory with Auditory Distractions, Auditory Memory with Auditory Distractions, Auditory Memory with Visual Distractions**). Another subtest, **Processing Speed**, evaluates an individual's ability to quickly scan and classify details of visual stimuli. It is influenced by attention to detail, task persistence, distractibility, and impulsivity. The final supplemental subtest, **Social Apperception**, taps social awareness and attention to facial nuances and to verbal expressions. Total administration time for all 5 supplemental subtests is approximately 15 minutes.

CHAPTER TWO:
THEORETICAL FOUNDATION

Contemporary Theoretical Perspective

Cattell first proposed the Gf-Gc model of human intelligence in 1941. Moving beyond Spearman's (1927) concept of one general functional unit (*g)*, he postulated that intelligence was not a single process but consisted of two separate and distinct abilities: **Fluid** (Gf) and **Crystallized** (Gc; Cattell, 1941, 1971). With the theoretical evidence accumulating over the next 50 years, the Cattell theory evolved beyond a two-factor approach into the extended Gf-Gc model of Horn and Cattell (Cattell, 1987; Cattell & Horn, 1978; Horn, 1965, 1968, 1972; 1976, 1985, 1988, 1989; Horn & Cattell, 1966, 1967; Horn & Stankov, 1982). In addition, Carroll (1972, 1989,1993) advocated for a three-stratum theory that specified more than 60 primary mental abilities at the third level, eight broad abilities at the second order, and one very broad ability (*g)* at the top strata. Incorporating the earlier work of Cattell and Horn, this approach is referred to as the Cattell-Horn-Carroll theory of cognitive abilities.

Additional structural equation modeling further refined the Cattell-Horn-Carroll theory to include nine primary dimensions: Fluid Reasoning (Gf), Crystallized Reasoning (Gc), Short-term Memory (Gsm), Long-term Retrieval (Glr), Processing Speed (Gs), Auditory Processing (Ga), Visual Processing (Gv), Quantitative Ability (Gq), and Decision/Reaction Time or Speed (Gt). The MEZURE subtests are intended to assess five of these broad abilities (Gf, Gc, Gsm, Gs, and Gv) with core subtests divided into two primary divisions: The **Fluid IQ Scale** and the **Crystallized IQ Scale**. It should be noted that many subtests tap into two dimensions (e.g., Gf-Gv, Gf-Gsm).

Fluid Reasoning (Gf)

This dimension of intelligence is measured by tasks that require age-appropriate inductive and deductive reasoning, concept formation, analysis-synthesis, combinatorial analysis, and symbolic classifications under novel conditions. To make such inferences, a person must concentrate and attend to details. Cognitive flexibility, motivation, perseverance, and carefulness are hypothesized to affect Gf. Fluid intelligence is not heavily influenced by formal schooling experiences or by one's cultural setting. MEZURE subtests that measure this domain are Analogies, Visual Closure, Visual Memory, Auditory Memory,

Auditory Memory with Visual Distractions, Auditory Memory with Auditory Distraction, and Visual Memory with Auditory Distractions.

Crystallized Reasoning (Gc)

By comparison, crystallized abilities reflect knowledge acquired from formalized learning experiences. This dimension taps word knowledge, verbal categorizations, fund of general information, behavioral functioning such as estimations of others' feelings, and mechanical and numerical facilities. Thus, crystallized intelligence reflects quality and quantity of formal education, educational opportunities such as travel and access to libraries, as well as acculturation. Crystallized competencies are reflected in the MEZURE subtests Categorization, Information, Vocabulary, and Social Apperception

Short-term Memory (Gsm)

This domain reflects the ability to immediately recall (within one minute or so) the order of a series of randomly related elements (e.g., letters, numbers, designs, grid locations). The modality of presentation (visual, auditory, tactile) is not relevant; instead, it is the ability to maintain awareness of, and then recall for, the correct sequence of the components. Freedom from distractibility and attention span may affect performance on Gsm tasks. MEZURE subtests that tap this area are Auditory Memory, Visual Memory, Auditory Memory with Visual Distractions, Auditory Memory with Auditory Distractions, and Visual Memory with Auditory Distractions,

Processing Speed (Gs)

Processing speed is defined as the ability to quickly perform simple scanning or matching tasks. The requirements are such that almost all people would get the correct answer if speed were not an issue. Concentration, effort, and attention to detail are important factors. The MEZURE subtest Processing Speed assesses this domain.

Visual Processing (Gv)

This area taps the ability to analyze and synthesize visual information. It is measured by such tasks as mental rotation of visual shapes, identification of shapes when parts of the whole are missing (visual closure), and completion of matrix or object analogies. This area is reflected in the MEZURE subtests Visual Closure and Visual Analogies.

Theoretical Validation of the MEZURE Model

The 7 core subtests of the MEZURE Standard Battery are divided into **Fluid** and **Crystallized Scales.** The assignment of each subtest to either the Fluid or Crystallized Scales is based on empirical validation via subtest intercorrelations and subsequent factor analysis. In factor analysis, subtests that cluster together represent a common abstract

and underlying dimension. The dimension is referred to as a *factor.* The subtests cluster together because they are highly correlated and are measuring a similar construct (i.e., Crystallized or Fluid cognitive functioning). The factors are further refined by *rotation*, which forces the factors to be relatively independent of one another. The factor loadings, which vary in value from 0.00 to +1.00, represent the degree to which each of the subtests correlates with the factor. Support for the Gf-Gc model of the MEZURE is validated by the subtest intercorrelations and robust factor loadings as illustrated in Tables 2.1 and 2.2.

Table 2.1 Intercorrelations Of MEZURE Subtests for The Entire Age Group

Subtest		VC	A	I	C	VM	V	AM
Visual Closure (VC)		--						
Analogies (A)		.35	--					
Information (I)	.08	.15	--					
Categorization (C)		.11	.18	.12	--			
Visual Memory (VM)		.30	.49	.12	.15	--		
Vocabulary (V)	.08	.17	.23	.12	.12	--		
Auditory Memory (AM)		.35	.46	.10	.15	.51	.12	--

Table 2.2 Exploratory Factor Analysis with Oblique Rotation for The Entire Age Group

Subtest	Fluid Factor	Crystallized Factor
Visual Closure	.66	
Analogies	.74	
Visual Memory	.78	
Auditory Memory	.80	
Information		.76
Categorization		.43
Vocabulary		.75

Multilingual Test Administration

The MEZURE includes versions in Spanish, Russian, and standard American English. Simply select the desired language from the language pull-down list found at the lower right corner on the main demographic screen. If no language is specified when completing the demographic information screen, the program will automatically administer the test in English. The MEZURE is administered in the identical manner in all languages.

The MEZURE was not merely ***translated*** into various languages but rather specifically ***adapted*** for those languages; each item in each language maintained a difficulty level that was comparable to that of its English counterpart. In fact, the final item pool of the MEZURE was only determined after the test was adapted into various languages to ensure that only those items which adapted well were included.

The Composite IQ will be calculated only if the entire Brief or Standard Battery has been completed. The Fluid-Crystallized cluster scores will only be displayed if the Standard Battery was completed. This ensures that Fluid-Crystallized Cluster Scores are only reported if they are based on enough subtests to be a reliable overall indicator of examinee performance. Hard copy of test results may be obtained by simply clicking the "PRINT" button at the center bottom of the scores screen.

Subtest Scatter Screen:

If the entire Standard Battery has been completed, the "**Subtest Scatter Screen"** will appear next. This screen shows the difference between the Fluid and Crystallized Domains, whether that difference is clinically significant or not, and at what level of significance. If the discrepancy is significant, a separate mean scaled score will be calculated for the Fluid and Crystallized Domains and used to determine subtest scatter within each domain; if the discrepancy is NOT significant, the overall mean will be used to determine subtest scatter. (For a detailed clinical explanation of Fluid-Crystallized Domain Discrepancy, please see **Analysis of Subtest Profile** in Chapter Four above)

Graphing subtest scores:

At the bottom of the "**MEZURE Scores Screen"** which appears at the end of a test administration (or when viewing scores from a previous test administration) a button marked "NEXT" appears on the lower right. Click on this button to view a graphic representation of the MEZURE scores. (If the entire Standard Battery has been administered, the **"Subtest Scatter**" screen will come first).

Graphing cluster scores:

At the bottom of the **Subtest Graph Screen** (see B above) a "NEXT" button appears on the lower right allowing the examiner the option to proceed to the **Cluster Score Graph Screen**. Click on this button to view a graphic representation of the Cluster Scores.

Summary of General Administration Guidelines

To insure the reliability of the test, please adhere to the following general guidelines:

- Review the contents of this test manual.
- Establish rapport with the examinee. Explain the purpose of the test and maintain a positive attitude throughout the testing session.
- The MEZURE includes all necessary instructions for the examinee. Do not provide any extra prompts during administration of the test.
- Be sure that the examinee knows how to use a mouse.
- Note: The MEZURE includes two simple training items before actual testing begins that allow the examinee to practice mouse skills and become familiar with the MEZURE's question-answer format.
- Administer the test in an environment that is quiet, well lit, well ventilated, comfortable, and free from distractions.
- To restart a subtest in the middle of the subtest, refresh your browser if needed.
- Warning: Once you exit a subtest that has already begun, you cannot re-administer that subtest again unless you begin a new test administration. (See above for more detailed information.)
- Because all response times are recorded, try not to allow the examinee to take a break during or between items. If a break is necessary, it is best to provide one between subtests only.

CHAPTER THREE:
Administration Directions / Item Descriptions

To facilitate examiner familiarity with the MEZURE, we have included the full script of the test here. Please note that this script is for reference purposes only, since all the instructions and item prompts listed are automatically read during MEZURE administration by professional, dialect-neutral voice-actors.

Introduction

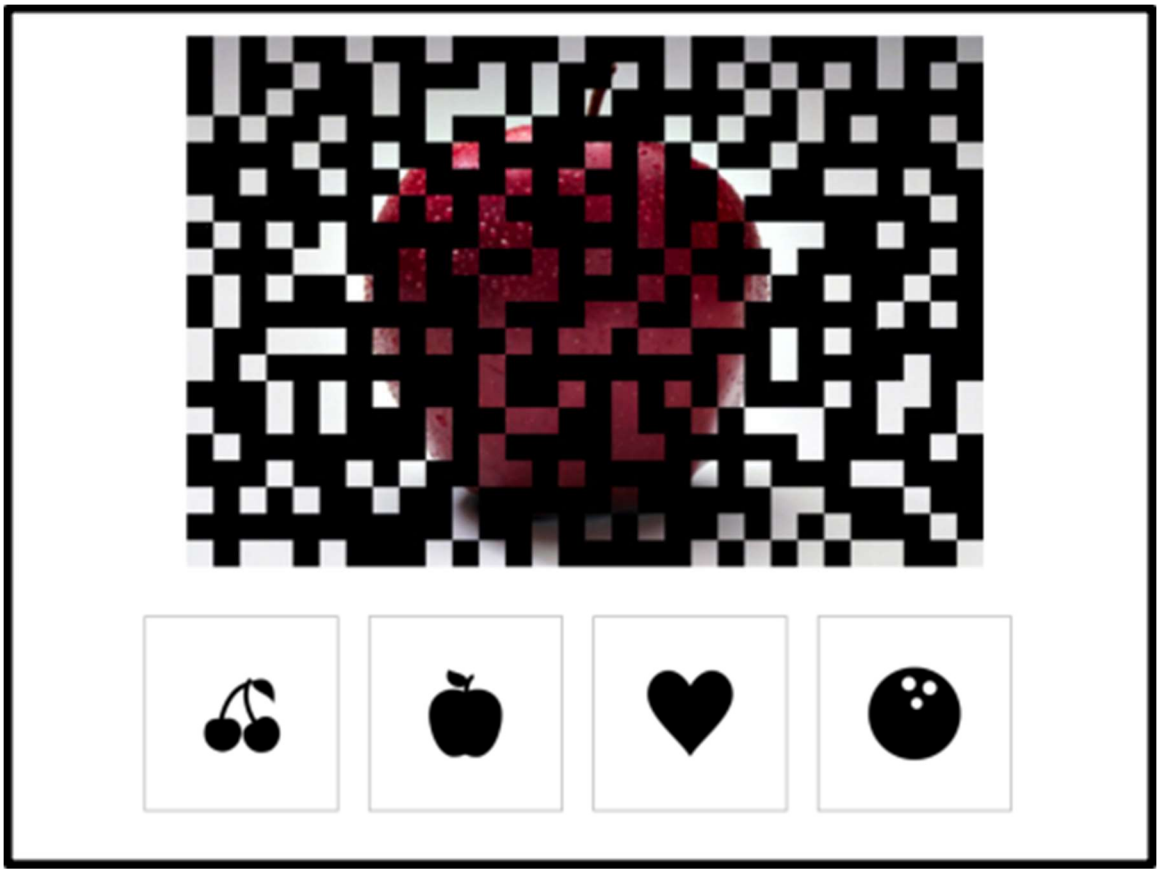
Directions:

Today we will be working on this computer. You will be asked questions from many different areas. Some things will be easy for you and some will be hard. Listen carefully and answer the questions as best as you can.

I will be asking you all the questions. After I finish each one, you will have a chance to choose your answer. If you think that you chose the wrong answer, you will have a few seconds to choose a different one. If you did not hear the question, you will have one chance to hear it again by clicking on the REPEAT button.

Let's try one before we begin.

Visual Closure



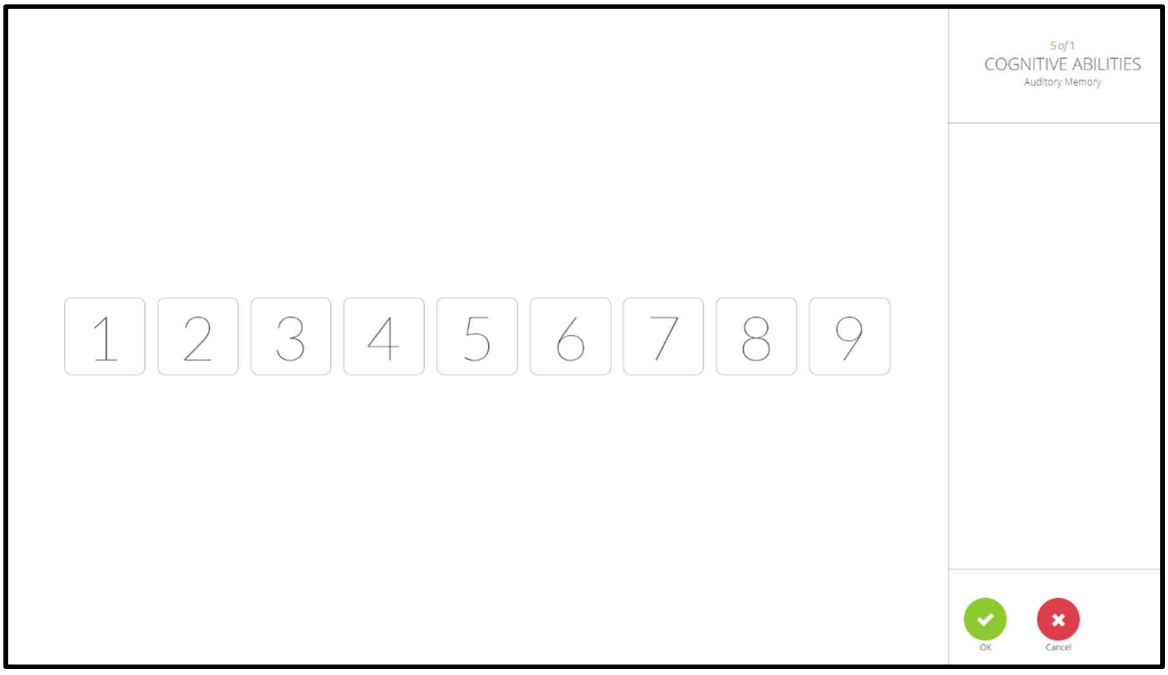
Directions:

Look at the screen. A picture is coming. When you know what the picture is, click on the screen. Here it comes.

