



**Watson-Glaser™ II**  
**Critical Thinking Appraisal**

**Technical Manual**  
**and User's Guide**

**Goodwin Watson and Edward M. Glaser**

Copyright © 2009 NCS Pearson, Inc. All rights reserved.

**Warning:** No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the copyright owner.

**Pearson**, the **Pearson** logo, **TalentLens** and **Watson-Glaser** are trademarks, in the U.S. and/or other countries, of Pearson Education, Inc. or its affiliate(s).

Portions of this work were previously published.

Printed in the United States of America.

P.O. Box 599700 San Antonio, TX 78259 800.211.8378 [www.TalentLens.com](http://www.TalentLens.com)



# Table of Contents

## Chapter 1

Introduction . . . . .	1
------------------------	---

## Chapter 2

Critical Thinking . . . . .	2
RED Model . . . . .	2
Recognize Assumptions . . . . .	2
Evaluate Arguments . . . . .	3
Draw Conclusions . . . . .	3

## Chapter 3

Development of the Watson-Glaser II . . . . .	4
Criteria for Item Selection . . . . .	4
Improve Subscale Structure and Interpretability . . . . .	4
Improve Score Distribution and Maintain Score Reliability . . . . .	4
Improve Scenarios and Items: Business Relevance, Global Applicability, Currency of Controversial Issues . . . . .	5
Business Relevance . . . . .	5
Global Applicability . . . . .	5
Currency of Controversial Scenarios and Items . . . . .	5
Maintain Reading Level . . . . .	5
Maintain Short Form Test Administration Time . . . . .	5
Test Development Stages . . . . .	5
Conceptual Development Stage . . . . .	6
Item Writing and Review Stage . . . . .	6
Pilot Stage . . . . .	6
Calibration Stage . . . . .	6
Tryout Stage: Form D . . . . .	7
Standardization Stage: Form D . . . . .	8
Standardization Stage: Form E . . . . .	8

## Chapter 4

Equivalent Forms . . . . .	9
Equivalence of Watson-Glaser II Forms D and E with Previous Forms . . . . .	9
Equivalent Raw Scores . . . . .	9
Equivalence of Computer-Based and Paper-and-Pencil Forms . . . . .	13

## Chapter 5

<b>Norms</b> . . . . .	<b>14</b>
Background on Norms . . . . .	14
Pearson Norms . . . . .	14

## Chapter 6

<b>Evidence of Reliability</b> . . . . .	<b>16</b>
Definitions of Reliability and Standard Error of Measurement . . . . .	16
Test-Retest Reliability . . . . .	16
Internal Consistency Reliability. . . . .	17

## Chapter 7

<b>Evidence of Validity</b> . . . . .	<b>19</b>
Content Validity . . . . .	19
Factor Structure. . . . .	19
Confirmatory Factor Analysis . . . . .	19
Subscale Intercorrelations . . . . .	23
Convergent Validity: Cognitive Ability Measures . . . . .	23
Watson-Glaser II and WAIS®-IV . . . . .	25
Convergent Validity: Open Mindedness . . . . .	27
Big 5. . . . .	27
Myers-Briggs Type Indicator (MBTI)® . . . . .	27
Golden Personality Type Profiler®. . . . .	28
Discriminate Validity . . . . .	28
Criterion-Related Validity. . . . .	29
Prior Evidence of Criterion-Related Validity. . . . .	30
Studies Using the Watson-Glaser II. . . . .	32
Summary . . . . .	33

## Chapter 8

<b>User's Guide</b> . . . . .	<b>34</b>
<b>Directions for Administration</b> . . . . .	<b>34</b>
Computer-Based Administration . . . . .	34
Preparing for Administration . . . . .	34
Testing Conditions . . . . .	34
Answering Questions. . . . .	34
Administering the Test . . . . .	34
Technical Issues . . . . .	35

Scoring and Reporting . . . . .	35
Paper-and-Pencil Administration . . . . .	35
Preparing for Administration . . . . .	35
Testing Conditions . . . . .	35
Materials Needed to Administer the Test . . . . .	35
Answering Questions . . . . .	36
Administering the Test . . . . .	36
Timed Administration . . . . .	37
Untimed Administration . . . . .	37
Concluding Administration . . . . .	37
Scoring . . . . .	38
Scoring With the Hand-Scoring Key . . . . .	38
Machine Scoring . . . . .	38
<b>Additional Considerations for Administration . . . . .</b>	<b>38</b>
Test Security . . . . .	38
Differences in Reading Ability—English as a Second Language . . . . .	39
Accommodating Examinees with Disabilities . . . . .	39

## Chapter 9

<b>Using the Watson-Glaser II as an Employment Selection Tool . . . . .</b>	<b>40</b>
<b>Employment Selection . . . . .</b>	<b>40</b>
<b>Fairness in Selection Testing . . . . .</b>	<b>40</b>
Legal Considerations . . . . .	40
Group Differences/Adverse Impact . . . . .	40
Monitoring the Selection System . . . . .	41

## Chapter 10

<b>Watson-Glaser II Reports . . . . .</b>	<b>42</b>
<b>The Watson-Glaser II Profile Report . . . . .</b>	<b>42</b>
Interpreting Test Results Using Norms and Percentiles . . . . .	42
Using Local Norms . . . . .	42
Using Pearson Norms . . . . .	42
Interpreting Percentiles . . . . .	43
<b>Score Ranges Used for Reports . . . . .</b>	<b>43</b>
<b>The Watson-Glaser II Interview Report . . . . .</b>	<b>43</b>
<b>The Watson-Glaser II Development Report . . . . .</b>	<b>44</b>

**List of Figures**

Figure 7.1 Three Factor Model (Model 2) For Subtests and Testlets (*N* = 306) . . . . . 21

Figure 7.2 Five Factor Model (Model 3) For Subtests and Testlets (*N* = 306) . . . . . 22