- *Please Note: The picture is revealed progressively and when the examinee guesses he clicks on the screen, the picture disappears, and he gets the four symbols to choose from.*

Trial item 1 (apple)
(Remember – click on the screen when you know what the picture is)
Now choose from these buttons to show what the picture is.
Good! Now watch the screen to see the whole picture.
You were right! It is a picture of an apple. Now try some more.
(That's not quite right. Watch the screen to see the whole picture.
You see, it is a picture of an apple. This is the correct answer. Now try more)

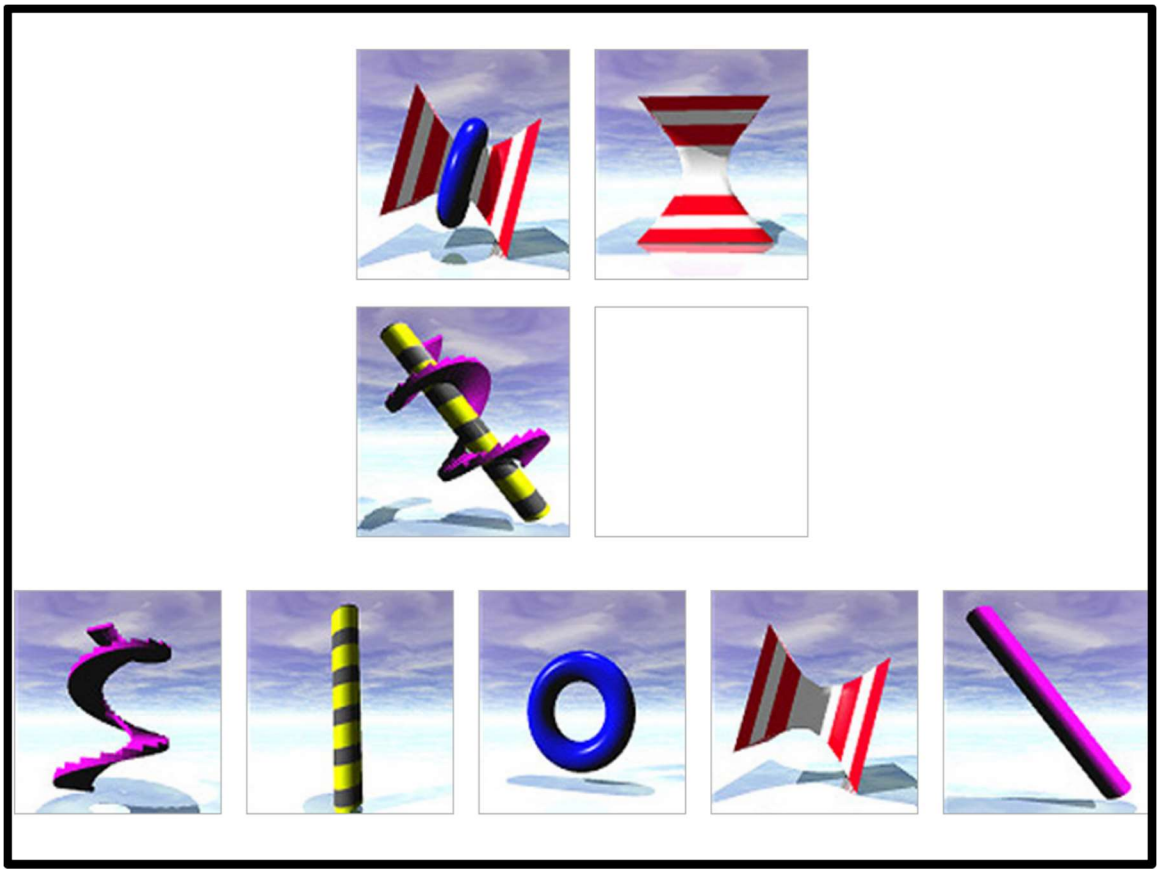
Auditory Memory



Directions:
You are going to hear numbers in a certain order. Pay close attention. When it is your turn, click on the numbers in the same order. Click on the OK button when you are done. If you want to change your answer, you will have one chance to click on the CANCEL button and begin your answer again. Let's try one together.

Trial item 1: 6-2
Good! Now try some more.
(That's not quite right. Watch how I do it. Now try some by yourself.)
(Remember – click on the OK button when you are done.)

Visual Analogies



Directions:

You are going to see two pictures that are related in some way. Choose a picture from the bottom to make another set that is related in the same way. Let's try one together.
This picture… is to this picture… as this picture is to… which one of these?
Click on your answer.

Trial item 1:
Good! Now try some more.
(That's not the best answer. You see…this is related to… this…because they are both the same color. This one is the same color as…this, so…this picture is the correct answer. Now try some more.).

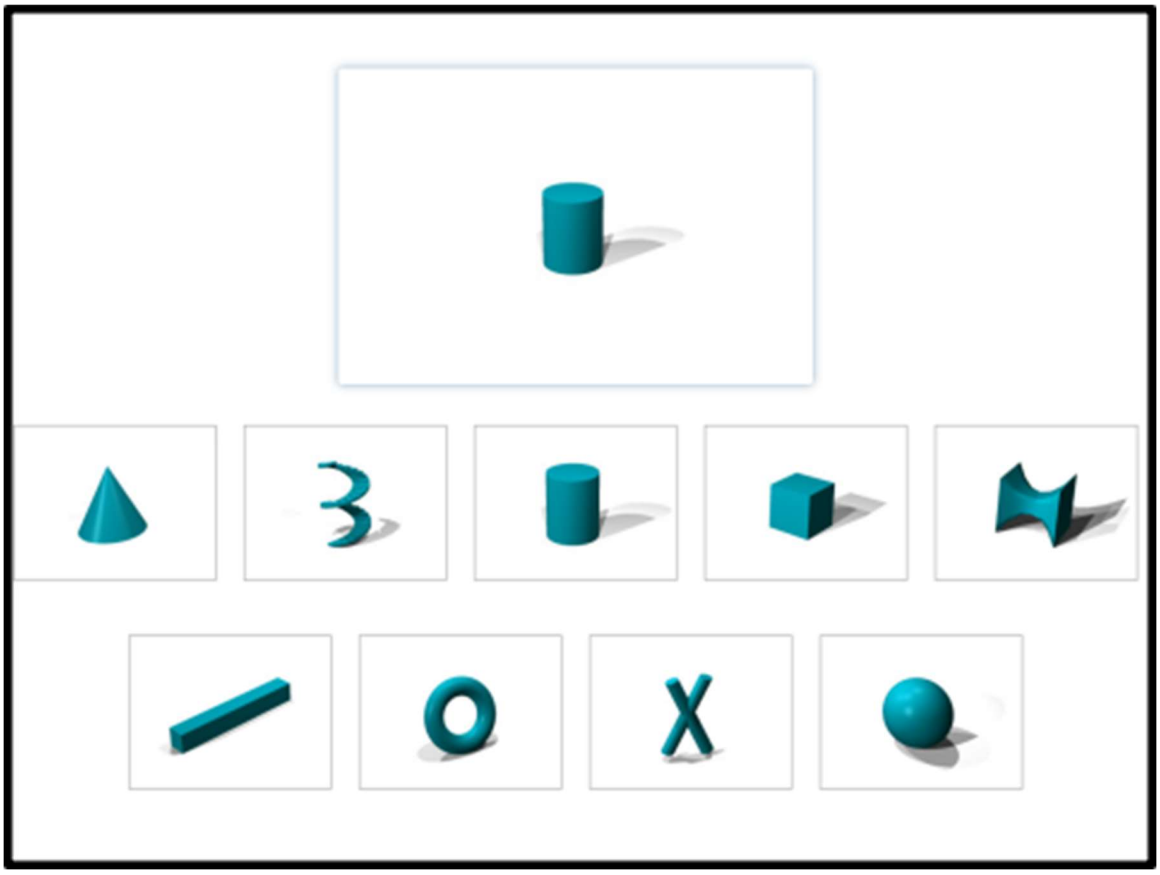
Vocabulary

Directions:

You are going to see some pictures. Then, you will hear a word. Click on the picture that goes with the word. You may use the same picture as your answer more than once. Let's try one together.

Trial item 1: Chair.
Good! Now, try some more.
(That's not quite right. This is the picture that goes with chair. Now, try more)

Visual Memory



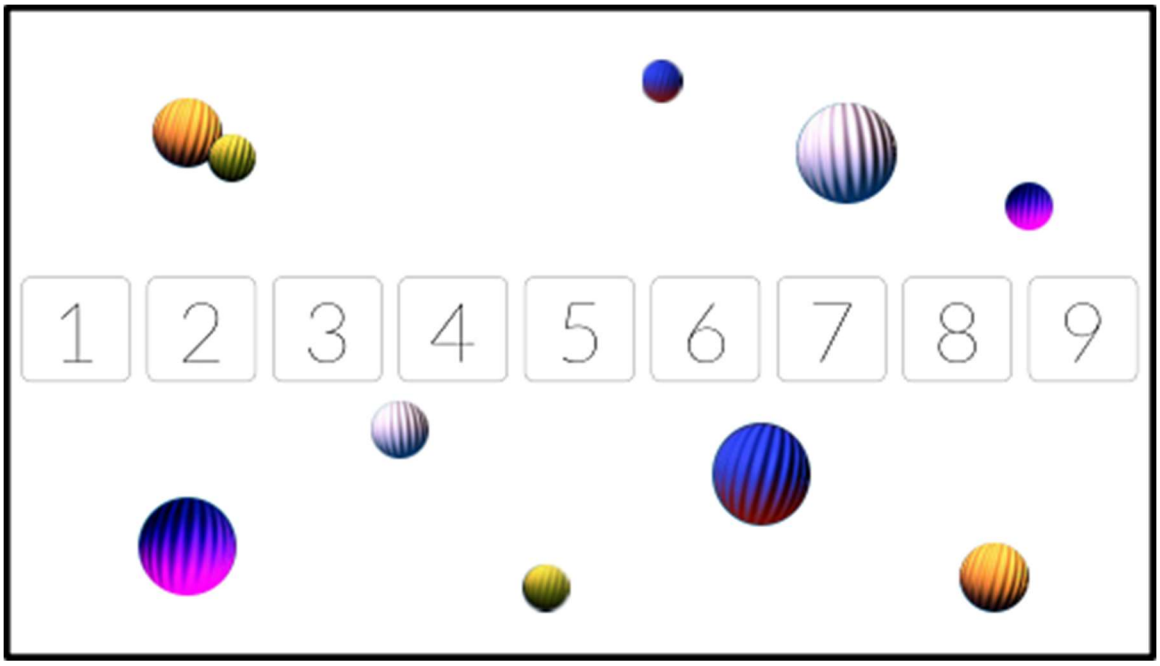
Directions:

Before we begin, let's get familiar with some new shapes. Each time you see a shape here… find it on the bottom.
(Shapes are presented individually.)
(That's not quite right. This is the right shape. Let's try it again)
Good! Now you are ready to begin.
(Now, ask the person helping you to choose one of these options.)

Watch the screen carefully. Shapes will be appearing in a certain order. Pay close attention. When it is your turn, click on these shapes in the same order. Click on the OK button when you are done. If you want to change your answer, you will have one chance to click on the CANCEL button… and begin your answer again. Let's try one together.

Trial item 1:
Good! Now try some more.
(That's not quite right. Watch how I do it. Now try some by yourself.)

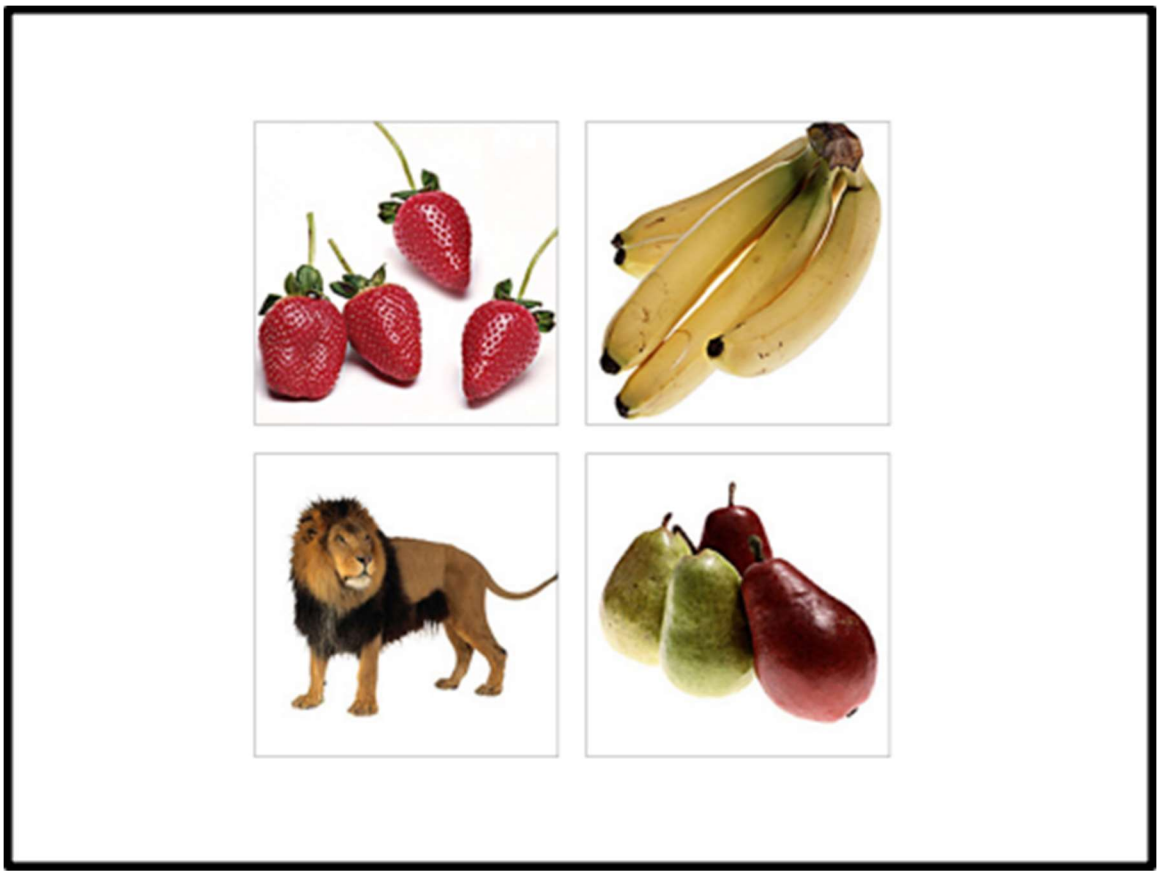
Auditory Memory with Visual Distractions



Directions:

Listen carefully. You are going to hear numbers in a certain order. Pay close attention. This time, you will see things on the screen as you hear the numbers. Click on the numbers in the same order they were said. Let's begin.

Categorization



Directions:

You are going to see four pictures. Three of the pictures are alike in some way and one picture is not like the others. Choose the picture that does not belong with the others. Let's try one together.

Trial item 1:
Good. Now try some more.
(That's not quite right. These three are alike because they are all fruits. This one is not a fruit, so it does not belong with the others. Now try more.)

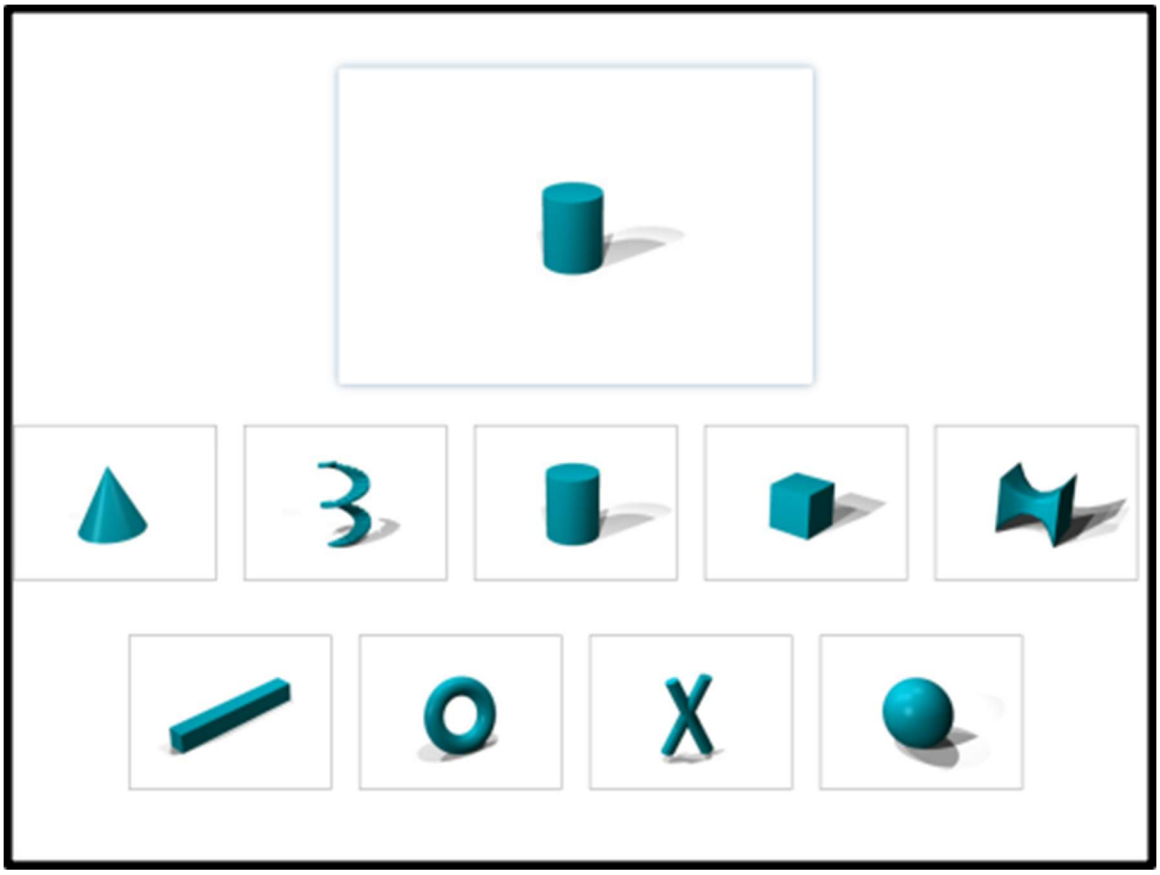
Trial item 2:
Good. Now try some more.
(That's not quite right. This is the correct answer. Now try more.)