**List of Tables**

Table 3.1 Demographic Data for Tryout Sample . . . . . 7

Table 3.2 Number of Cases Gathered for New Watson-Glaser II Form E Items . . . . . 8

Table 4.1 Descriptive Statistics and Correlations for Watson-Glaser II Form D and Watson-Glaser Short Form Scores . . . . . 9

Table 4.2 Total Raw Score Equivalencies for Forms D, Short, and A/B . . . . . 10

Table 4.3 Total Raw Score Equivalencies for Forms E, Short, and A/B . . . . . 11

Table 4.4 Total Raw Score Equivalencies for Forms D and E . . . . . 12

Table 4.5 Equivalency of Paper and Online Modes of Administration . . . . . 13

Table 5.1 Selected List of Watson-Glaser II Normative Samples . . . . . 15

Table 6.1 Test-Retest Reliability of the Watson-Glaser Short Form . . . . . 17

Table 6.2 Watson-Glaser II Internal Consistency Reliability Coefficients (*r*) and Standard Errors of Measurement (*SEM*) . . . . . 17

Table 6.3 Demographic Characteristics of the Sample Used to Calculate Form D Internal Consistency Coefficients . . . . . 18

Table 7.1 Confirmatory Factor Analyses of Watson-Glaser II Form D . . . . . 20

Table 7.2 Intercorrelations Among Watson-Glaser II Form D Subscale Scores . . . . . 23

Table 7.3 Watson-Glaser Convergent Validity Evidence . . . . . 24

Table 7.4 Descriptive Statistics and Correlations for the Watson-Glaser II Form D Raw and WAIS-IV Scaled and Composite Scores . . . . . 26

Table 7.5 Demographic Data for Convergent and Criterion-Related Validity Studies . . . . . 29

Table 7.6 Previous Studies Showing Evidence of Criterion-Related Validity . . . . . 31

Table 7.7 Descriptive Statistics and Correlations for Watson-Glaser II Scores and Performance Ratings . . . . . 32

Table 7.8 Mean Performance of Highly Ranked Critical Thinkers and a Contrast Group . . . . . 33

**References . . . . . 46**

**Glossary . . . . . 48**

The *Watson-Glaser Critical Thinking Appraisal*<sup>®</sup> has a distinguished history, dating back to its initial development in 1925. Designed to measure important abilities and skills involved in critical thinking, it has been used in organizations as a selection and development tool and in academic settings as a measure of gains in critical thinking resulting from specific coursework or instructional programs. A *Mental Measurement Yearbook* review noted that the Watson-Glaser is distinguished by its voluminous research and validity studies (Geisenger, 1998).

The Watson-Glaser<sup>™</sup> II Critical Thinking Appraisal (hereafter referred to as Watson-Glaser II) is the newest revision. This revision was undertaken to incorporate enhancements requested by customers while maintaining the qualities that have made the Watson-Glaser the leading critical thinking appraisal over the last 85 years. Specific enhancements include:

- More contemporary and business relevant items
- Better face validity and applicability of items for individuals from countries other than the United States
- Inclusion of a higher proportion of difficult items to better separate individuals along a continuum of critical thinking
- Development of two 40-item forms that can be administered in approximately the same time as the previous Short Form, while discriminating among candidates as effectively as the previous 80-item forms
- New reports, including a basic Profile Report, Interview Report, and Development Report
- Interpretable subscale scores that provide information about three critical thinking skill domains; the ability to Recognize Assumptions, Evaluate Arguments, and Draw Conclusions

This manual describes the steps taken to update the test and create the new reports. The manual is divided into two broad areas. The first section addresses:

- Construct conceptualization and content development. Chapter 2 describes the underlying conceptualization of the Watson-Glaser II and introduces the new Forms D and E. Chapter 3 describes the criteria used to select items and stages of test development, including information on data collection procedures, sample characteristics, item writing and item analysis.
- Linking Watson-Glaser II to prior Watson-Glaser Forms. To ensure that scores on all Watson-Glaser forms can be directly compared, studies were conducted to link scores for the new Forms D and E to the existing Forms Short, A, and B. Chapter 4 describes and presents results for these procedures, and chapter 5 describes the creation of norms.
- Reliability and validity. Evidence of reliability is presented in chapter 6 and evidence of validity is presented in chapter 7.

The second section focuses on application topics:

- Test Administration. Chapter 8 presents guidelines for computer-based and paper-and-pencil administration.
- Pre-Employment Selection. Use and interpretation of the Watson-Glaser II for pre-employment selection, with references to specific legal standards and best practices, is presented in chapter 9.
- Reports. Chapter 10 describes the content and interpretation of the Watson-Glaser II Profile, Interview, and Development reports.

# Critical Thinking 2

Watson and Glaser (Glaser, 1937; Watson & Glaser, 1994) believed that critical thinking includes:

- attitudes of inquiry that involve an ability to recognize the existence of problems and an acceptance of the general need for evidence in support of what is asserted to be true,
- knowledge of the nature of valid inferences, abstractions, and generalizations in which the weight or accuracy of different kinds of evidence are logically determined, and
- skills in employing and applying the above attitudes and knowledge.

Consistent with this conceptualization, the Watson-Glaser II has maintained the same approach to measuring critical thinking. Each Watson-Glaser II subtest is composed of reading passages or scenarios that include problems, statements, arguments, and interpretations of data similar to those encountered on a daily basis at work, in the classroom, and in newspaper or magazine articles. Each scenario is accompanied by a number of items to which the participant responds.

There are two types of scenario/item content: *neutral* and *controversial*. Neutral scenarios and items deal with subject matter that does not cause strong feelings or prejudices, such as the weather, scientific facts, or common business situations. Scenarios and items having controversial content refer to political, economic, and social issues that frequently provoke emotional responses.

As noted in the critical thinking research literature, strong attitudes, opinions, and biases affect the ability of some people to think critically (Klaczynski, Gordon, & Fauth, 1997; Nickerson, 1998; Sa, West, & Stanovich, 1999; Stanovich & West, 1997, 2008; West, Tolplak, & Stanovich, 2008). Though controversial scenarios are included throughout the Watson-Glaser II, the majority are included in the Evaluate Arguments subtest. Evaluate Arguments scores are, therefore, expected to reflect people's ability to think critically about controversial issues.

## RED Model

The Watson-Glaser II introduces one notable change to Watson and Glaser's original work. Factor analyses of the existing instrument (Forms Short, A, B) consistently revealed a structure in which three scales, Inference, Deduction and Interpretation—all related to drawing conclusions—factored together. Recognition of Assumptions and Evaluation of Arguments remained as independent factors. Based on this finding and the logical appeal and interpretational ease of the three factor model, a new subscale composition was proposed.



## Recognize Assumptions

Assumptions are statements that are assumed to be true in the absence of proof. Identifying assumptions helps in discovery of information gaps and enriches views of issues. Assumptions can be unstated or directly stated. The ability to recognize assumptions in presentations, strategies, plans, and ideas is a key element in critical thinking. Being aware of assumptions and directly assessing their appropriateness to the situation helps individuals evaluate the merits of a proposal, policy, or practice.

## Evaluate Arguments

Arguments are assertions that are intended to persuade someone to believe or act a certain way. Evaluating arguments is the ability to analyze such assertions objectively and accurately. Analyzing arguments helps in determining whether to believe them or act accordingly. It includes the ability to overcome a confirmation bias—the tendency to look for and agree with information that confirms prior beliefs. Emotion plays a key role in evaluating arguments as well. A high level of emotion can cloud objectivity and the ability to accurately evaluate arguments.

## Draw Conclusions

Drawing conclusions consists of arriving at conclusions that logically follow from the available evidence. It includes evaluating all relevant information before drawing a conclusion, judging the plausibility of different conclusions, selecting the most appropriate conclusion, and avoiding overgeneralization beyond the evidence.

To better measure these factors in the Watson-Glaser II, more items were added to the Recognize Assumptions and Evaluate Arguments scales than had been in the Watson-Glaser Short Form. In the Short Form, Recognize Assumptions included 8 items and Evaluate Arguments included 9 items. The new equivalent forms, D and E, include 12 Recognize Assumptions items, 12 Evaluate Arguments items, and 16 Draw Conclusions items.

## Development of the Watson-Glaser II 3

The Watson-Glaser II Form D is a revision of Short Form/Form A and Form E is a revision of Form B. Historical and test development information for the Short Form is available in the Watson-Glaser, Short Form Manual, 2006 edition and historical and test development information for Forms A and B is available in the Watson-Glaser, Forms A and B Manual, 1980 edition.

### Criteria for Item Selection

The following criteria were used to guide the selection of Watson-Glaser II items:

- Improve subscale structure and interpretability
- Improve total score distribution and maintenance of total score reliability
- Improve scenarios and items: business relevance, global applicability, currency of controversial issues
- Maintain 9th grade reading level
- Maintain Short Form administration time

### Improve Subscale Structure and Interpretability

Watson-Glaser II development began with investigation of the factor structure of Forms A, B, and Short. A series of exploratory factor analyses were conducted using Form A ( $n = 2,844$ ), B ( $n = 2,706$ ), and Short ( $n = 8,508$ ) testlet scores. (A testlet is 1 scenario and a set of 2 to 6 questions.) Testlet scores were generated by summing the number of correct responses for items associated with each scenario. Based upon these analyses and a confirmatory test of Form D (see chapter 7), it was determined that Watson-Glaser II scores could best be represented by a three-subscale structure: Recognize Assumptions, Evaluate Arguments, and Draw Conclusions.

Interpretability was improved by organizing Watson-Glaser II subscale scores according to the empirically verified three-subscale structure and adjusting the number of items in each subscale to improve reliability. Specifically, each subscale is composed of a minimum of 12 items (Recognize Assumptions and Evaluate Arguments) and a maximum of 16 items (Draw Conclusions). Interpretability was improved by conducting validation analyses to better understand similarities and differences in the meaning of subscale scores (see chapter 7).

### Improve Score Distribution and Maintain Total Score Reliability

Improving score distribution through item difficulty was of central importance throughout the item selection process. Previous versions of the Watson-Glaser contained a majority of items that had high passing rates, resulting in negatively skewed distributions. To better normalize total score distributions, and improve discrimination along the continuum of critical thinking, items with lower than average passing rates and high discrimination were included in the Watson-Glaser II. Items with high discrimination at different points across the full range of ability for the target population also were chosen. Both Classical Test Theory (CTT) statistics ( $p$ -values, item-total and item-subscale correlations) and Item Response Theory (IRT) statistics ( $a$  and  $b$  parameter estimates) were considered in selecting items based on difficulty and discrimination.

## Improve Scenarios and Items: Business Relevance, Global Applicability, Currency of Controversial Issues

### Business Relevance

To meet the goal of improving the business relevance of scenarios and items, the proportion of business relevant items was increased from 8%, 4% and 23% on Forms Short, A and B, respectively, to 45% and 58% on Watson-Glaser II Forms D and E, respectively. Business relevant scenarios and items were those that involved common workplace events (e.g., training), organizational challenges (e.g., increasing productivity), government practices toward industry (e.g., trade restrictions), and employee/consumer rights issues (e.g., workplace safety).

### Global Applicability

Global applicability was increased by having a cross-cultural team representing eight countries review all existing and experimental scenarios for relevance and appropriateness in their countries. Based on the results of this process, 100% of the items on Forms D and E are relevant or could be easily adapted for use in Australia, Canada, Mexico, the Netherlands, the United Kingdom, and the United States. More than 85% of the items on both forms were appropriate for use in China and France.