Information

Directions:

You are going to hear some questions. Click on the picture that best answers each question. You may use the same picture as your answer more than once. Let's try one together.

Visual Memory with Auditory Distractions



Please note: This time there are sound distractions.

Directions:

Watch the screen carefully. This time, you will hear sounds as the shapes appear. Click on the shapes in the same order that they were shown. Let's begin.

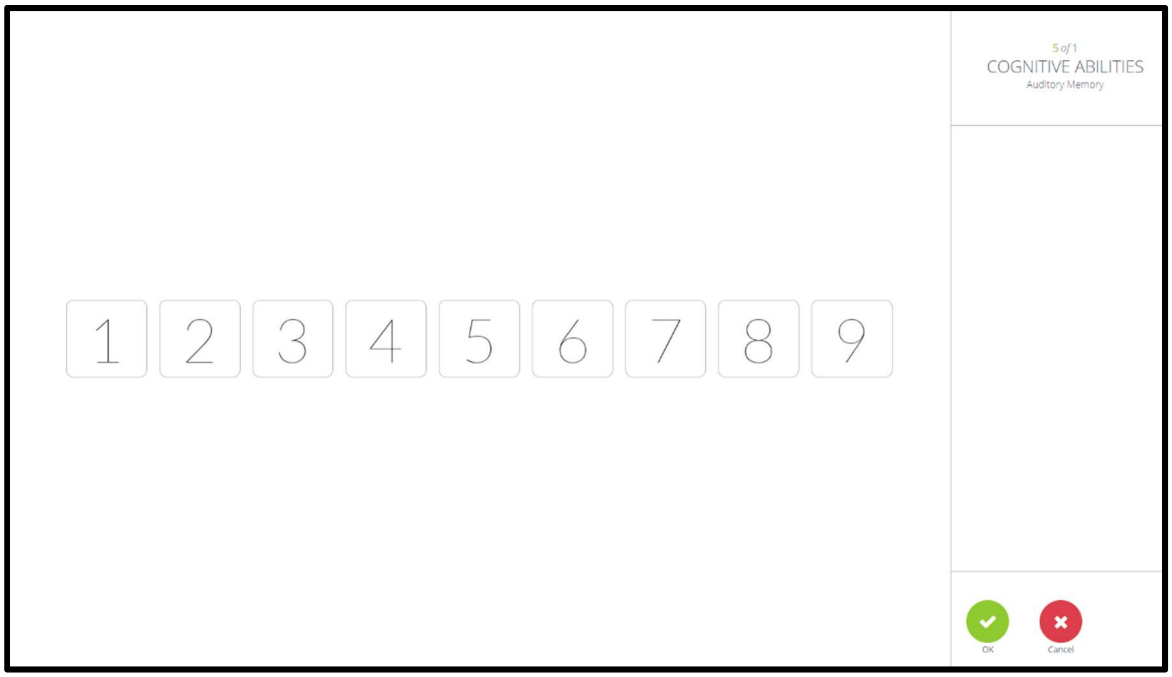
Processing Speed

Directions:

I want to see how quickly you can work. Look at the picture on top.
Now look at these pictures. Some of them are exactly the same as the one on top; others are different. You will be asked to choose all the pictures that are the same as the one on top. If you make a mistake, you can change your answer by clicking on it a second time. Watch how this is done.

Trial item 1:
This picture is not the same as the one on top. Try canceling this choice by clicking on it again.
Good! Now you're ready to start. Go ahead. Find all the pictures that match the one on top. Click on the OK button when you are done.
Good! Now try some more by yourself. Remember - work as quickly and carefully as you can.
(That's not quite right. These are the correct answers. Now try some more. Remember – work as quickly and carefully as you can. Now try some more.)

Auditory Memory with Auditory Distractions

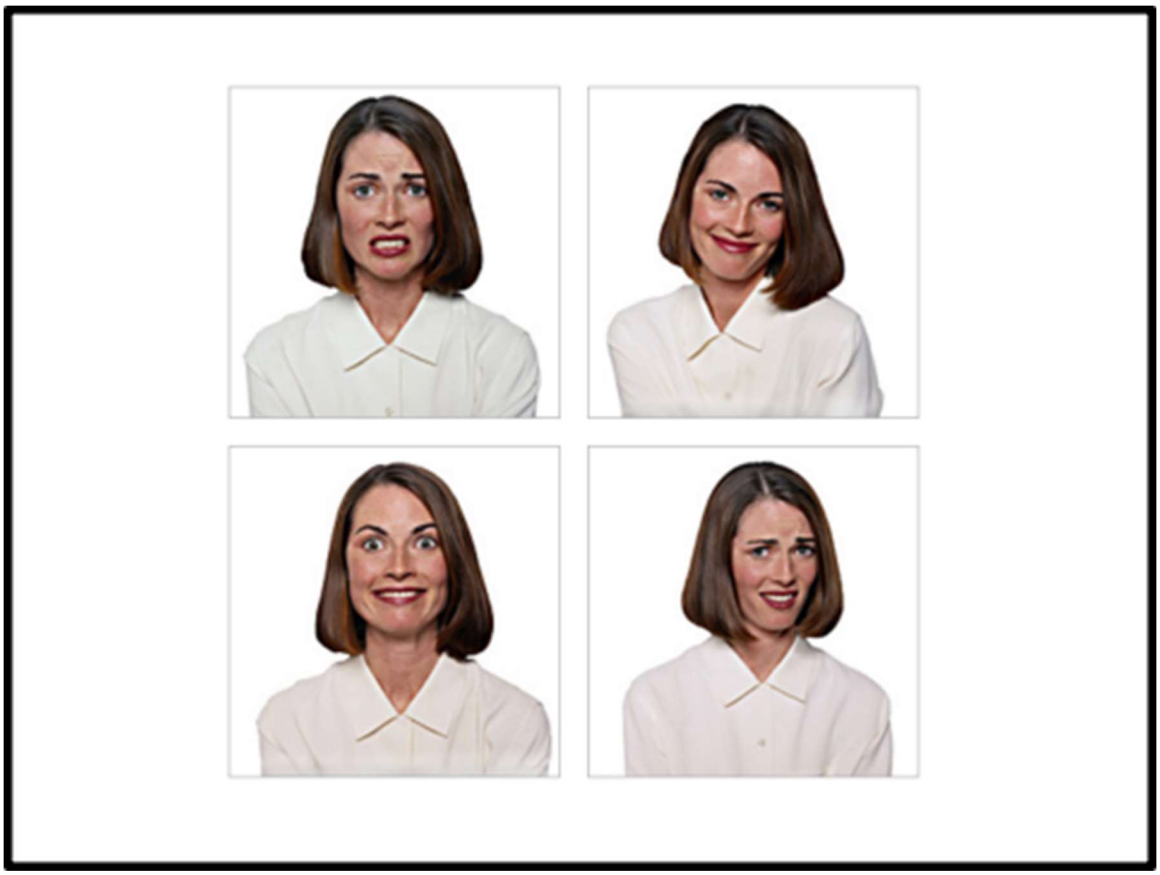


Please note: This time there are sound distractions.

Directions:

Listen carefully. You are going to hear numbers in a certain order. Pay close attention. This time, you will hear other sounds as you hear the numbers. Click on the numbers in the order they were said. Let's begin.

Social Apperception



Directions:

You are going to see pictures of people who are thinking or feeling many different things. Then you will hear someone speak. Choose the person that goes with what you heard. You may use the same picture as your answer more than once. Let's try one together.

Trial item 1: I don't care what you think!
Who do you think said that?
Good! Now try some more. Work as quickly and as carefully as you can.
(That's not quite right. Listen again. The boy who is speaking sounds angry. This picture is the right answer because it shows a boy that is angry. Now try more.)

CHAPTER FOUR:
INTERPRETING TEST RESULTS

Profile Printout

Upon completion of the selected battery (Screening or Standard) and/or supplemental subtests, a MEZURE profile will be printed. Because all scoring is computed by the MEZURE program, the raw scores, scaled scores, standard scores, percentile ranks, and confidence intervals are reported automatically.

Interpreting Confidence Intervals

All test scores are subject to errors of measurement. Such errors occur because assessment is an imprecise science, especially when evaluating complex areas of functioning such as intelligence. It is routine to apply a *standard error of measurement (SEM)* band around the individual's obtained score. This band or **confidence interval** communicates that the *true score* falls somewhere within the calculated range. A confidence interval of 90 percent is reported on the printout. A 90 percent interval provides an ample amount of confidence for most testing purposes.

Analysis of Fluid and Crystallized IQ Discrepancy

It is recommended that examiners routinely compare the Fluid IQ with the Crystallized IQ to determine if there is a statistically significant difference between the two scores. The existence of such a disparity suggests that the individual has true ability differentiation between these two domains. The discrepancy, however, must be large enough to be meaningful and not occur by chance. To facilitate an interpretation of such a difference, the MEZURE printout will indicate whether the Fluid-Crystallized score differentiation is not significant (ns) or is significant at the .05 or .01 level of significance. The existence of statistical significance means that the difference is too large to be attributable to chance fluctuations (i.e., measurement errors).

The Fluid-Crystallized standard score differences necessary for statistical significance are shown in Table 4.1 for children. Values are presented for the 0.01 and 0.05 levels of confidence for the Standard Battery. Values are presented for each one-year age group. These values are graphically displayed by the MEZURE and are flagged if the differences between Fluid and Crystallized intelligence scores are, in fact, significant. Table 4.2 shows the frequency distributions of the Fluid and Crystallized standard score differences obtained in the normative population.

Table 4.1 Crystallized-Fluid Significant Difference Requirements by Age (Standard Battery)

	C-F Diff Scores at	
AGE	**0.05 level**	**0.01 level**
AGE 6	19	25
AGE 7	15	20
AGE 8	14	18
AGE 9	13	17
AGE 10	12	16
AGE 11	13	17
AGE 12	12	15
AGE 13	13	17
AGE 14	13	18
AGE 15	12	16
AGE 16	13	17
AGE 17	12	16
ADULT	12	15

Table 4.2 Cumulative Percentages of Normative Sample with Crystallized-Fluid Difference Scores (Standard Battery)

C-F Diff.	Cum Pct
0	100
1	96.5
2	92.9
3	89.7
4	84.8
5	80.4
6	75.9
7	71.4
8	66.3
9	61.1
10	56.3
11	50.8
12	45.8
13	41
14	35.8
15	31.8
16	27.6
17	24.1
18	21
19	18.1
20	15.5
21	13.2
22	11.3
23	10
24	8.2
25	6.8
26	5.7
27	5.1
28	4.3
29	3.7
30	3.3
31	3
32	2.6
33	2.5
34	2.2
35	2.1
36	2
37	1.8
38	1.8
39	1.7
40	1.7
41	1.6

If the Fluid IQ score is significantly higher, it indicates that the individual is better at solving novel problems that do not require formal training as opposed to completing tasks that are highly influenced by educational and cultural experiences. A significantly higher

Crystallized IQ than Fluid IQ implies the opposite conclusion. (For a more in-depth description of these two domains, refer to Chapter Two.) If no such discrepancy exists, then the individual's abilities are equally developed.

Composite IQ

The Composite IQ score is viewed as a summative index of general intellectual functioning. When the Fluid and Crystallized IQ scores are not significantly different, the Composite IQ is viewed as the most reliable and valid measure of a youngster's global cognitive functioning. **When the Fluid and Crystallized IQ scores are statistically different, then the Composite IQ score should not be used as a measure of general functioning**. That is, the Fluid and Crystallized IQ scores should be treated separately.

Analysis of Subtest Profile

To determine if fluctuations among the subtests are meaningful, we recommend the use of **ipsative** comparisons rather than normed evaluations. That is, specific strengths and weaknesses should be identified for each person relative to his/her performance, not relative to the average performance of children or adolescents of the same age group. To do so, the MEZURE will first calculate the **mean scaled score** for the administered subtests. **If the profile indicates that there is a significant difference between the Fluid and Crystallized IQ standard scores, then the two domains will be calculated separately**. That is, the determination of the subtest mean should be calculated first for the subtests within the Fluid Scale, and then for the subtests within the Crystallized Scale. If there is not a significant difference between the Fluid and Crystallized IQ scores, the mean scaled score will be calculated based on all administered subtests (i.e., 7 subtests for the Standard Battery or 4 subtests for the Screening Battery). The MEZURE requires **at least one standard deviation difference** (i.e., 3 or more scaled score points) to indicate a significant difference between the calculated mean scaled score and an individual subtest. This requirement is to ensure that the disparity is empirically valid and not due to chance.

When the subtest scatter is significant, an individual's personal cognitive strengths and weaknesses (as assessed by the MEZURE) can be interpreted by evaluating the Cattell-Horn-Carroll dimension being assessed (i.e., Gf, Gc, Gsm, Gs, and Gv) as well as the primary factors that influence performance on each subtest (e.g., concentration, distractibility, attention to detail, motivation, richness of educational experiences). Such an approach allows for empirically driven interpretation of the test profile. To facilitate such an analysis of subtest scatter, a description of each subtest is given, followed by a listing of key aspects that affect performance.

Visual Closure (Gf, Gv)

This subtest requires the examinee to identify an object as a picture as it gradually becomes visible on the screen. The examinee is instructed to click on the screen as soon as he/she knows what the picture represents. The examinee then chooses an answer from the four choices that appear on the bottom of the screen.

Visual Closure is a computer unique subtest that accurately measures the examinee's visual closure performance. Performance on this subtest may be influenced by an individual's ability to focus on a task, visual inferencing skills and prior knowledge.

Analogies (Gf, Gv)

This subtest requires the examinee to choose pictures that will complete visual analogies. The examinee selects an answer from five choices displayed below each analogy.

Analogies tap the ability to conceptualize relationships and engage in perceptual reasoning and associative thinking. The examinee is required to view objects from different perspectives, deduce the relationship that exists between them, and apply this information to other objects. Performance may be affected by attention to visual details, concentration, and cognitive flexibility.

Information (Gc)

The Information subtest requires the examinee to answer questions that are based upon a broad range of general knowledge. The examinee must respond to each question by choosing a picture from a template of six. Templates presented during this subtest include landmarks, professions, body parts, foods, animals and environments.

The Information subtest focuses on knowledge, long-term memory, and verbal comprehension. A unique feature of this subtest is the incorporation of various sound effects along with visual prompts. Performance in this subtest may be affected by educational background, interests, and the scope of the examinee's knowledge.

Categorization (Gc)

In this subtest, the subject is asked to identify a picture that is conceptually unrelated to the others in a group. Each item includes four pictures from which the examinee chooses the one that does not have an attribute shared by the remaining three.

Categorization taps into an individual's ability to conceptualize relationships. It involves logical and associative thinking as well as general knowledge.

Visual Memory (Gf, Gsm)

This subtest requires the examinee to recall the order in which a series of shapes appears. After reviewing a sequence of shapes, the examinee responds by clicking on the shapes in the same order as they were shown. The Visual Memory subtest is preceded by a reinforcement task in which the examinee is required to match identical shapes. This familiarizes the examinee with the shapes used in the subtest and insures his/her ability to differentiate between them. If the examinee cannot match the shapes correctly, the examiner will be given the option of omitting this subtest.

This subtest measures short-term visual memory as well as rote memory. Attention to detail and the ability to attend to a task may affect performance on this subtest.

Vocabulary (Gc)

This subtest requires the examinee to identify the picture that corresponds to a word presented auditorily. For each item, the examinee chooses his/her answer from a template of six pictures.

The skills assessed in vocabulary include the examinee's language development, verbal comprehension, and ability to form associations. The quality of an individual's education and his/her prior knowledge may also influence performance on this subtest.