### Currency of Controversial Scenarios and Items

Inclusion of controversial scenarios allows for assessment of an individual's critical thinking effectiveness when dealing with emotionally laden versus neutral subject topics. To ensure that a proportion of Watson-Glaser II scenarios and items reflect current controversial issues, a panel of six test developers with degrees in applied psychology independently rated the 49 testlet scenarios from Forms, A, B, and Short. A subset of four of these raters rated an additional 25 experimental testlet scenarios. Raters were instructed to rate scenarios on a seven-point scale from 1 = "Neutral/Not Controversial" to 4 = "Moderately Controversial" to 7 = "Highly Controversial" based on how a typical Watson-Glaser examinee would be likely to interpret the scenario. A scenario was considered controversial if a majority of raters (e.g., 4 out of 6) rated the testlet as moderately controversial or higher. Based on this analysis, Form D had 6 testlets and Form E had 8 testlets that could be considered controversial by today's standards. For comparison, the Short Form had 5 testlets, and Forms A and B each had 8 testlets that could be considered controversial by today's standards.

### Maintain Reading Level

Like previous Watson-Glaser forms, Watson-Glaser II instructions, scenarios, and items were written at or below the ninth-grade reading level. Reading level was assessed using *EDL Core Vocabulary in Reading, Mathematics, Science, and Social Studies* (Taylor, et al., 1989).

### Maintain Short Form Test Administration Time

To maintain the administration time established for the 40-item Short Form, both Forms D and E were developed as 40-item forms. In the standardization sample ( $n = 636$ ), completion times for the Watson-Glaser Short Form (median = 21.92 minutes) and Watson-Glaser II Form D (median = 22.48 minutes) were similar, supporting relatively equal administration times.

## Test Development Stages

Test development occurred in six stages: conceptual development, item writing and review, pilot, calibration, tryout, and standardization. The following sections provide a brief description of each of these stages, including the purpose and relevant methodology.

## Conceptual Development Stage

During conceptual development, the revision goals were identified and research plans established. Consultation with Watson-Glaser customers including internal human resources professionals from large- and mid-size organizations, human resources consultants, and educational instructors provided an initial set of revision goals. These goals were supplemented with information from extensive literature reviews and input sessions with Pearson customer service and sales representatives.

## Item Writing and Review Stage

The purpose of item writing was to generate enough items to create two 40-item short forms, Forms D and E, that each contained approximately 35% new items and 65% existing items. Toward this end, 200 new experimental items (approximately 40 testlets, each testlet = 1 scenario followed by 5 items) were drafted.

Item writing was conducted by individuals with extensive prior experience writing critical thinking/reasoning items. Detailed guidelines for writing items were provided to each writer, and each writer had to submit items for review prior to receiving approval to write additional items. Writers were instructed to write items at a 9th grade reading level.

Subject matter experts with experience writing and reviewing general mental ability/reasoning items reviewed and provided feedback on how well each experimental item measured the target construct, clarity and conciseness of wording, and difficulty level.

In addition, a separate group of subject matter experts reviewed and provided feedback on how well experimental items and items from existing forms could be expected to transport or be adapted to other countries/cultures. These subject matter experts included one Pearson employee born in the U.S., as well as twelve Pearson employees born and raised in countries other than the U.S. All subject matter experts were familiar with the Watson-Glaser and test development principles. Countries represented by the panel included Australia, Canada, China, France, Mexico, the Netherlands, the United Kingdom, and the United States.

As a final step, experimental scenarios and items intended to be business relevant were reviewed for use of appropriate business language and situations by Pearson's U.S. Director of Talent Assessment.

## Pilot Stage

The primary goal of the pilot stage was to do a preliminary check on the psychometric properties of the experimental items, and determine which experimental items merited further data collection. Sets of 5 to 15 items were administered as experimental items to people who completed Forms A, B, and Short. These individuals were part of the customer data base, which is comprised primarily of business professionals (approximately 90%) with a smaller proportion of college students (approximately 10%). After a minimum of 100 cases, and typically more (average  $n = 203$ ), were collected on an experimental item set, a new experimental item set was inserted into the test and the previous set was rotated out. Classical Test Theory (CTT) item analysis procedures were used to evaluate each experimental item.

## Calibration Stage

To achieve stable CTT and Item Response Theory (IRT) parameter estimates, experimental items that had adequate psychometric properties based on pilot stage data were administered to a larger sample until a cumulative total of at least 400 participants had taken each item. Item adequacy was evaluated using IRT and CTT information (e.g.,  $a$  and  $b$  parameter, item difficulty, item to scale correlation).

IRT item analysis procedures were also used to calibrate experimental items and put them on a common scale with Forms A, B, and Short items. Chapter 4 describes the methodology used to calibrate experimental items.

## Tryout Stage: Form D

The primary goal of the tryout stage was to evaluate a preliminary version of the Watson-Glaser II, Form D, including factor structure and subscale reliabilities. Results from the factor analyses, which are presented in detail in Chapter 7, supported the three factor model—Recognize Assumptions, Evaluate Arguments, and Draw Conclusions. Tryout data were obtained from a sample of 306 examinees who had at least a Bachelor's degree. College degree was used as a proxy to represent the ability level expected for the Watson-Glaser II target population. Table 3.1 provides demographic data for this sample. Results from the tryout analyses on Form D confirmed the item quality and subscale structure, so Form D item content was finalized.