Auditory Memory (Gf, Gsm)

In this subtest, the examinee is required to listen to a series of digits which increases in difficulty as the subtest progresses. When the series is completed, numerals 1-9 appear on the screen. The examinee then clicks on the numbers in the order they were said.

The Auditory Memory subtest assesses short-term auditory memory as well as rote memory. Factors that influence performance on this subtest include concentration, attention, and freedom from distractibility.

Visual Memory with Auditory Distractions (Gf, Gsm)

This subtest is the same as the Visual Memory subtest with the addition of real-life auditory distracters accompanying visual stimuli presentation.

This subtest measures the examinee's visual memory in the presence of auditory distracters. The distractions were designed to simulate those typically encountered in daily life. It requires more attention, concentration, and freedom from distractibility than the Visual Memory subtest.

Auditory Memory with Visual Distractions (Gf, Gsm)

This subtest is the same as the Auditory Memory subtest with the added dimension of visual distracters accompanying digit presentation. It requires more attention, concentration, and freedom from distractibility than the Auditory Memory Subtest.

Auditory Memory with Auditory Distractions (Gf, Gsm)

This subtest is the same as the Auditory Memory subtest with the addition of real-life auditory distracters accompanying digit presentation. It requires more attention, concentration, and freedom from distractibility than the Auditory Memory Subtest.

Processing Speed (Gf, Gs)

This subtest is a timed activity designed to measure an individual's mental processing speed. The examinee is required to identify all the pictures on the screen that are identical to the picture displayed on top.

This subtest is primarily a task of visual matching and visual memory, and the use of a computer to both generate test stimuli and record the response time allows analyses of both accuracy and speed. Accuracy is determined by considering the number of items identified correctly ("hits"), as well the number of items erroneously identified as having been seen previously ("false alarms"). The scoring of this subtest is a composite score considering both factors (hits and false alarms) as well as the timing of the response. Performance may be influenced by attention to detail as well as the ability to concentrate and attend to a task while being timed.

Social Apperception (Gc)

This subtest measures an individual's ability to associate facial and gestural expressions with real-life verbal expression. Items in this subtest require the examinee to listen to someone speak, then choose the person that was the speaker.
Social Apperception probes the examinee's attention to the nuances of social and emotional expression. Knowledge of implied meanings in a variety of verbal and visual prompts is necessary. Attention to detail, social awareness, and range of social experiences may influence performance on this subtest.

CHAPTER FIVE:
DEVELOPMENT AND STANDARDIZATION

Pilot Study I

A preliminary version of MEZURE was pilot-tested with 195 subjects in New York City. The sample included subjects from grade 1 through adult and consisted of 106 females and 89 males with ranging achievement levels. Subjects represented a variety of ethnic backgrounds (approximately 35% European American, 32% Asian American, 15% African American and 10% Hispanic American), socioeconomic levels, geographic regions, and urban/suburban/rural locations, with distributions on each factor reflecting recent demographic data from the US Census Bureau. The test was comprised of 14 subtests collaboratively designed by a team of psychologists, educators, and speech-language pathologists over a 4-year period. Twelve subtests from the pilot test were eventually incorporated into the final standardized version. The Spatial Memory subtest, which was designed to assess short-term spatial memory, was discarded due to its failure to demonstrate significant age-group effects. The Math subtest was eliminated due to lack of enough evidence supporting its role in overall intelligence.

All subjects were administered the MEZURE in its computerized format. Subjects viewed items on a 14-inch monitor and responded via a standard mouse device. Data was automatically collected and stored by the computer, then subjected to a variety of analyses.

Analysis of Pilot Study Data

Analyses performed on the pilot study data included the following:

1) *Classical item analyses* for each subtest yielded item difficulty indices which indicated the proportion of participants who had obtained the correct answer for each test item.
2) *Mean scores* for each subtest were compared for grade levels (Grades 1-2, Grades 7-9, and Grades 10 through Adult). Mean score comparisons were additionally drawn between ethnic groups and genders within each grade level, using Student's t-test to detect any statistically significant differences in group performance.
3) *Answer choice frequencies* (A,B,C,D) was computed for selected subtests in order to identify any possible discrepancies between the intended correct responses and the answers chosen by subjects.

Classical Item Analyses

Traditional item analyses yield indices of the difficulty level of each test item (i.e., how many subjects correctly responded to each item). To effectively differentiate between subjects with varying degrees of understanding in an area, difficulty indices should ideally range from 30-80% for most test items, with several easier items (difficulty indices of 80-100%) and several harder items (difficulty indices of 0-20%). Results of the initial pilot study showed a near 'text-book' range of item difficulties, with most items falling within the 30-80% difficulty range, in addition to several relatively easier and harder items. Items clustering around a specific level of difficulty (within the same 10% range) were subjected to further scrutiny, resulting in the deletion of items considered less representative of the ability in question.

Developmental Appropriateness - Mean Score Comparisons by Age

As one would expect for any assessment of cognitive processes across ages, MEZURE subtest scores were lower for young children than for older subjects. All subtests demonstrated this pattern, indicating that the subtests are assessing either learned skills or maturational processes that develop over time or with experience.

Further analyses compared performances between grades 1 and 2, grades 7-9, and grades 10 through adulthood. Results showed appropriate developmental differences, with younger students scoring lower than older subjects on all subtests. In addition, scores on memory subtests showed age-appropriate digit spans in all groups of subjects.

Gender Comparisons

Comparisons of answer patterns between genders at each age level showed only one subtest, Social Apperception, as having consistent sex differences, with girls scoring higher than boys in elementary and junior high school but demonstrating near-equal performance in high school. These differences reflect patterns of socialization that are well-documented in developmental research literature suggesting that girls / women are "more attuned" at earlier ages to social nuances than are boys / men.

Mean Score Comparisons by Ethnicity

The variety of ethnic backgrounds in the pilot study allowed a cursory examination of the response patterns from each group. For each subtest, the mean scores of each ethnicity were compared to see if there were consistent differences, allowing a preliminary estimation of ethnic bias in the content of test items. Results showed no evidence of ethnic bias. More extensive and reliable bias studies using Mantel-Haenszel analyses (which require a much larger sample size than was available from the pilot population) are described later in this section.

Frequency of Answer Choices

To determine whether there were discrepancies in the answers chosen by students and the choices designated as correct by the authors, the frequency of each answer choice for each item was analyzed. There were no instances where an answer designated as correct by the authors was not chosen by most of the subjects; there were only a few instances where a different answer choice was chosen with a slightly lesser frequency than the choice designated as correct. This suggested that the questions were well-designed, without ambiguous answers.

Supplemental Memory Subtests

Three supplemental memory subtests (Auditory Memory with Auditory Distractions, Auditory Memory with Visual Distractions, and Visual Memory with Auditory Distractions) were of interest in that memory was assessed using the familiar digit-span paradigm, but with the presence of "real-life" distracters and, as such, may provide a more realistic view of a student's processing abilities since the world is rarely devoid of distraction. As expected, mean retention scores in both visual and auditory modalities were lower in young children (grades 1 and 2) than for older subjects.

Test Modifications

A total of 20 items were deleted from the Vocabulary subtest and 13 items from the Information subtest based on collective analyses performed on the pilot study data. Remaining items were reorganized according to their demonstrated order of difficulty. Starting points were established for each level based on the percentage of correct responses for initial items in each subtest. At least 90% of subjects at a given age level had to have passed an item for it to be assigned as the starting point for an age group.

Several modifications were made to visual and auditory stimuli based on feedback collected from examiners and subjects. Specific graphical adjustments included close-up renderings of characters featured in the Information and Vocabulary templates and photo retouching or replacements to incorporate a greater racial and ethnic variety. Revisions were also made to several audio files to clarify instructions. Finally, additional sample questions were created for all subtests to provide more trials for subjects showing difficulty on the initial practice item.

Pilot Study II

A second pilot study was conducted with an additional 212 subjects from New York and Oklahoma, to further assess the MEZURE's reliability. Data was combined with the previous study, yielding a total of 407 subjects, with 178 in Grades 1-2, 116 from Grades 3-6, and 113 from Grades 7 thru adulthood. Gender, ethnicity, socioeconomic levels, geographic regions, and location types were represented in accordance with US Census data.

Homogeneity of Test Items

Classical item analyses from Pilot Study I had already shown item difficulty indices to be within the desired 30-80% difficulty range, with a few easier and harder items. The combined pilot study data (from pilot studies I and II) allowed analyses of ***unidimensionality*** in order to determine how homogeneous each subtest is in terms of what it assesses. Two unidimensionality measures were utilized:

1) ***Cronbach's coefficient alpha*** estimates the "internal consistency" of a test or subtest in terms of the variability of responses given for each item. The underlying premise is that if all items within a subtest are tapping the same construct (or measuring the same ability), then all differences (or "variances") seen in the scoring of that subtest would be due primarily to differences within the test-taking population's true ability and would not be due to ambiguity or poorly constructed items.

2) ***Biserial correlation***, or the item discrimination index, shows the relationship between each item response and the subtest total. A well-constructed test is expected to show a high correlation between each individual item and the subtest.

Both measures of internal consistency are reported as correlational values between 0 and 1, with 1 indicating the highest internal consistency. Subjects of the pilot studies were grouped according to age groups. The MEZURE data yielded values very close to 1.00, indicating that each subtest is carefully constructed and tapping only one construct.

Standardization

Standardization of the MEZURE (based on pilot testing results) was administered to over 5,000 subjects between age 6 and adult. Normative data was derived from 4184 subjects who completed the full array of MEZURE subtests in a standardized fashion. Participants were from a total of over 100 sites representing all types of educational settings, including public, private, parochial, alternative and home schools. Students from both regular and special education classrooms were included in the normative sample.

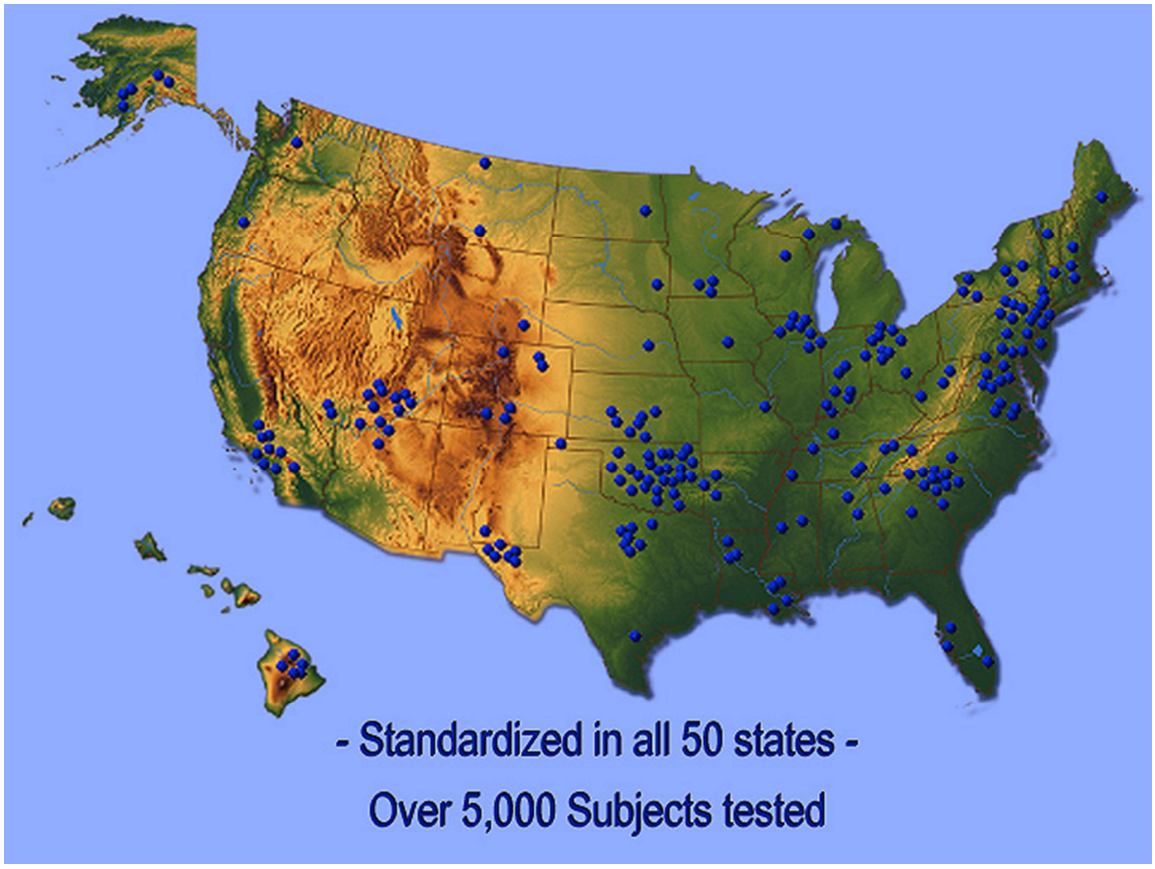
All 50 states were represented in the standardization of the MEZURE and sample proportions of each geographic region within the United States (North Central, Northeast, South, and West) closely matched population proportions reported by the U.S. Census Bureau. Subjects of all ethnicities were included in the sample and were categorized as Asian American, African American, Hispanic American, European American, or Other, which included Native Americans and Eskimo/Aleut Islanders. A special and unprecedented effort was made to include Eskimos / Aleut Islanders in the study. Three professionals traveled with laptops via small plane to remote Alaskan villages to test those remote populations and include them in the sample. As a result, a total of 60 Eskimos / Aleut Islanders and over 140 Native American Indians were included in the normative sample.

To obtain such a large and diverse sample, school psychologists, clinical psychologists, doctoral students, private clinicians, and other trained professionals were queried as to their willingness to participate in the development of the MEZURE. As knowledge of this innovative test became widespread, many additional professionals contacted Assessment Technologies, Inc. on their own initiative to ask permission to participate in the national standardization procedure.

In keeping with established informed consent practices for participation in clinical research, written permission was obtained from all parents and guardians for minor subjects prior to testing. Consent was also elicited informally from the minors themselves, who could terminate the testing at any time. All individual test results were kept strictly confidential.

Multilingual Modules

The MEZURE was adapted for use with examinees speaking languages other than English by professional interpreters and linguists who adapted the original items into equivalent translations while maintaining intended levels of difficulty. All items passed through three stages of editing in which translation accuracy, as well as possible social, cultural, and linguistic differences of each item were considered by an array of translators individually. Interpreters from different areas speaking a language participated in the adaptation process to screen for the possibility of regional differences in dialects or culture. Lastly, the final item pool was not determined until all the language adaptations were completed to allow the culling of items that did not adapt sufficiently well into any language. This allowed the final item pool in the MEZURE to include only those items that were culture-fair and adaptable into multiple language versions.


- Standardized in all 50 states -
Over 5,000 Subjects tested

Demographic Characteristics of Normative Sample

The following tables, 5.1 and 5.2, show the demographic characteristics of the normative population, including sex, age, grade, ethnicity, residence location, geographic region, school type, and parental education. These proportions are compared to, and closely match, those reported by the U.S. Census Bureau data.

Table 5.1
Demographic Characteristics of Normative Sample (N=4184)

N	Sample %	Sex
2066	49.4	F
2118	50.6	M

N	Sample %	Ethnicity
202	4.8	Asian
497	11.9	African-American
2841	67.9	Caucasian
441	10.5	Hispanic
203	4.9	Other

N	Sample %	School
19	.5	Homeschool
42	1.0	Alternative
260	6.2	Private
3863	92.3	Public

N	Sample %	Location
1341	32.1	Rural
2843	67.9	Urban

N	Sample %	Age
154	3.7	6
177	4.1	7
226	5.5	8
267	6.5	9
256	6.3	10
226	5.4	11
304	7.1	12
337	8.2	13
480	11.7	14
509	12.0	15
569	13.6	16
429	10.1	17
250	5.8	Adult Group

N	Sample %	Grade
269	6.4	1
238	5.7	2
215	5.1	3
270	6.5	4
227	5.4	5
290	6.9	6
355	8.5	7
397	9.5	8
639	15.3	9
499	11.9	10
465	11.1	11
320	7.6	Adult Group

N	Sample %	Region
954	22.8	NC
824	19.7	NE
1522	36.4	S
884	21.1	W

N	Sample %	Parent Education
254	6.1	<High School
994	23.8	HS Graduate
604	14.4	1-3 years College
1350	32.3	College Graduate
982	23.5	Unknown

Table 5.2 Comparison of Normative Sample to US Population (N=4184)

Sex	Sample %	US %
F	49.4	48.9
M	50.6	51.1

Ethnicity	Sample %	US %
Asian	4.8	3.5
African-American	11.9	12.1
Caucasian	67.9	72.7
Hispanic	10.5	11.0
Other	4.9	.7

Location	Sample %	US %
Rural	32.1	24.8
Urban	67.9	75.2

Region	Sample %	US %
NE	22.8	23.4
NC	19.7	19.4
S	36.4	35.1
W	21.1	22.1

Region	Sample %	US %
NE	22.8	23.4
NC	19.7	19.4
S	36.4	35.1
W	21.1	22.1

Parent Education	Sample %	US %
<High School	6.1	13.7
HS Graduate	23.8	29.1
1-3 years College	14.4	29.7
College Graduate	32.3	27.5
Unknown	23.1	

Norms Derivation

Little useful or accurate interpretative information about a student's ability can be gathered from raw test scores. It is more informative to interpret test scores in relation to those of other subjects of the same age. To facilitate accurate and meaningful comparisons, raw scores are transformed to scaled scores, standard scores, percentile ranks, and age equivalents. These transformations allow results of various tests to be compared reliably using a common metric, regardless of the lengths of the different tests. MEZURE subtest scores are reported in terms of scaled scores and percentiles, while the summed scaled scores for Brief and Standard Batteries are reported in terms of standard scores and percentile ranks. The methods used to calculate those types of transformed scores are described in this section.