**Table 3.1** Demographic Data for Tryout Sample

	<b>N</b>	<b>Percent</b>
	306	100.00
<b>EDUCATION LEVEL</b>		
Bachelor's	181	59.15
Master's	108	35.29
Doctorate	17	5.56
<b>SEX</b>		
Female	98	32.03
Male	208	67.97
<b>ETHNICITY</b>		
White non-Hispanic	231	75.49
Hispanic, Latino/a	43	14.05
Black, African American	10	3.27
Asian/Pacific Islander	9	2.94
Native American	1	0.33
Multiracial	6	1.96
No Response	6	1.96
<b>AGE</b>		
20–29	56	18.30
30–39	89	29.08
40–49	55	17.97
50–59	63	20.59
60+	41	13.40
No Response	2	0.65
<b>POSITION</b>		
Professional/Individual Contributor	179	58.50
Manager	28	9.15
Hourly/Entry-Level	28	9.15
Executive/Director	19	6.21
Skilled Trades & General Labor	5	1.63
Supervisor	3	0.98
Other	44	14.38
<b>INDUSTRY</b>		
Education	134	43.79
Health Care	26	8.50
Publishing, Printing	25	8.17
Manufacturing & Production	19	6.21
Information Technology, High-Tech, Telecommunications	18	5.88
Government, Public Service, Defense	19	6.21
Professional, Business Services	12	3.92
Financial Services, Banking, Insurance	9	2.94
Other	44	14.38

## Standardization Stage: Form D

The standardization stage focused on the creation of linking tables for the Watson-Glaser II Form D with Forms A, B, and Short, and additional reliability and validity analyses. Validity data were obtained across six sampling sites. To examine convergent and discriminate validity, examinees were administered Form D and one more of the following instruments: *Wechsler Adult Intelligence Scale-IV*, *Myers-Briggs Type Indicator*, *Golden Personality Type Profiler*, and the *Workplace Big Five Profile*. To examine criterion-related validity, Form D was administered to job incumbents in two different organizations in which job performance data was collected.

## Standardization Stage: Form E

After Form D was finalized, items were selected for Form E using item selection criteria, information from the pilot and calibration stages, and item difficulty levels corresponding with Form D items. Table 3.2 shows the number of cases used to evaluate the psychometric properties of each new Form E item.

**Table 3.2** Number of Cases Gathered for New Watson-Glaser II Form E Items

Item Number	<i>N</i>
3	945
4	945
5	945
14	855
15	855
18	2659
19	2659
23	695
24	694
31	689
32	689
33	678
34	708

### Equivalence of Watson-Glaser II Forms D and E with Previous Forms

Numerous steps were taken to ensure that the new Watson-Glaser II forms measure the same constructs in the same way as the previous forms. First, the psychometrically strongest items from previous forms were used to form the majority of items in the new forms. Second, the item format used for Watson-Glaser II Forms D and E was identical to the format used for previous Forms A, B, and Short. Third, item writers were instructed to write items aimed at measuring the same constructs tapped by previous forms. Fourth, new items were selected based in part on their correlation with subscales from previous forms. Finally, following the assembly of Watson-Glaser II Form D, correlation coefficients were computed between total raw scores on Form D and the Short Form.

During the standardization stage, Watson-Glaser II Form D and Watson-Glaser Short Form items were administered to 636 examinees, as part of a single 68-item form. The results, which are presented in Table 4.1, reveal a correlation of .85 between the Watson-Glaser II and Short Form total scores. To estimate subscale correlations, three scales on the Short Form (Deduction, Inference, and Interpretation) were combined to form a Draw Conclusions subscale. The correlations between the subscales of the two instruments were .88 for Recognize Assumptions, .82 for Draw Conclusions, and .38 for Evaluate Arguments. Because Evaluate Arguments was the psychometrically weakest subscale in previous forms of the Watson-Glaser, the low correlation was not surprising.

**Table 4.1** Descriptive Statistics and Correlations for Watson-Glaser II Form D and Watson-Glaser Short Form Scores

Short Form Scores	Watson-Glaser II Form D Scores				Watson-Glaser Short Form		
	Total Score	Recognize Assumptions	Evaluate Arguments	Draw Conclusions	Mean	SD	N
Total Score	.85	.68	.43	.80	29.2	5.7	636
Recognizing Assumptions	.74	.88	.26	.48	5.8	2.2	636
Evaluating Arguments	.41	.24	.38	.36	6.9	1.3	636
Draw Conclusions	.75	.50	.37	.82	17.9	4.0	636
	Watson-Glaser-II Form D						
Mean	27.1	8.2	7.7	11.2			
SD	6.5	3.1	2.2	3.1			
N	636	636	636	636			

### Equivalent Raw Scores

To establish equivalent raw scores across forms, raw-score-to-ability estimates were generated for all forms using Rasch-model difficulty parameters. Raw scores corresponding to the same ability estimate were considered equivalent (i.e., represent the same ability level).

Table 4.2 presents raw score equivalents for Forms D to Short and A/B at the total score level. To convert a Form D raw score to Short Form raw score, find that score in the Form D column in Table 4.2, then look to the right at the Short Form raw score column. For example, a score of 28 on Form D is equivalent to a Short Form score of 30. Table 4.3 presents the raw score equivalents for Form E to Short and A/B and Table 4.4 presents the raw score equivalents for Form D to Form E.

**Table 4.2 Total Raw Score Equivalencies for Form D, Short and A/B**

	<b>Raw Score</b>		
	<b>Form D</b>	<b>Short Form</b>	<b>Forms A/B</b>
40	40	79	
39	40	78	
38	38	76	
37	38	75	
36	37	73	
35	36	71	
34	35	69	
33	34	68	
32	34	66	
31	33	64	
30	32	62	
29	31	61	
28	30	59	
27	29	57	
26	28	55	
25	27	54	
24	26	52	
23	25	50	
22	25	48	
21	24	47	
20	23	45	
19	22	43	
18	21	41	
17	20	39	
16	19	37	
15	18	35	
14	17	33	
13	16	32	
12	15	30	
11	14	28	
10	13	25	
9	12	23	
8	11	21	
7	9	19	
6	8	17	
5	7	14	
4	6	12	
3	5	9	
2	3	7	
1	2	4	

Table 4.3 Total Raw Score Equivalencies for Forms E, Short and A/B

Form D	Raw Score	
	Short Form	Forms A/B
40	40	79
39	40	78
38	39	77
37	38	75
36	37	74
35	37	72
34	36	71
33	35	69
32	34	67
31	34	66
30	33	64
29	32	62
28	31	61
27	30	59
26	29	57
25	28	55
24	27	53
23	26	52
22	25	50
21	24	48
20	23	46
19	22	44
18	21	42
17	20	40
16	19	38
15	18	36
14	17	34
13	16	32
12	15	29
11	14	27
10	12	25
9	11	23
8	10	21
7	9	18
6	8	16
5	7	14
4	5	11
3	4	9
2	3	6
1	2	3

**Table 4.4 Total Raw Score Equivalencies For Forms D and E**

<b>Raw Score</b>	
<b>Form D</b>	<b>Form E</b>
40	40
39	39
38	38
37	37
36	35, 36
35	34
34	33
33	32
32	31
31	30
30	29
29	28
28	27
27	26
26	25
25	24
24	23
23	22
22	21
20, 21	20
19	19
18	18
17	17
16	16
15	15
14	14
13	13
12	12
11	11
10	10
9	9
8	8
7	7
6	6
5	5
4	4
3	3
2	2
1	1

## Equivalence of Computer-Based and Paper-and-Pencil Forms

Occasionally, customers inquire about the equivalence of on-line versus paper administration of the Watson-Glaser. Studies of the effect of test administration mode have generally supported the equivalence of paper and computerized versions of non-speeded cognitive ability tests (Mead & Drasgow, 1993). To ensure that these findings held true for the Watson-Glaser, in 2005, Pearson conducted an equivalency study using paper-and-pencil and computer-administered versions of the Short Form (Watson & Glaser, 2006). This study is presented in this manual for the reader's convenience. Given these results no equivalency study was conducted for the Watson-Glaser II.

In this study, a counter-balanced design was employed using a sample of 226 adult participants from a variety of occupations. Approximately half of the group ( $n = 118$ ) completed the paper form followed by the online version, while other participants ( $n = 108$ ) completed the tests in the reverse order. Table 4.3 presents means, standard deviations, and correlations obtained from an analysis of the resulting data. As indicated in the table, neither mode of administration yielded consistently higher raw scores, and mean score differences between modes were less than one point (0.5 and 0.7). The variability of scores also was very similar, with standard deviations ranging from 5.5 to 5.7.

The correlation coefficients indicate that paper-and-pencil raw scores correlate very highly with online administration raw scores (.86 and .88, respectively). Notably, the correlations across administration modes were similar to those found in test-retest studies that used the same administration mode across testings (.81 for paper and .89 for online; see chapter 6). The high correlations provide further support that the two modes of administration can be considered equivalent. Thus, raw scores on one form (paper or online) may be interpreted as having the same meaning as identical raw scores on the other form.

**Table 4.5**      **Equivalency of Paper and Online Modes of Administration**

Administration Order	<i>N</i>	Paper		Online		<i>r</i>
		Mean	<i>SD</i>	Mean	<i>SD</i>	
Paper followed by Online	118	30.1	5.7	30.6	5.5	.86
Online Followed by Paper	108	29.5	5.5	28.8	5.7	.88
Total	226	29.8	5.6	29.7	5.6	.87

The raw score on the Watson-Glaser II (Forms D and E) is calculated by adding the total number of correct responses. The maximum raw score is 40. Raw scores can be used to rank examinees in order of performance, but little can be inferred from raw scores alone. It is important to relate the scores to specifically defined normative groups to make the test results meaningful.

## Background on Norms

Norms provide a basis for evaluating an individual's score relative to the scores of other individuals who took the same test. Norms allow for the conversion of raw scores to more useful comparative scores, such as percentile ranks. Typically, norms are constructed from the scores of a large sample of individuals who took a test. This group of individuals is referred to as the normative group.

The characteristics of the sample used for preparing norms are critical in determining the usefulness of those norms. For some purposes, such as intelligence testing, norms that are representative of the general population are essential. For other purposes, such as pre employment selection, information derived from a relevant and well-defined group is most useful (e.g., occupation-specific group). Typically, occupational norms provided by the publisher are applicable. However, a variety of situational factors, including job demands and local labor market conditions impact an organization. Therefore, organizations need to consider their own context before deciding to implement commercially published norms.

The ideal norm group is one that is representative of those who will be taking the test in the local situation. It is best, whenever possible, to prepare local norms by accumulating the test scores of applicants, trainees, employees, or students. One of the factors that must be considered in preparing norms is sample size. Data from smaller samples tend to be unstable and the use of standard scores like percentile ranks presents an unwarranted impression of precision. To avoid unstable results, it may be preferable to use Pearson norms until a sufficient and representative number of local cases has been collected (preferably 100 or more) to create a local norm.

## Pearson Norms

The type of norms available and their composition characteristics are updated frequently, so it is best to contact an Account Manager (1.888.298.6227) or access [Talents.com](https://www.talents.com) for the most current offerings.

The Watson-Glaser II norms were derived from the existing Watson-Glaser norms through an extrapolation process (described in Chapter 4). The raw scores on the Watson-Glaser II and Watson-Glaser Form A were converted to ability estimates using Rasch-model difficulty parameters. These ability estimates were then converted to a common scale (i.e., scaled scores), facilitating the comparison of scores across forms. This link across forms was used to allow the normative samples collected for Form A to be converted for use with the Watson-Glaser II.

Fourteen occupational or level groups created for Watson-Glaser Form A were selected as the normative samples. These groups, which contained relatively large numbers (average  $n = 967$ ), are presented in Table 5.1.