Scaled and Standard Scores

Scaled scores describe a student's test performance relative to that of a normative sample. The transformation of raw scores to scaled scores entails fitting the distribution of raw scores to a distribution having a known mean (10) and standard deviation (3), with scaled score values ranging from 0 to 19. The method described in Angoff (1971) was used to derive scaled scores. In this method, cumulative frequencies and corresponding percentile ranks were computed for each subtest raw score in each 1-year age interval. The raw scores were plotted against the percentile ranks and the resulting curve was smoothed to lessen sampling irregularities. New percentile ranks corresponding to each raw score were then derived from the smoothed curves. For each percentile rank, Z scores were computed, providing the basis for a scaled score distribution having a mean of 10 and standard deviation of 3. Scaled score values for the three-month-intervals were

interpolated from the full-year data. Data for the youngest and oldest ages were extrapolated from the norms curves to allow three-month scores to be interpolated.

Standard scores also describe a student's test performance relative to that of a normative sample, except that the normative distribution had a mean of 100 and standard deviation of 15. Standard scores were derived from raw score data using the same method described above but are based on summed scaled scores. For each child in the normative sample, scaled scores were obtained for each subtest. For the Brief Battery standard scores, the scaled scores for Visual Closure, Visual Analogies, Categorization and Information subtests were summed. For the Standard Battery standard scores, the scaled scores for all seven subtests were summed. Cumulative frequencies and percentile rank of summed scaled scores were obtained, and the distributions plotted and smoothed as described above. As with the scaled scores, mid-year values were interpolated from the full-year data.

In traditionally administered ("paper and pencil") tests, the examiner would consult a norms table to look up the scaled or standard score corresponding to a raw score. The MEZURE program automatically translates the raw scores to scaled and standard scores.

Percentile Ranks

Percentile ranks correspond directly to the normal curve distribution. The use of score transformations described above yields scores that also correspond directly with the normal curve, providing a common metric that can be used reliably to compare scores from different tests (see Anastasi & Urbina, 1998). Percentiles are also provided automatically by the MEZURE program.

Age Equivalents

It is common and useful in scholastic and clinical settings to interpret a student's test performance in terms of functional age, or "age equivalent," which is based on the median scores of a one-year age group. MEZURE age-equivalents were calculated for all possible raw scores for each of the MEZURE subtests. To obtain these scores, the median raw score for each age interval was plotted against the midpoint of that age interval. As for other transformed scores, the curve was smoothed and raw scores corresponding to each one-month age interval were read from the graph; smoothed median scores and standard deviations are shown in Table 5.3. Age equivalents for MEZURE subtests are provided automatically by the MEZURE program.

TABLE 5.3 Smoothed Medians and Standard Deviations for All Subtests, All Ages

	Information Smoothed		Categorization Smoothed		Vocabulary Smoothed	
AGE	*Median*	*SD*	*Median*	*SD*	*Median*	*SD*
6	28	7.9	9	3.9	14	3.8
7	32	9.7	11	4.3	15	4.2
8	36	10.6	12	4.6	17	3.7
9	40	9.8	13	4.7	19	3.6
10	44	10.1	15	4.5	20	3.6
11	48	8.5	15	4.2	22	3.7
12	51	8.4	17	4.4	23	4.1
13	54	7.5	18	3.9	24	4.2
14	56	7.7	19	3.8	25	3.9
15	59	8.9	19	3.9	26	4.6
16	61	8.1	20	4.1	27	4.3
17	64	8.6	20	4.1	28	4.4
Adult	66	9.0	21	4.2	29	4.7

	Visual Analogies Smoothed		Auditory Memory Smoothed		Visual Memory Smoothed	
AGE	*Median*	*SD*	*Median*	*SD*	*Median*	*SD*
6	5	2.9	3	.9	2	.8
7	7	3.9	3	1.0	2	.9
8	9	5.5	4	1.1	3	1.0
9	10	6.8	4	.9	3	1.1
10	12	7.9	4	1.0	3	1.1
11	14	8.2	5	1.1	3	1.1
12	16	8.7	5	.9	4	1.0
13	18	8.2	5	1.1	4	1.2
14	20	7.8	6	1.1	4	1.2
15	22	8.1	6	1.2	4	1.2
16	23	8.2	6	1.2	5	1.3
17	25	8.2	6	1.2	5	1.4
Adult	26	7.6	7	1.3	6	1.4

	Visual Closure Smoothed	
AGE	*Median*	*SD*
6	46	21.75
7	56	21.85
8	66	25.22
9	75	21.51
10	84	21.28
11	92	28.70
12	100	25.35
13	108	20.27
14-Adult	115	21.83

Test Modifications

Based on data collected during the standardization of the MEZURE, ceiling points were established to limit testing time while maintaining the accuracy of test results. Several ceiling variations were applied to subject scores to determine the most effective point at which testing could be terminated. Results supported the use of three consecutive incorrect responses as the ceiling in all subtests aside from the Auditory and Visual Memory subtests in which incorrect responses on both items at any level ends administration of that subtest. Raw score comparisons before and after implementation of subtest ceilings showed high correlations, as demonstrated in Table 5.4.

Table 5.4 Correlations Between Raw Scores Before and After Ceilings

SUBTEST	r
CATEGORIZATION	.99
INFORMATION	.99
VIS. ANALOGIES	.99
VOCABULARY	.99
MEMORY - AUD	-- --*
MEMORY - VIS	-- --*
VIS. CLOSURE	-- --**

* ceilings were not modified in these subtests
** ceilings were not applied to this subtest

CHAPTER SIX:
TECHNICAL DATA

Item Development

Item analyses, utilizing both classical and Item Response Theory methods (Crocker & Algina, 1986), were performed at several times during test development to determine which items to retain for the standardization edition of MEZURE and to determine item sequencing and starting points (within each subtest) for each age group. For each item, the *item difficulty* (proportion of students who correctly answered each item) was calculated. Within an age interval, item difficulties between 0.2 (relatively hard) and 0.8 (relatively easy) are indicative of effective item differentiating power (Anastasi & Urbina, 1997). Also calculated was the *item discrimination index*, determined by the biserial correlation of item response to total score, a measure of how well an item discriminates between students of differing ability that is not dependent on item difficulty (Henryssen, 1971). Items that proved to be problematic were removed from the item pool.

A series of item analyses performed on the pilot study data enabled the authors to streamline the subtests by eliminating the number of items per subtest while still maintaining the desired range of item difficulties. The number of items in each subtest was substantially reduced (by about 70%) to achieve the item composition of the final edition of the MEZURE. A final item analysis was later performed utilizing the entire normative sample to confirm the earlier item selection and sequencing decisions. The large sample population permitted Rasch analyses to be performed at the time of the final item analyses; this confirmed the final item sequencing.

Bias analysis is an important aspect of test development to ensure that a test can be used fairly with children of all backgrounds, ethnicities, and locations. Mantel-Haenszel analyses (reviewed in Nandakumar, et al, 1993) were utilized to determine whether there was evidence of Differential Item Functioning (DIF) between subgroups of the normative population. The effect of DIF, if it is found, is item or test bias - where groups of equal ability but differing on some group characteristic (such as race) will perform differently on the same item. The determinant for item response is then group membership, not individual skill.

Item response comparisons were made between the following groups: African-American/Caucasian, Asian/Caucasian, Hispanic/Caucasian, and Another Ethnicity/Caucasian. Mantel-Haenszel analyses of MEZURE test scores showed that when compared to responses from ethnic majority (Caucasians), none of the test items demonstrated any statistically significant bias against any minority group.

Item response comparisons were additionally drawn between students from the four geographic regions in the United States (Northeast, North Central, South, and West) and

between different residence locations (Urban, Rural). Results of both analyses showed no evidence of either regional or residence location biases.

Reliability

A test's reliability is the degree to which one person's scores on the same test are consistent between different testing occasions (test-retest reliability) or with different examiners (inter-scorer reliability), or the degree to which items are consistent within the test (internal consistency). It is common to define this in terms of temporal stability (test-retest reliability), scorer or examiner stability (inter-rater reliability), and the homogeneity of items in sampling the subject domain (internal consistency). Coefficients greater than .80 are quite acceptable, although values of .90 or greater are extremely strong (Anastasi & Urbina, 1997).

Test-retest reliability for MEZURE was established by administering the test to a sample of students on two occasions (n = 40 to 81; not all students finished all subtests). The time between testing sessions ranged from 3 weeks to 3 months. The resulting correlations between scores from the two testing range from .64 to .92 and are shown in Table 6.1.

Table 6.1 Test-Retest Reliability Coefficients

	r	N
Vocabulary	.92	81
Information	.85	75
Categorization	.70	79
Visual Analogies	.82	79
Visual Closure	.64	79
Auditory Memory	.88	68
Visual Memory	.90	40

Inter-scorer reliability is defined by the degree of consistency in different examiners obtaining the same results with a given student or set of students. This type of reliability determination is not an issue with MEZURE since the only administrator and scorer is the computer, which will not alter in either the procedures used to administer or to score the test. Utilizing the computer in this way effectively eliminates one source of error which has been inherent in traditional testing methods, thereby enhancing the overall reliability of the test.

Internal consistency was determined by two methods. *Split-half reliability* was calculated from correlations between halves of the test, usually between odd-numbered and even-numbered items; these are shown in Table 6.2. *Cronbach's Coefficient Alpha* is another index of internal consistency and is essentially the mean of all possible split-half combinations; the greater the degree of internal consistency, the higher the coefficient (Anastasi & Urbina, 1997). The internal consistency coefficients are shown in Table 6.2. The correlations from both methods are quite high at most ages, an indication that the domains of items sampled by each subtest are homogeneous.

TABLE 6.2 Cronbach's Alpha and Split Half Correlations

	CATEGORIZATION		INFORMATION		VOCABULARY		VISUAL ANALOGIES	
	Alpha	*Split Half**	*Alpha*	*Split Half**	*Alpha*	*Split Half**	*Alpha*	*Split Half**
AVERAGE	.81	.85	.90	.93	.81	.83	.91	.94
AGE 6	.82	.85	.87	.91	.77	.82	.74	.85
AGE 7	.83	.86	.90	.94	.81	.81	.85	.91
AGE 8	.84	.86	.92	.94	.80	.79	.91	.92
AGE 9	.85	.88	.91	.93	.76	.73	.93	.95
AGE 10	.83	.85	.91	.94	.76	.81	.94	.97
AGE 11	.81	.83	.88	.91	.77	.81	.94	.96
AGE 12	.83	.86	.89	.92	.80	.83	.95	.97
AGE 13	.78	.83	.87	.90	.82	.81	.93	.95
AGE 14	.77	.78	.88	.91	.80	.82	.93	.95
AGE 15	.78	.83	.91	.94	.85	.89	.93	.95
AGE 16	.80	.85	.89	.91	.83	.88	.94	.94
AGE 17	.80	.87	.90	.93	.84	.89	.94	.95
ADULT	.82	.85	.92	.94	.87	.90	.92	.94

Split Half * = Spearman-Brown

Standard Error of Measurement

Another index of test reliability is the standard error of measurement (SEM). According to classical test theory, any test score is composed of a person's "true ability" and some error inherent in the measurement techniques (Crocker & Algina, 1987). In order to interpret individual test scores, a measure of this error is useful. Using a reliability coefficient (usually either the test-retest coefficient or Cronbach's Coefficient Alpha), the SEM can be computed with the formula shown below. In that formula, "*SD*" is the standard deviation of test scores for the sample and "r_{tt} "is the reliability coefficient.

$$SEM = SD \sqrt{1- r_{tt}}$$

Determination of the SEM also allows the calculation of confidence intervals within which each child's score can be interpreted. Confidence intervals are based on the premise that if a person were to take a test multiple times (say, 100 times), the test scores would be normally distributed, so that 68% of the time the person's score would be within one SD of the mean of all their scores. Usually, a small degree of error, 5% or 10%, is accepted; the corresponding confidence intervals are calculated accordingly. Relatively low SEMs are an indication of greater test reliability. The SEM and confidence intervals for all subtests are shown in Table 6.3.

Table 6.3 SEM And Confidence Intervals

	CATEGORIZATION		INFORMATION		VOCABULARY		VISUAL ANALOGIES	
	SEM	*95% CI*	*SEM*	*95% CI*	*SEM*	*95% CI*	*SEM*	*95% CI*
OVERALL	1.85	3.62	2.91	5.71	1.96	3.84	2.05	4.03
AGE 6	1.69	3.32	2.91	5.70	1.83	3.58	1.47	2.87
AGE 7	1.77	3.48	3.07	6.02	1.85	3.63	1.54	3.02
AGE 8	1.83	3.58	3.02	5.91	1.65	3.24	1.71	3.36
AGE 9	1.80	3.53	3.03	5.93	1.73	3.39	1.83	3.58
AGE 10	1.85	3.63	2.99	5.85	1.75	3.44	1.94	3.80
AGE 11	1.83	3.59	2.93	5.74	1.81	3.54	2.02	3.96
AGE 12	1.85	3.62	2.78	5.46	1.85	3.62	2.04	4.00
AGE 13	1.82	3.57	2.71	5.32	1.80	3.52	2.10	4.11
AGE 14	1.83	3.59	2.69	5.27	1.75	3.42	2.11	4.14
AGE 15	1.83	3.59	2.73	5.35	1.76	3.46	2.13	4.17
AGE 16	1.84	3.60	2.66	5.22	1.76	3.45	2.10	4.11
AGE 17	1.81	3.54	2.67	5.24	1.75	3.43	2.08	4.08
ADULT	1.78	3.50	2.63	5.15	1.70	3.33	2.07	4.06

*NOTE: VISUAL CLOSURE and MEMORY SUBTESTS did not lend themselves to alphas

Validity

A test's validity is the degree to which the test measures the constructs or traits it purports to measure (Anastasi & Urbina, 1997). Validity is established by examining several empirical parameters that indicate whether the test results obtained in the test's standardization study can be generalized to other populations. The data presented to support the validity of a test enables the practitioner to make the appropriate inferences from test results. Validity data is always viewed in terms of the constructs the test intends to measure.

To establish the validity of MEZURE, comparisons were made between it and other existing, established tests that tapped the same constructs. Several examiners provided scores from other intelligence and achievement tests. The most meaningful comparisons are those made between standardized scores, so those types of scores were used in the validity analyses.

In this section, data are presented that explore three types of validity issues: whether the test items are representative of the intended subject domains (content validity), the degree of correlation between MEZURE scores and other related test scores (criterion-related validity), and the extent to which the MEZURE measures the abilities it was designed to measure (construct validity).

Content Validity

Content validity is the extent to which the test items adequately sample the traits or abilities to be measured and is usually built into the test by the choice of items selected for each subtest. Several psychologists and educators prepared test items that tapped, as much as possible, the discrete skills named by the MEZURE subtests. Item selection was fine-tuned several times during test development by periodic item analyses (detailed in the previous section) to determine which items were kept in the test item pool.

Criterion-Related Validity

Criterion-related (concurrent) validity was established based on correlations between performance on MEZURE and other tests known to tap the same constructs. Correlations between MEZURE and overall cognitive ability and between MEZURE and subtest scores tapping the same skills were derived using scores from *Wechsler's Intelligence Scales for Children – Third Edition (WISC 3)*.

1. MEZURE overall scores (Standard Battery), which indicate general cognitive ability should correlate strongly with other measures of general cognitive ability, such as other intelligence tests. Correlations between the MEZURE and WISC-III are strong, ranging from .70 to .79; these are shown in Table 6.4.

Table 6.4 Correlations **Between Overall Cognitive Scores**

	MEZURE Total
WISC-III VIQ	.70
WISC-III-PIQ	.72
WISC-IIIFSIQ	.79

2. MEZURE scores should correlate strongly with academic achievement test scores, both for subject-related portions of the tests and for overall ability indices. Correlations between the MEZURE subtest scores and achievement test scores were derived with two studies using the Iowa Test of Basic Skills. These comparisons were moderate to strong, ranging from .54 to .74, as shown in Table 6.5.

Table 6.5 Correlations Among Subtest Scores on The MEZURE And the Iowa Test Of Basic Skills: Study I (N = 27)

		MEZURE				
		Vocabulary	*Information*	*Visual Analogies*	*Auditory Memory*	*Visual Memory*
ITBS	*Vocabulary*	.64	.59			
	Reading	.57	.57			
	Social Studies	.65	.54			
	Math			.65	.58	.74

Study II (N = 33)

		MEZURE	
		BRIEF	*STANDARD*
ITBS	*NPR Reading*	.50	.51
	NCE Reading	.48	.54
	NPR Vocabulary	.58	.59
	NCE Vocabulary	.52	.58
	NPR Math	.63	.65
	NCE Math	.59	.63
	NPR Social Studies	.62	.65
	NCE Social Studies	.59	.63
	NPR Math Computation	.55	.53
	NCE Math Computation	.47	.46

		MEZURE						
		Categorization	*Information*	*Auditory Memory*	*Visual Memory*	*Visual Analogies*	*Visual Closure*	*Vocabulary*
ITBS	*NPR Reading*	.61	.43	.17	.26	.38	.04	.36
	NCE Reading	.61	.39	.26	.29	.32	.10	.41
	NPR Vocabulary	.62	.61	.31	.03	.31	.24	.44
	NCE Vocabulary	.57	.55	.37	.04	.24	.25	.50
	NPR Math	.64	.50	.36	.23	.60	.08	.35
	NCE Math	.62	.40	.38	.21	.56	.11	.36
	NPR Social Studies	.61	.61	.27	25	.54	.06	.45
	NCE Social Studies	.61	.57	.30	.20	.49	.05	.50
	NPR Math Computation	.50	.40	.40	.06	.58	.05	.18
	NCE Math Computation	.41	.27	.35	.05	.53	.07	.15

MCOMP = math computation
NCE = normal curve equivalent
NPR = national percentile rank

Construct Validity

Construct validity was established by compiling data from several analyses. Since the skills assessed by the MEZURE are assumed to be the result of exposure to formal learning settings as well as cognitive processes that are the result of neurophysiological maturation, some predictions can be made regarding these comparisons.

1. MEZURE subtest scores should increase as children get older. This relationship is confirmed with the data shown in Table 6.6.

Table 6.6 Smoothed Medians By Age

AGE	*Information*	*Categorization*	*Vocabulary*	*Visual Analogies*	*Auditory Memory*	*Visual Memory*	*Visual Closure*
6	28	9	14	5	3	2	46
7	32	11	15	7	3	2	56
8	36	12	17	9	4	3	66
9	40	13	19	10	4	3	75
10	44	15	20	12	4	3	84
11	48	15	22	14	5	3	92
12	51	17	23	16	5	4	100
13	54	18	24	18	5	4	108
14	56	19	25	20	6	4	115
15	59	19	26	22	6	4	115
16	61	20	27	23	6	5	115
17	64	20	28	25	6	5	115
Adult	66	21	29	26	7	6	115

2. MEZURE subtest and overall scores should be lower for persons having known cognitive impairments as compared to persons without such impairments. Table 6.7 shows that for a small group of students previously diagnosed with the WISC-III as mentally retarded, median MEZURE subtest and battery scores were indeed lower than expected.

Table 6.7 Median MEZURE Scaled and Standard Scores For A Sample Of Mentally Retarded Students (N=9)

	Subtest	Minimum	Maximum	Mean	SD	Expected Scores
STANDARD SCORES	*Categorization*	1	8	4.17	2.04	10
	Auditory Memory	1	13	6.83	3.63	10
	Visual Memory	1	12	6.28	2.61	10
	Visual Analogies	1	8	2.83	2.57	10
	Vocabulary	1	12	4.11	3.36	10
	Visual Closure	1	19	6.11	5.21	10
	Information	1	6	2.44	2.41	10
	BRIEF BATTERY	50	98	76.06	11.46	100
	STANDARD BATTERY	54	79	70.39	6.67	100

3. MEZURE subtest and overall scores should be higher for gifted persons (scores within the 97th percentile or above on standardized cognitive measures) as compared to the normative population. Table 6.8 confirms that for a small group of students previously diagnosed as gifted, median MEZURE subtest and battery scores are higher than expected.

Table 6.8 Median MEZURE Scaled and Standard Scores For A Sample Of Gifted Students (N=15)

	Subtest	Minimum	Maximum	Mean	SD	Expected Scores
STANDARD SCORES	*Information*	11	19	14.87	2.53	10
	Vocabulary	10	19	17.07	2.6	10
	Visual Memory	9	19	14.27	3.97	10
	Auditory Memory	8	19	14.53	3.04	10
	Categorization	7	15	12.07	2.15	10
	Visual Analogies	4	19	15.93	4.04	10
	Visual Closure	8	19	15	2.85	10
	BRIEF BATTERY	113	140	125.47	8.21	100
	STANDARD BATTERY	118	143	128.13	7.66	100

1. MEZURE subtest and overall scores should not differ between students diagnosed as learning disabled and those without learning disabilities since the current practice defines learning disability as a performance deficit in the presence of age-appropriate cognitive ability. Table 6.9 demonstrates that for a group of 44 learning disabled students, median MEZURE scores were not significantly different from the expected medians.

Table 6.9 Median MEZURE Scaled and Standard Scores for A Sample of Learning-Disabled Students (N=44)

Subtest		Minimum	Maximum	Mean	SD	Expected Scores
STANDARD SCORES	*Categorization*	2	14	9.05	3.05	10
	Auditory Memory	2	19	10.34	3.65	10
	Visual Memory	1	16	8.77	3.26	10
	Visual Analogies	1	18	7.55	4.31	10
	Vocabulary	1	17	8.68	3.48	10
	Visual Closure	1	14	7.75	3.70	10
	Information	1	16	8.68	3.92	10
	BRIEF BATTERY	50	138	102.82	20.29	100
	STANDARD BATTERY	59	120	92.86	13.48	100

Internal Validity

Whether there is empirical evidence for a test's score structure is demonstrated by internal validity. This is done by examining the intercorrelations between the subtests, the correlations between the subtests and the total score for the test (the Standard overall scores), and the factor analyses. The subtest intercorrelations and factor analyses are described in Chapter 2. Correlations between the subtests and Standard overall scores are strong, ranging from .53 to .81. These are shown in Table 6.10 below.

Table 6.10 Correlations Between MEZURE Subtest and MEZURE Total Scores

	Categorization	*Vocabulary*	*Information*	*Visual Closure*	*Visual Analogies*	*Visual Memory*	*Auditory Memory*	*MEZURE Total*
Categorization	-- --							
Vocabulary	.52	-- --						
Information	.55	.69	-- --					
Visual Closure	.22	.35	.45	-- --				
Visual Analogies	.44	.53	.51	.31	-- --			
Visual Memory	.33	.30	.27	.04	.35	-- --		
Auditory Memory	.21	.37	.36	.11	.30	.35	-- --	
MEZURE Total	.67	.80	.81	.53	.75	.55	.59	-- --

Correlations between standard scores obtained for the Brief and Standard Batteries should be high if both versions of the test are measuring the same constructs. The correlations obtained from MEZURE's normative sample are high for all ages (r = .086 to 0.92) and for the overall sample (r = 0.91). These values are shown in Table 6.11. The magnitude of these correlations allows MEZURE users to be confident with the test results, regardless of which version of the test was be used.

TABLE 6.11 Correlations Between Brief and Standard Battery Standard Scores

Age	r	N
Overall	0.91	4172
Age 6	0.87	148
Age 7	0.91	175
Age 8	0.93	225
Age 9	0.91	266
Age 10	0.92	256
Age 11	0.93	224
Age 12	0.92	304
Age 13	0.90	339
Age 14	0.88	481
Age 15	0.91	509
Age 16	0.87	567
Age 17	0.90	429
Adults	0.86	249

CHAPTER SEVEN:
SUPPLEMENTAL SUBTESTS

Processing Speed, Social Apperception and Distraction Resistance Scales are not included in either the Brief Battery or Standard Battery but are designed to stand alone. Each subtest reflects distinct processing modalities that may prove helpful in in-depth psychoeducational, neuropsychological, or clinical assessments and for this reason were normed separately. Three subtests, **Visual Memory with Auditory Distractions, Auditory Memory with Auditory Distractions, Auditory Memory with Visual Distractions**, measure visual and auditory short-term memory acquisition under various distracting stimuli. **Processing Speed** evaluates an individual's ability to quickly scan and classify pictures. It is influenced by attention to detail, task persistence, distractibility, and impulsivity. The final supplemental subtest, **Social Apperception**, taps social awareness and attention to facial nuances and to verbal expressions. Administration time for all 5 supplemental subtests is approximately 15 minutes. Further details regarding the supplemental subtests and the types of scores derived for each are discussed below.

Processing Speed

This subtest is a timed activity designed to measure an individual's mental processing speed. The examinee is required to identify all the pictures on the screen that are identical to the picture displayed on top.

This subtest is primarily a task of visual matching; the use of a computer to both generate test stimuli and record the response time allows analyses of both accuracy and speed. Accuracy is determined by considering the number of items identified correctly ("hits"), as well the number of items erroneously identified ("false alarms"). The scoring of this subtest is a composite score considering both factors (errors of omission and errors of comission) as well as the timing of the response. Performance may be influenced by attention to detail as well as the ability to concentrate and attend to a task while being timed.

The Processing Speed subtest was standardized with a sample of 4416 subjects. Demographic characteristics of the sample are shown in Table 7.1.

Table 7.1 Demographics of Processing Speed Sample (N=4416)

Age	N	Sample %
6	243	5.5
7	257	5.8
8	323	7.3
9	415	9.4
10	377	8.5
11	291	6.6
12	296	6.7
13	330	7.5
14	457	10.3
15	465	10.5
16	446	10.1
17	320	7.2
ADULT	196	4.4

Grade	N	Sample %
1	510	11.5
2	340	7.7
3	304	6.9
4	376	8.5
5	318	7.2
6	291	6.6
7	323	7.3
8	353	8
9	516	11.7
10	426	9.6
11	389	8.8
12	270	6.1

Sex	N	Sample %
F	2230	50.5
M	2186	49.5

Race	N	Sample %
Asian	123	2.8
African-American	444	10.1
Caucasian	3350	75.9
Hispanic	390	8.8
Other	109	2.5

Deriving Processing Speed Scores

The nature of the Processing Speed tasks was such that several factors had to be taken into consideration; scoring was not simply a matter of tallying the number of correct responses (hits), or the number of items incorrectly chosen (false alarms), or the time to completion of the subtest (response time). Each of those variables reflect various aspects of cognitive functioning and maturation. Task accuracy (errors of omission vs. the number of correct responses) can reflect a student's attentiveness. The number of errors of commission can reflect impulsivity (or lack thereof). Response time can be an index of underlying information processing mechanisms. For example, if two students of the same age had the same number of hits and false alarms, but differed only in response time, one could infer that the student with the faster response time was a "more efficient" processor of information. Coupled with task accuracy, decreased response time could indicate efficient information processing capabilities; if coupled with errors or omission misses, decreased response time could indicate distractibility; if coupled with errors, decreased response time could indicate impulsivity. Response time is also an index of maturation, in that a younger child, operating optimally, can be expected to have a longer response time than an optimally-operating teen-ager; even when the number of hits and false alarms are the same. The differences in scores among children reflect the cognitive and neurophysiological maturation of the information processing system. As expected, the response times for the MEZURE Processing Speed subtest showed a steady decline as children aged, leveling after age 14; this is shown in Table 7.2.

Table 7.2 Processing Speed Total Time by Age

	6	7	8	9	10	11	12
N	243	257	323	415	377	291	296
Mean	286.87	262.74	232.83	223.25	193.67	185.06	166.65
SD	126.38	118.99	84.86	100.09	61.75	57.87	103.37
	13	**14**	**15**	**16**	**17**	**ADULT**	
N	330	457	465	446	320	196	
Mean	150.18	148.18	145.97	152.07	147.68	146.3	
SD	39.14	41.88	42.65	57.2	48.62	45.98	

Examination of Processing Speed subtest data showed that the Processing Speed Index values increased with age, as would be expected if the statistic is an index of information processing efficiency that is a function of the child's maturation. The mean scores are shown in Table 7.3.

Table 7.3 Processing Speed Index Scores by Age

	6	7	8	9	10	11	12
N	243	257	323	415	377	291	296
Mean	32.24	36.72	37.57	42.11	46.58	49.93	56.1
SD	22.78	24.97	18.06	18.79	16.09	15.61	15.18
	13	**14**	**15**	**16**	**17**	**ADULT**	
N	330	457	465	446	320	196	
Mean	60.29	60.97	62.29	61.47	62.91	62.37	
SD	14.08	15.16	16.07	18.21	17.14	17.37	

Processing Speed Standardization

Standardization of Processing Speed subtest data utilized the methods outlined by Angoff (1971). These are the same methods as used with other MEZURE subtests and Batteries and are previously detailed in this manual. For each one-year age group, a frequency distribution of the Processing Speed Index values was obtained. The median Index values (across all ages) at Z-distribution points (-3Z, -2Z, -1Z, 0Z, +1Z, +2Z, and +3Z) were recorded, plotted and smoothed. A 3rd order polynomial trend line was fitted to the line. Since scaled scores have the same distribution as the Z-distribution, the Processing Speed Index values were read for each scaled score interval. Similarly, Processing Speed Index values were plotted against age for derivation of age equivalents.

As with other MEZURE subtests, the Processing Speed Index' values, and the corresponding scaled scores and age equivalents are computed automatically by MEZURE software.

Social Apperception

This subtest measures an individual's ability to associate facial and gestural expressions with real-life verbal expression. Items in this subtest require the examinee to listen to someone speak, then choose the facial gesture that is most indicatory of the emotion which was expressed auditorily.

Social Apperception probes the examinee's attention to the nuances of social and emotional expression. Knowledge of implied meanings in a variety of verbal and visual prompts is necessary. Attention to detail, social awareness, and range of social experiences may influence performance on this subtest.

The Social Apperception utilized a subset of 4397 students. The demographic characteristics of the sample are shown in Table 7.4.

TABLE 7.4 Demographics of Social Apperception - Normative Sample

Age	N	Sample %
6	264	6
7	258	5.9
8	317	7.2
9	410	9.3
10	376	8.6
11	291	6.6
12	281	6.4
13	322	7.3
14	440	10
15	469	10.7
16	452	10.3
17	318	7.2
ADULT	199	4.5

Grade	N	Sample %
1	520	11.8
2	341	7.7
3	304	6.9
4	369	8.4
5	322	7.3
6	276	6.3
7	316	7.2
8	335	7.6
9	516	11.7
10	436	9.9
11	389	8.8
12	273	6.2

Sex	N	Sample %
F	2214	50.4
M	2183	49.6

Race	N	Sample %
Asian	127	2.9
African-American	447	10.2
Caucasian	3322	75.6
Hispanic	393	8.9
Other	108	2.5

Deriving Social Apperception Scores

Scoring of the Social apperception tasks was based on the empirical difference seen across the various ages and whether the child made the correct identification, as well as the time it took to do so. Of those two variables, the time to make the response reflected the child's level of maturation. The response time decreased as children matured, as expected, and can be thought of as reflective of both the developmental progression of the ability to discern nuances of social and emotional expression, as well as the child's (learned) experiences within his or her environment. The median response times for the MEZURE Social apperception subtest showed a steady decrease as children aged, leveling after age 14. The median response times are shown in Table 7.5.

Table 7.5 Social Apperception Mean Time By Age

	6	7	8	9	10	11	12
N	264	258	317	410	376	291	281
Mean	220.19	196.91	182.3	179	169	163.62	157.81
SD	92.85	67.23	55.93	67.07	43.46	33.75	37.87
	13	**14**	**15**	**16**	**17**	**ADULT**	
N	322	440	469	452	318	199	
Mean	158.02	151.36	151.88	152.07	155.39	159.37	
SD	34.28	29.43	35.11	33.77	34.5	65.33	

To account for both accuracy and timing in the scoring scheme, a Social Index was computed. The Social Index was based on a base score which reflected accuracy plus bonus points based on each item's response time. Bonus points were only awarded if the time was more than one (or two) standard deviation faster than the mean time.