**Table 5.1 Selected List of Watson-Glaser II Normative Samples**

<b>Group</b>	<b><i>N</i></b>
<b>Occupation</b>	
Accountant	368
Consultant	473
Engineer	677
Human Resource Professional	562
Information Technology Professional	702
Sales Representative	507
<b>Position Type/Level</b>	
Executive	1,389
Director	1,468
Manager	3,243
Professional/Individual Contributor	2,234
Supervisor	922
Hourly/Entry-Level	306
<b>Norms by Occupation Within Specific Industry</b>	
Engineer in Manufacturing/Production	140
Manager in Manufacturing/Production	584

# Evidence of Reliability 6

Reliability refers to a test's stability or internal consistency. In this chapter, the concepts of reliability and standard error of measurement are defined. Then evidence of test-retest reliability is introduced, followed by a description of the analyses used to derive internal consistency estimates for Watson-Glaser II Forms D and E.

## Definitions of Reliability and Standard Error of Measurement

The reliability of a test is expressed as a correlation coefficient that represents the consistency of scores that would be obtained if a test could be given an infinite number of times. In actual practice, however, we do not have the luxury of administering a test an infinite number of times, so we can expect some measurement error. Reliability coefficients help us to estimate the amount of error associated with test scores. Reliability coefficients can range from .00 to 1.00. The closer the reliability coefficient is to 1.00, the more reliable the test. The U.S. Department of Labor (1999) provides the following general guidelines for interpreting a reliability coefficient: above .89 is considered "excellent," .80–.89 is "good," .70–.79 is considered "adequate," and below .70 "may have limited applicability."

The methods most commonly used to estimate test reliability are test-retest (the stability of test scores over time) and internal consistency of the test items (e.g., Cronbach's alpha coefficient and split-half). Occasionally, alternate forms analysis (the consistency of scores across alternate forms of a test) is used.

Since repeated testing always results in some variation, no single test event ever measures an examinee's actual ability with complete accuracy. We therefore need an estimate of the possible amount of error present in a test score, or the amount that scores would probably vary if an examinee were tested repeatedly with the same test. This error is known as the standard error of measurement (*SEM*). The *SEM* decreases as the reliability of a test increases; a large *SEM* denotes less reliable measurement and less reliable scores.

The *SEM* is a quantity that is added to and subtracted from an examinee's test score to create a confidence interval or band of scores around the obtained score. The confidence interval is a score range that, in all likelihood, includes the examinee's hypothetical "true" score which represents the examinee's actual ability. Since the true score is a hypothetical value that can never be obtained because testing always involves some measurement error, any obtained score is considered only an estimate of the examinee's "true" score. Approximately 68% of the time, the observed score will lie within +1.0 and –1.0 *SEM* of the true score; 95% of the time, the observed score will lie within +1.96 and –1.96 *SEM* of the true score; and 99% of the time, the observed score will lie within +2.58 and –2.58 *SEM* of the true score.

## Test-Retest Reliability

Cognitive ability is a stable trait (Deary, Whalley, Lemmon, Crawford, & Starr, 2000), and prior versions of the Watson-Glaser Short Form have demonstrated an acceptably high level of test-retest reliability. In light of this evidence, we did not undertake a test-retest reliability study for Forms D and E. Instead, we refer to previous research.

In 1994, a study investigating the test-retest reliability of the Watson-Glaser Short Form was conducted using a sample of 42 adults who completed the Short Form two weeks apart. The test-retest correlation was .81 ( $p < .001$ ) and the mean score was 30.5 ( $SD = 5.6$ ) at the first testing and 31.4 ( $SD = 5.9$ ) at the second testing. The difference in mean scores between the first testing and the second testing was statistically small ( $d = 0.16$ ).

In 2006, test-retest reliability was evaluated using a sample of 57 job incumbents drawn from various organizational levels and industries. The test-retest intervals ranged from 4 to 26 days, with a mean interval of 11 days. As shown in Table 6.1, the Watson-Glaser Short Form total score demonstrated acceptable test-retest reliability ( $r_{12} = .89$ ). The difference in mean scores between the first testing and the second testing was statistically small ( $d = 0.17$ ).

**Table 6.1** Test-Retest Reliability of the Watson-Glaser Short Form

Test-Retest Study	First Testing		Second Testing		$r_{12}$	Cohen's $d$	$N$
	Mean	$SD$	Mean	$SD$			
1994	30.5	5.6	31.4	5.9	.81	0.16	42
2006	29.5	7.0	30.7	7.0	.89	0.17	57

## Internal Consistency Reliability

Cronbach's alpha and the standard error of measurement ( $SEM$ ) were calculated for Watson-Glaser II Form D total and subscale scores using Classical Test Theory. Because Form E was developed using a common-item approach (i.e., no single examinee had data on all 40 items), traditional methods of estimating internal consistency were not applicable. The split-half reliability estimation for Form E was carried out using a method based on Item Response Theory (IRT) since IRT has more flexibility to deal with missing data. The reliability was calculated based on the ability estimates calibrated for the odd and even half of the test using the 27 items for which all examinees had complete data (i.e., items retained from Form B). The calculations were completed using a sample drawn from the customer data base ( $N = 2,706$ ). A correction was then applied to estimate the reliability of the 40-item form using the Spearman-Brown prophecy formula.

Results are presented in Table 6.2 and descriptions of the sample used to estimate reliability for Form D are presented in Table 6.3. Internal consistency reliabilities for the total scores were .83 and .81 for Forms D and E, respectively. Consistent with research on previous Watson-Glaser forms, these values indicate that Forms D and E total scores possess adequate reliability.

Internal consistency reliabilities for the Form D subscales Recognize Assumptions (.80) and Draw Conclusions (.70) were both adequate. Internal consistency reliability for the Form D Evaluate Arguments subscale was .57, which is low. It is possible that this subscale is measuring a multidimensional construct (see chapter 2). Overall, subscale scores showed lower estimates of internal consistency reliability as compared to the total score, suggesting that the subscale scores alone should not be used when making selection decisions.