Even though the Social Apperception score incorporates time, it became clear when comparing the scores from Social Apperception and Processing Speed (the two subtests that take time into consideration as a scoring factor) that the Social Apperception scores were not simply measuring the time it took to complete the task. During normative testing, examiners noted that those who seemed to have trouble interpreting the faces in the Social Apperception task took longer to answer even when the answer was correct, while others giving correct answers responded quickly. These observations were substantiated by a study comparing the times and scores of 70 subjects (mean age = 10.77, median = 10.0, SD = 3.88). The Pearson's Product-Moment correlation coefficient between the times to complete the subtests was 0.17, suggesting that students who were fast on one test were not necessarily fast on the other. A Subject's t-Test comparison showed that the differences between the times for the two tests were significant (t_{69}=2.83, p = 0.01). This timing difference can also be seen by inspection of mean times,

seen in Tables PS2 and SA2. These results suggest that the response time component in the Social Index is not due to processing speed efficiency.

Examination of Social apperception subtest data showed that the median Social Index values increased with age. These data are shown in Table 7.6.

Table 7.6 Mean Social Apperception Scores by Age

	6	7	8	9	10	11	12
N	264	258	317	410	376	291	281
Mean	132.64	140.61	147.48	159.11	165.61	174.8	179.22
SD	29.45	27.27	26.61	25.21	24.28	15.51	16.77
	13	**14**	**15**	**16**	**17**	**ADULT**	
N	322	440	469	452	318	199	
Mean	182.06	183.43	181.97	182.84	178.86	183.21	
SD	11.86	16.91	21.79	20.4	27.67	18.06	

Social Apperception Standardization

Standardization of Social apperception subtest data utilized the same methods, outlined by Angoff (1971), as were used with the other MEZURE subtests. For each one-year age group, a frequency distribution of the Social Index values was obtained. The median Social Index values (across all ages) at Z-distribution points (-3Z, -2Z, -1Z, 0Z, +1Z, +2Z, and +3Z) were recorded and plotted and a smoothed trend line was fitted to the plotted line. Since scaled scores have the same distribution as the Z-distribution, the Social Index values could be read for each scaled score interval. Similarly, Social Index values were plotted against age for derivation of age equivalents.

As with other MEZURE subtests, the Social Index values and the corresponding scaled scores and age equivalents are computed automatically by MEZURE software.

Distraction Resistance Scales

Visual Memory with Auditory Distractions (Gf, Gsm)

This subtest is the same as the Visual Memory Subtest with the addition of real-life auditory distracters accompanying visual stimuli presentation.

This subtest measures the examinee's visual memory in the presence of auditory distracters. The distractions were designed to simulate those typically encountered in daily life. It requires more attention, concentration, and freedom from distractibility than the Visual Memory subtest.

Auditory Memory with Visual Distractions (Gf, Gsm)

This subtest is the same as the Auditory Memory subtest with the added dimension of visual distracters accompanying digit presentation. It requires more attention, concentration, and freedom from distractibility than the Auditory Memory Subtest.

Auditory Memory with Auditory Distractions (Gf, Gsm)

This subtest is the same as the Auditory Memory subtest with the addition of real-life auditory distracters accompanying digit presentation. It requires more attention, concentration, and freedom from distractibility than the Auditory Memory Subtest.

The Distraction Resistance Scales utilized a sample of 3977 subjects. The demographic characteristics of the sample are shown in Table 7.7.

Table 7.7 Demographics of Distractibility Scales Normative Sample

Age	N	Sample %
6	153	3.8
7	173	4.4
8	223	5.6
9	264	6.6
10	251	6.3
11	215	5.4
12	281	7.1
13	326	8.2
14	454	11.4
15	461	11.6
16	541	13.6
17	401	10.1
ADULT	234	5.9

Grade	N *	Sample %
1	267	6.9
2	189	4.9
3	211	5.5
4	266	6.9
5	221	5.7
6	263	6.8
7	311	8.1
8	383	10
9	578	15
10	467	12.1
11	413	10.7
12	277	7.2

Sex	N	Sample %
F	1958	49.2
M	2019	50.8

Race	N	Sample %
Asian	196	4.9
African-American	489	12.3
Caucasian	2656	66.8
Hispanic	411	10.3
Other	224	5.6

* some examinees in the normative sample did not have their grades indicated

Deriving Distractibility Scores

The Distraction Resistance Scales are extensions of the Visual and Auditory Memory tasks. In addition to the "pure" memory subtests included in the Standard Battery, memory performance was assessed in the presence of distracters. With the auditory modality, distracters were either auditory noise or a visual presentation. With the visual modality, distracters were auditory noise.

To derive the distraction scales, raw scores for each of the five conditions (pure auditory, auditory with noise, auditory with visual distracters, pure visual, visual with noise) had to

be first normed separately and then transformed to five sets of scaled scores. The mean raw scores for each of the five initial conditions are shown in Table 7.8.

Table 7.8 Mean Distractibility Scores by Age - Initial Conditions

Age 6	Pure Auditory	Aud with Aud Distractions	Aud with Vis Distractions	Pure Visual	Vis with Aud Distractions
N	153	153	153	153	153
Mean	3.88	3.37	3.55	2.49	2.14
SD	0.92	1.34	1.27	0.87	0.86
Age 7	Pure Auditory	Aud with Aud Distractions	Aud with Vis Distractions	Pure Visual	Vis with Aud Distractions
N	173	173	173	173	173
Mean	4.27	3.9	3.98	2.71	2.41
SD	1.01	1.38	1.39	0.93	0.91
Age 8	Pure Auditory	Aud with Aud Distractions	Aud with Vis Distractions	Pure Visual	Vis with Aud Distractions
N	223	223	223	223	223
Mean	4.45	4.04	4.45	3.01	2.7
SD	1.05	1.52	1.36	1.03	1.09
Age 9	Pure Auditory	Aud with Aud Distractions	Aud with Vis Distractions	Pure Visual	Vis with Aud Distractions
N	264	264	264	264	264
Mean	4.82	4.58	4.83	3.36	3.15
SD	0.94	1.44	1.09	1.14	1.18
Age 10	Pure Auditory	Aud with Aud Distractions	Aud with Vis Distractions	Pure Visual	Vis with Aud Distractions
N	251	251	251	251	251
Mean	5.19	4.96	5.18	3.81	3.59
SD	1.01	1.35	1.07	1.14	1.29
Age 11	Pure Auditory	Aud with Aud Distractions	Aud with Vis Distractions	Pure Visual	Vis with Aud Distractions
N	215	215	215	215	215
Mean	5.34	5.13	5.52	3.92	3.69
SD	1.11	1.33	1.13	1.11	1.17
Age 12	Pure Auditory	Aud with Aud Distractions	Aud with Vis Distractions	Pure Visual	Vis with Aud Distractions
N	281	281	281	281	281
Mean	5.55	5.51	5.72	4.25	4.05
SD	0.97	1.2	1.16	1.05	1.14
Age 13	Pure Auditory	Aud with Aud Distractions	Aud with Vis Distractions	Pure Visual	Vis with Aud Distractions
N	326	326	326	326	326
Mean	5.88	5.75	5.87	4.36	4.07

SD	1.13	1.29	1.21	1.23	1.29
Age 14	Pure Auditory	Aud with Aud Distractions	Aud with Vis Distractions	Pure Visual	Vis with Aud Distractions
N	454	454	454	454	454
Mean	6.11	5.95	6.16	4.51	4.34
SD	1.17	1.33	1.4	1.23	1.37
Age 15	Pure Auditory	Aud with Aud Distractions	Aud with Vis Distractions	Pure Visual	Vis with Aud Distractions
N	461	461	461	461	461
Mean	6.01	5.88	6.08	4.59	4.21
SD	1.22	1.46	1.29	1.25	1.46
Age 16	Pure Auditory	Aud with Aud Distractions	Aud with Vis Distractions	Pure Visual	Vis with Aud Distractions
N	541	541	541	541	541
Mean	6.16	6	6.27	4.57	4.26
SD	1.27	1.52	1.41	1.38	1.54
Age 17	Pure Auditory	Aud with Aud Distractions	Aud with Vis Distractions	Pure Visual	Vis with Aud Distractions
N	401	401	401	401	401
Mean	6.35	6.02	6.42	4.69	4.48
SD	1.28	1.54	1.36	1.41	1.54
ADULT	Pure Auditory	Aud with Aud Distractions	Aud with Vis Distractions	Pure Visual	Vis with Aud Distractions
N	234	234	234	234	234
Mean	6.41	6.03	6.36	4.74	4.49
SD	1.37	1.46	1.48	1.51	1.47

Differences were then computed for scaled scores of memory performance without distractions (the "pure" auditory or visual score) and the student's scaled scores in the presence of distracters (auditory with noise, auditory with visual, or visual with noise). These scores were expressed in absolute difference units and can be seen in Tables 7.9 – 7.11.

The resulting three difference scores were then summed to yield a Summed Distraction score which was then normed in the same manner as all other MEZURE subtests. The median Summed Distraction score at each age was plotted and smoothed; scaled scores were read from that curve to yield the Distraction Resistance Index. The mean Distraction Resistance Index can be seen in Table 7.12.

Table 7.9 Scaled Score Absolute Differences Between Conditions: Pure Auditory And Auditory With Noise

	Age 6	Age 7	Age 8	Age 9	Age 10	Age 11	Age 12
N	153	173	223	264	251	215	281
Minimum	0	0	0	0	0	0	0
Maximum	15	11	11	14	16	17	13

Mean	4.17	3.07	3.06	3.64	3.27	2.74	2.41
SD	2.93	2.48	2.57	2.75	2.49	2.49	2.11
	Age 13	**Age 14**	**Age 15**	**Age 16**	**Age 17**	**ADULT**	
N	326	454	461	541	401	234	
Minimum	0	0	0	0	0	0	
Maximum	17	13	16	18	14	15	
Mean	3.02	2.86	3.23	3.09	3.19	3.01	
SD	2.45	2.44	2.46	2.94	2.69	2.91	

Table 7.10 Scaled Score Absolute Differences Between Conditions: Pure Auditory and Auditory with Visual Distractors

	Age 6	**Age 7**	**Age 8**	**Age 9**	**Age 10**	**Age 11**	
N	153	173	223	264	251	215	
Minimum	0	0	0	0	0	0	
Maximum	14	12	11	12	11	12	
Mean	4.16	2.98	2.43	3.22	3.11	2.59	
SD	3.11	2.32	2.35	2.32	2.16	2	
	Age 12	**Age 13**	**Age 14**	**Age 15**	**Age 16**	**Age 17**	**ADULT**
N	281	326	454	461	541	401	234
Minimum	0	0	0	0	0	0	0
Maximum	12	14	13	14	17	14	15
Mean	2.93	3.22	2.75	3.81	3.4	3.62	3.25
SD	2.32	2.32	2.42	2.67	2.81	2.81	2.68

Table 7.11 Scaled Score Absolute Differences Between Conditions: Pure Visual and Visual with Noise

	Age 6	**Age 7**	**Age 8**	**Age 9**	**Age 10**	**Age 11**	
N	153	173	223	264	251	215	
Minimum	0	0	0	0	1	0	
Maximum	12	11	12	12	12	17	
Mean	3.18	3.11	3.72	3.36	4.45	4.05	
SD	2.59	2.05	3.01	2.63	2.97	2.96	
	Age 12	**Age 13**	**Age 14**	**Age 15**	**Age 16**	**Age 17**	**ADULT**
N	281	326	454	461	541	401	234
Minimum	0	0	0	0	0	0	0
Maximum	12	16	18	18	16	17	16
Mean	3.18	5.02	4.31	4.48	5.24	5.07	5.13
SD	2.46	3.35	3.17	3.09	3.45	3.62	3.93

Table 7.12 Summed Scaled Score Differences by Age

	Age 6	**Age 7**	**Age 8**	**Age 9**	**Age 10**	**Age 11**	**Age 12**
N	153	173	223	264	251	215	281
Minimum	1	3	0	1	2	1	0
Maximum	36	29	29	30	31	40	26
Mean	11.51	9.16	9.2	10.22	10.83	9.38	8.52
SD	6.72	4.84	5.97	5.86	5.99	5.45	4.71

	Age 13	**Age 14**	**Age 15**	**Age 16**	**Age 17**	**ADULT**
N	326	454	461	541	401	234
Minimum	2	2	0	0	1	0
Maximum	36	41	36	43	38	41
Mean	11.27	9.93	11.53	11.72	11.88	11.39
SD	6.07	5.83	6.21	6.86	6.79	7.08

Summed Scaled Score Differences = the sum of 3 scaled score differences (absolute differences)

A low Summed Distraction Score yields a high Distraction Resistance Index Scaled Score and indicates that distractions do not impair the subject's task performance. A high Summed Distraction Score yields a low Distraction Resistance Index Scaled Score and indicates that distractions markedly impair the subject's performance. Frequency distributions of the Summed Distraction Score within the normative sample showed that Summed Distraction Score of 14 or greater were associated with the lower 25th percentile (lower quartile), and scores of 18 or greater were associated with the lower 10th percentile. These levels of low performance are of educational and clinical importance, and scores within the lower 10^{th} percentile are flagged by the MEZURE software.

Distraction Resistance Scales Standardization

Standardization of Distraction Resistance Scales subtest data utilized the same methods, outlined by Angoff (1971), as were used with the other MEZURE subtests. First, the five distraction condition scores were normed. For each one-year age group, a frequency distribution of the distraction condition scores was obtained. The median values (across all ages) at Z-distribution points (-3Z, -2Z, -1Z, 0Z, +1Z, +2Z, and +3Z) were recorded and plotted and a smoothed trend line was fitted to the plotted line. Since scaled scores have the same distribution as the Z-distribution, the distraction condition scores values could be read for each scaled score interval. Next, three sets of difference scores were derived for each student based on the absolute difference between the "pure" and the distracted states. These scaled score differences were summed to yield a Summed Distraction score, which was then normed as above and scaled scores were then read for each Distraction Resistance Index value.

As with other MEZURE subtests, all Distraction Resistance Scales score values and the corresponding scaled scores and age equivalents are computed automatically by MEZURE software.

Reliability and Validity

Reliability

A test's reliability is the degree to which one person's scores on the same test are consistent between different testing occasions (test-retest reliability) or with different examiners (inter-scorer reliability), or the degree to which items are consistent within the test (internal consistency). It is common to define this in terms of temporal stability (test-

retest reliability), scorer or examiner stability (inter-rater reliability), and the homogeneity of items in sampling the subject domain (internal consistency). Coefficients greater than .80 are quite acceptable, although values of .90 or greater are extremely strong (Anastasi & Urbina, 1997).

Test-retest reliability for MEZURE was established by administering the test to a sample of students on two occasions (n = 40 to 81; not all subjects finished each of the subtests). The time between testing sessions was between 1 month and 3 months. The resulting correlations between scores from the two tests are 0.81 (Social Apperception) and 0.85 (Processing Speed). These are shown in Table 7.13.

Table 7.13 Test-Retest Coefficients

Subtest	r
Social Apperception	0.81
Processing Speed	0.85

Inter-scorer reliability is defined by the degree of consistency in different examiners obtaining the same results with a given student or set of students. This type of reliability determination is not an issue with MEZURE since the only administrator and scorer is the computer, which will not alter in either the procedures used to administer or to score the test. Utilizing the computer in this way effectively eliminates one source of error which has been inherent in traditional testing methods, thereby enhancing the overall reliability of the test.

To further test inter-computer reliability, comparisons were made between results obtained from a separate group of subjects (n=47) using different hardware configurations, namely whether scores were affected using laptop computers rather than desktop computers, small monitors rather than large monitors, as well as whether different speakers, mice or computer processor speed had any effect on the results of the MEZURE. Pearson product-moment correlations (r = 0.06) showed that there was virtually no relationship between the type of computer hardware used and the scores achieved by the students.

Internal consistency is usually determined by two methods, Split-half reliability and Cronbach's Coefficient Alpha, neither of which are appropriate for tests in which time determines the subtest score (Anastasi & Urbina, 1997).

Standard Error of Measurement

Another index of test reliability is the standard error of measurement (SEM). According to classical test theory, any test score is composed of a person's "true ability" and some error inherent in the measurement techniques (Crocker & Algina, 1987). To interpret individual test scores, a measure of this error is useful. Using a reliability coefficient (in the cases of the Supplemental subtests which are speeded tests, the test-retest coefficient was used). The SEM can be computed with the formula shown below. In that

formula, "*SD*" is the standard deviation of subtest scaled scores and "r_{tt} "is the reliability (test-retest) coefficient; the SEMs can be seen in Table 7.14.

$$SEM = SD \sqrt{1 - r_{tt}}$$

Table 7.14 Supplemental Tests SEM

	Processing Speed	Social Apperception
AGE 6	8.82	15.05
AGE 7	9.67	13.37
AGE 8	6.99	12.74
AGE 9	7.28	12.46
AGE 10	6.23	11.57
AGE 11	6.05	8.70
AGE 12	5.88	8.68
AGE 13	5.45	6.46
AGE 14	5.87	8.29
AGE 15	6.22	11.03
AGE 16	7.05	10.09
AGE 17	6.64	13.69
ADULT	6.73	9.08

Validity

A test's validity is the degree to which the test measures the constructs or traits it purports to measure (Anastasi & Urbina, 1997). Validity is established by examining several empirical parameters that indicate whether the test results obtained in the test's standardization study can be generalized to other populations. The data presented to support the validity of a test enables the practitioner to make the appropriate inferences from test results. Validity data is always viewed in terms of the constructs the test intends to measure.

Usually, at least three types of validity issues are of importance: whether the test items are representative of the intended subject domains (content validity), the degree of correlation between MEZURE scores and other related test scores (criterion-related validity), and the extent to which the MEZURE measures the abilities it was designed to measure (construct validity).

Content Validity

Content validity is the extent to which the test items adequately sample the traits or abilities to be measured and is usually built into the test by the choice of items selected for each subtest. Several psychologists and educators prepared test items that tapped, as much as possible, the discrete skills named by the MEZURE subtests. Item selection

was fine-tuned several times during test development by periodic item analyses (detailed in the previous section) to determine which items were kept in the final test item pool.

Criterion-Related Validity

Concurrent validity (a type of criterion-related validity) is usually established based on correlations between performance on MEZURE and other tests known to tap the same constructs. Social Apperception seems to have no directly comparable companion test currently, nor is the exact paradigm employed by Distraction Resistance Scales used by other tests.

A separate sample of subjects (n = 37, mean age = 14.8 years) was administered both the MEZURE and WISC-3 so that scores could be compared. MEZURE Processing Speed scores correlated moderately (r = 0.63) with the Processing Speed Factor score of the WISC-3. A comparison between MEZURE's Processing Speed and Wechsler's Processing Speed Factor scores may not be appropriate since the tasks are structured quite differently. It is important, always, to look at the tasks themselves, not just at the subtest titles when determining whether subtests are equivalent.

Comparison of Normative Sample to Clinical Sample

To further establish subtest validity, the performance of the normative sample was compared to that of a clinical sample consisting of 98 individuals with formal diagnoses for such conditions including, but not limited to, Learning Disabilities, Mental Retardation, Attention Deficit Disorder, and emotional disturbances. The only subgroup large enough to yield statistically sound analyses was the Learning-Disabled group (LD, N = 28). The demographic characteristics of the entire clinical sample are shown in Table 7.15.

Table 7.15 Demographics of Distraction Resistance Scales: Clinical Sample

Age	N	Sample %
6	6	6.1
7	7	7.1
8	19	19.4
9	19	19.4
10	5	5.1
11	6	6.1
13	3	3.1
14	6	6.1
15	4	4.1
16	10	10.2
17	11	11.2
ADULT	2	2

Grade	N	Sample %
1	7	5.1
2	17	17.3
3	23	23.5
4	7	7.1
5	6	6.1
6	2	2
8	3	3.1
9	11	11.2
10	10	10.2
11	10	10.2
ADULT	2	2

Sex	N	Sample %
F	41	41.8
M	57	58.2

Race	N	Sample %
Asian	6	6.1
African-American	16	16.3
Caucasian	69	70.4
Hispanic	5	5.1
Other	2	2

Scaled scores for each of the five conditions, absolute differences between the scaled scores of "pure" and distracted conditions, and Distraction Resistance Index are shown in Table 7.16. While the Distraction Resistance Index of the clinical group is not markedly different from the normative population, the scores for the LD group are nearly 1 sd below the mean. The Distraction Resistance Index scores and the corresponding scaled scores and percentile ranks are shown in Table 7.17.

In addition, Processing Speed and Social Apperception scores for the clinical sample were compared to those of the normative population. Both subtests' mean scores in the clinical group were about 1 standard deviation below the normative mean for the groups' mean age (12.8 years). These score comparisons are seen in Tables 7.18 and 7.19.

Table 7.16 Distraction Resistance Scales: Clinical Sample
Scaled Scores for Five Initial Conditions

OVERALL	Pure Auditory	Aud with Aud Distractions	Aud with Vis Distractions	Pure Visual	Vis with Aud Distractions
N	98	98	98	98	98
Minimum	1	1	1	1	1
Maximum	19	18	18	16	19
Mean	8.61	7.64	8	8.79	7.02
SD	4.16	3.94	3.82	3.41	3.89

* all clinical diagnostic categories together

LD SAMPLE	Pure Auditory	Aud with Aud Distractions	Aud with Vis Distractions	Pure Visual	Vis with Aud Distractions
N	28	28	28	28	28
Minimum	1	1	1	4	1
Maximum	14	18	18	14	13
Mean	8.86	7.25	7.46	7.86	4.82
SD	3.32	4.58	4.26	2.56	3.15

Table 7.17 Distraction Resistance Scales: Clinical Sample
Absolute Differences of Scaled Scores for Three Distracted States

	OVERALL			LD SAMPLE		
	Aud-Aud w/N	Aud-Aud w/V	Vis-Vis w/N	Aud-Aud w/N	Aud-Aud w/V	Vis-Vis w/N
N	98	98	98	28	28	28
Minimum	0	0	0	0	0	0
Maximum	11	12	12	11	10	12
Mean	3.05	2.94	3.49	3.96	3.54	5.75
SD	2.7	2.42	3.21	3.19	2.7	3.41

Table 7.18 Processing Speed: Clinical Sample

OVERALL	TimeSUM	HitSUM	MissSUM	FASUM	PS_SCORE
N	94	94	94	94	94
Minimum	13	0	0	0	0
Maximum	434	93	93	30	101.85
Mean	184.21	82.59	9.34	7.87	45.39
SD	81.17	18.25	16.83	5.33	18.45

Table 7.19 Social Apperception - Clinical Sample

OVERALL	BASE	BONUS	SOC_SCORE
N	90	90	90
Minimum	44	7	74
Maximum	164	41	203
Mean	138.84	26.13	164.98
SD	22.22	8.89	23.92

ACKNOWLEDGEMENTS

We would like to express our deep gratitude to the Director of the Clinical Development Team, Dr. LeAdelle Phelps for her constant and essential guidance throughout this project; the MEZURE would not be what it is today without her foresight and expertise.

Director of the Clinical Development Team

LeAdelle Phelps

Brief Biographical Statement

LeAdelle Phelps, PH.D. is Professor and Director of the School Psychology Program in the Department of Counseling and Educational Psychology at the State University of New York at Buffalo. She is a Fellow of APA Division 16 (School Psychology) and a member of Division 40 (Clinical Neuropsychology) and 54 (Pediatric Psychology). Her scholarship is evident in the more then 50 journal articles and book chapters she has published on such diverse health-related topics as eating disorders, prenatal alcohol and cocaine exposure, and lead poisoning, Likewise, she has written extensively on assessment and measurement issues. She authored the **Phelps Kindergarten Readiness Scale,** ***a nationally standardized assessment tool evaluating learning readiness aptitudes predictive of later school achievement. Her latest book entitled:*** **Health Related Disorders in Children and Adolescents: A Guidebook for Understanding and Educating** ***was published in May 1998. Another book*** **Pediatric Psychopharmacology: Facilitating Collaborative Practices** ***is scheduled for publication in 2001. She is Editor of*** **Psychology in the Schools** ***and serves on the editorial boards of*** **School Psychology Quarterly, School Psychology Review,** ***and*** **Journal of Psychoeducational Assessment.** ***National leadership roles include chairing the Council of Directors of School Psychology Programs (CDSPP), chairing the American Psychological Association Division 16's Task Force on Training Standards in School Psychology, serving as a liaison to the APA Board of Educational Affairs, and being a member of the APA Council of Chairs of Training Councils. She teaches such graduate courses as*** **Counseling with Children** ***and*** **Advanced Personality Assessment.** ***She maintains a private practice specializing in neuropsychological assessment and therapeutic interventions with children and adolescents. Previously, she was Chief Psychologist at the Traumatic Head Injury Clinic located at Still Hospital, Jefferson City, Missouri.***

Director of the Statistical Analyses Team

Nancy A. Martin

Brief Biographical Statement

Nancy A. Martin, Ph.D., is currently an adjunct faculty member at Dominican University, San Rafael, CA, in the departments of Psychology and Biology, where she teaches Neuroanatomy, Physiological Psychology, Learning and Cognition, Statistics, and Research Writing. Her research at University of California, Davis, CA, documented the latent effects of prenatal drug exposure on children's cognitive development (problem solving) and brain function (as measured by evoked response potentials). She is currently a member of Cognitive Neuroscience Society and has presented her research at meetings of the Cognitive Neuroscience Society and the American Academy of Child and Adolescent Psychiatry.

For the last 12 years she has been a test development consultant for Academic Therapy Publications, Novato, CA, and, most recently, with Assessment Technologies Inc., New York, NY. She is co-author of the **Quick Neurological Screening Test-II** ***and has provided normative and statistical assistance for*** **Learning Efficiency Test-II, Webster Pre-Kindergarten Screen, Expressive One-Word Picture Vocabulary Test-3rd Edition, Spadafore ADHD Rating Scale, Figurative Language Interpretation Test, Motor-Free Visual Perception Test-Revised, and Motor-Free Visual Perception Test-Vertical,** ***among others.***

She has provided test interpretation workshops for school psychologists, educational specialists and teachers and has also provided parent workshops exploring child behavior in relation to brain development.

She has taught Psychological Assessment, Cognitive Psychology, and Developmental Psychology at University of California, Davis, and Developmental Psychology at San Francisco State University, San Francisco, CA.

We would like to thank the entire MEZURE Development Team for their excellent work and dedication throughout this project. Special mention needs to be made of the following individuals:

Michael Hirsch, Project Coordinator
Marian Green, Director Special Education
Tzippy Friedlander, Director Computer Development
Naomi Horn – Director Speech and Hearing Clinitians
Mary Krepel– Director of Graphic Design
Linda Halperin – Director of School Psychology

We would like to express our gratitude to Dr. Jerome M. Sattler, Dr. Bruce Bracken, and Dr. Donald D. Hammill for their invaluable advice. We would like to thank the following professionals in each state for their outstanding effort in data collection for the MEZURE norming:

Alaska	Pat Patterson	*North Carolina*	Todd Morton
	Jeff Tysinger	*North Dakota*	Ken Carlenson
Alabama	Martha Hardin	*Nebraska*	Marylyn Bechtel
Arkansas	Sandra Sanders	*New Hampshire*	Jane Plamondon
Arizona	Clay Mills	*New Jersey*	Bill Challenger
California	Scott Savage		Linda Halperin
Colorado	David Cantrell	*New York*	Beth Krieger
	Achilles Bardos		William Cohen
Connecticut	Kent Gemmell		Steven Sage
Delaware	Allen Kleeban	*New Mexico*	Enedina Vazquez
Florida	Robert Silver	*Nevada*	Clay Mills
Georgia	Dovida Levine	*Ohio*	Brian Barnhart
	William Blackerby	*Oklahoma*	Cynthia Boykin
Hawaii	Jacqueline Brittian		Linda Palmer
Iowa	Rex Shariari		Lin Cagel
Idaho	Virginia Allen		Heather Adams
Illinois	Lisa Corradino	*Oregan*	Phil Bowser
	Robert Clark	*Pennsylvania*	Lawrence Billardi
Indiana	Cythnia Jenner	*Rhode Island*	Rina Jurkowitz
	Susan McDowell	*South Carolina*	Todd Morton
	Glenda Love	*South Dakota*	James Kappen
Kansas	Pam Bush	*Tennessee*	William Tracy
Kentucky	Jan Roberson		Robertson
Louisiana	Gail Gillespie		Peter Hodges
	Gerry Tobacyk	*Texas*	Megan Hudson
Massachusetts	Ken Durant		Trey Asbury
Maryland	Gregory Ford	*Utah*	Clay Mills
	Ellen Hickey	*Virginia*	Duane Harrell
Maine	Margarita Marnick		Barbara Lafever
Michigan	Lynette Borree	*Vermont*	Deborah Wallis
Minnesota	Ralph Kudela	*Washington*	Peiling Farajallah
Missouri	Michelle Lubbert	*Wisconsin*	Paul Hesse
Mississippi	Anita Craft	*West Virginia*	Stephanie Oberly
Montana	Evelyn Lamont	*Wyoming*	Bill McLean

We would also like to mention our appreciation to all the others who participated in the nationwide data collection for the MEZURE but have not been listed here by name.

REFERENCES

Anastasi, A. & Urbina, S. (1997). *Psychological Testing, 7th Edition.* Upper Saddle River, NJ: Prentiss-Hall, Inc.

Anastasi, A. (1988). *Psycyhological testing* (6th ed.). New York: Macmillan.

Angoff, W. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement, 2nd Edition.* Washington, D.C.: National Council on Education.

Carroll, J. B. (1972). Stalking the wayward factors. *Contemporary Psychology, 17,* 321-324..

Carroll, J. B. (1989). Factor analysis since Spearman: Where do we stand? What do we know? In R. Kanfer, P. L. Ackerman, & R. Cudeck (Eds.). *Abilities, motivation, and methodology The Minnesota symposium on learning and individual differences* (pp. 43-67). Hillsdale, NJ: Erlbaum.

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

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin, 38,* 592.

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action.* Boston: Houghton Mifflin.

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action.* New York: North-Holland.

Cattell, R. B., & Horn, J. L. (1978). A check on the theory of fluid and crystallized intelligence with description of new subtest designs. *Journal of Educational Measurement, 15,* 139-164.

Cole, N. & Moss, P. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational Measurement, 3rd Edition.* New York, NY: Macmillan.

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* New York, NY: Harcourt Brace Jovanovich College Publishers.

Guilford, J. P. (1954). *Psychometric Methods, 2nd Edition.* New York, NY: McGraw-Hill.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991*). Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Publications.

Henryssen, S. (1971). Gathering, analysing, and using data on test items. In R. L. Thorndike (Ed.), *Educational Measurement, 2nd Edition.* Washington, D.C.: National Council on Education.

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H.Wainer & H.I.Braum (Eds.), *Test Validity.* Mahwah, NJ: Lawrence Erlbaum Associates.

Horn, J. L. (1965). *Fluid and crystallized intelligence: A factor analytic and developmental study of the structure among primary abilities.* Unpublished Doctoral dissertation. University of Illinois, Champaign.

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

Horn, J. L. (1972). The structure of intellect: Primary abilities. In R. M. Dreger (Ed.), *Multivariate personality research* (pp. 451-511). Baton Rouge, LA: Claitor's.

Horn, J. L. (1976). Human abilities: A review of research and theory in the 1970s.

Annual Review of Psychology, 27, 437-485.
Horn, J. L. (1985). Remodeling old models of intelligence: Gf-Gc theory. In B. B. Wolman (Ed.), *Handbook of multivariate psychology* (pp. 645-685). New York: Academic Press.
Horn, J. L. (1989). Models of intelligence. In R. Linn (ed), *Intelligence: Measurement, Theory, and public policy* (pp. 29-73). Urbana, IL" University of Illinois Press.
Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligence. *Journal of Educational Psychology, 57,* 253-270.
Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica, 26,* 107-129.
Horn, J. L., & Stankov, L. (1982). Auditory and visual factors of intelligence. *Intelligence, 6,* 165-185.
Iowa Tests of Basic Skills. (1996). Itasca, IL: Riverside Publishing.
MacMillan, N. A. & Creelman, C. D. (1991). Detection theory: A user's guide. Cambridge: Cambridge University Press.
Messick, S. (1989). Validity. In In R. L. Linn (Ed.), *Educational Measurement, 3rd Edition.* New York, NY: Macmillan.
Miller, J. (1996). The sampling distribution of d'. Perception & Psychophysics, 58 (1), 65-72.
Nandakumar, R., Guttling, J. & Oakland, T. (1993). Mantel-Haenszel methodology for detecting item bias: An introduction and example using the guide to the assessment of test behavior. *Journal of Psychoeducational Assessment, 11,* 108-119.
Spearman, C. (1927). *The abilities of man: Their nature and measurement.* London: Macmillan.
Swaminathan, H. & Rogers, J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361- 370.
U. S. Bureau of the Census. (1998). *Statistical Abstract of the United States.* Washington, D.C.: U. S. Department of Commerce.
Wechsler, D. (1991). *Wechsler Intelligence Scale for Children, 3rd Edition.* San Antonio, TX: The Psychological Corporation.