**Table 6.2** Watson-Glaser II Internal Consistency Reliability Coefficients ( $r$ ) and Standard Errors of Measurement ( $SEM$ )

Form D	$N$	$r_{\text{alpha}}$	$SEM$
Total Score	1011	0.83	2.63
Recognize Assumptions	1011	0.80	1.33
Evaluate Arguments	1011	0.57	1.45
Draw Conclusions	1011	0.70	1.68
Form E	$N$	$r_{\text{split}}$	$SEM$
Total Score	2706	0.81	2.78*

Note. \*  $SEM$  is estimated, based on the variance of Form D score.

**Table 6.3 Demographic Characteristics of the Sample used to Calculate Form D Internal Consistency Coefficients**

	<b>N</b>	<b>Percent</b>
	1011	100.00
<b>Education Level</b>		
HS/GED	32	3.17
1–2 yrs college	92	9.10
Associate's	40	3.96
3–4 yrs college	55	5.44
Bachelor's	463	45.80
Master's	210	20.77
Doctorate	31	3.07
No Response	88	8.70
<b>Sex</b>		
Female	493	48.76
Male	436	43.13
No Response	82	8.11
<b>Ethnicity</b>		
White non-Hispanic	771	76.26
Hispanic, Latino/a	76	7.52
Black, African American	26	2.57
Asian/Pacific Islander	23	2.27
Multiracial	14	1.38
Native American	2	0.20
Other	5	0.49
No Response	94	9.30
<b>Age</b>		
16–24	128	12.66
25–34	223	22.06
35–39	123	12.17
40–49	214	21.17
50–59	174	17.21
60+	62	6.13
No Response	87	8.61
<b>Position</b>		
Manager	294	29.08
Professional/Individual Contributor	290	28.68
Supervisor	122	12.07
Hourly/Entry-Level	119	11.77
Executive/Director	70	6.92
Skilled Trades/General Labor	27	2.67
Not Applicable	89	8.80
<b>Industry</b>		
Financial Services, Banking, Insurance	313	30.96
Education	175	17.31
Health Care	124	12.27
Retail & Wholesale	59	5.84
Manufacturing & Production	57	5.64
Publishing, Printing	29	2.87
Hospitality, Tourism	28	2.77
Professional, Business Services	27	2.67
Information Technology, High-Tech, Telecommunications	23	2.27
Government, Public Service, Defense	21	2.08
Other	155	15.33

Validity refers to the degree to which specific data, research, or theory support the interpretation of test scores (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). “Validity is high if a test gives the information the decision maker needs” (Cronbach, 1970). To establish the utility of the Watson-Glaser II, components of construct validity, including content validity, internal factor structure, and convergent and discriminate validity are presented.

## Content Validity

Evidence of content validity exists when the content of a test includes a representative sample of tasks, behaviors, knowledge, skills, or abilities of the identified construct. The critical thinking skills measured by the Watson-Glaser were articulated many years ago by Watson and Glaser (Glaser, 1937; Watson & Glaser, 1952), and they still correspond to critical thinking skills articulated in current models of critical thinking (Facione, 1990; Fisher & Spiker, 2004; Halpern, 2003; Paul & Elder, 2002).

In employment settings, the principal concern is with making inferences about how well the test samples a job performance *domain*—a segment or aspect of job performance which has been identified and about which inferences are to be made (Lawshe, 1975). Because most jobs have several performance domains, a standardized test generally applies only to one segment of job performance. Thus, the judgment of whether content-related evidence exists depends upon an evaluation of whether the same capabilities are required in both the job performance domain and the test (Cascio & Aguinis, 2005). In an employment setting, evidence based on test content should be established by demonstrating that the jobs for which the test will be used require the critical thinking abilities and skills measured by the Watson-Glaser II. In classroom and instructional settings the course content and objectives of such instructional programs should correspond to the constructs measured by the Watson-Glaser II.

## Factor Structure

A series of factor analyses were run to evaluate the factor structure and pattern of the Watson-Glaser. As discussed in chapter 2, exploratory factor analyses had revealed three definable factors. A maximum likelihood extraction method with oblique rotation had been used to analyze the Watson-Glaser Short Form ( $N = 8508$ ) Form A ( $N = 2,844$ ), and Form B ( $N = 2,706$ ). Initial exploration resulted in three stable factors and additional factors (four or five) that could not be interpreted. These additional factors included psychometrically weak testlets and were not stable across forms. Follow-up analyses that specified three factors revealed the configuration of Recognize Assumptions, Evaluate Arguments, and Draw Conclusions (i.e., Inference, Deduction, and Interpretation loaded onto one factor). Given this evidence, and logical appeal and interpretational ease, the three factor model was proposed for the Watson-Glaser II.

## Confirmatory Factor Analysis

Confirmatory factory analysis (CFA) can be used to determine how well a specified theoretical model explains observed relationships among variables. Common indices used to evaluate how well a specified model explains observed relationships include the goodness-of-fit index (GFI), adjusted goodness-of-fit index (AGFI), and the root mean squared error of approximation (RMSEA). GFI and AGFI values each range from 0 to 1, with values exceeding .9 indicating a good fit to the data (Kelloway, 1998). RMSEA values closer to 0 indicate better fit, with values below .10 suggesting a good fit to the data, and values below .05 a very good fit to the data (Steiger, 1990). CFA can also be used to evaluate the comparative fit of several models. Smaller values of chi-square relative to the degrees of freedom in the model indicate relative fit.

During the tryout stage, a series of confirmatory models were compared: Model 1 specified critical thinking as a single factor; Model 2 specified the three factor model; and, Model 3 specified the historical five-factor model.

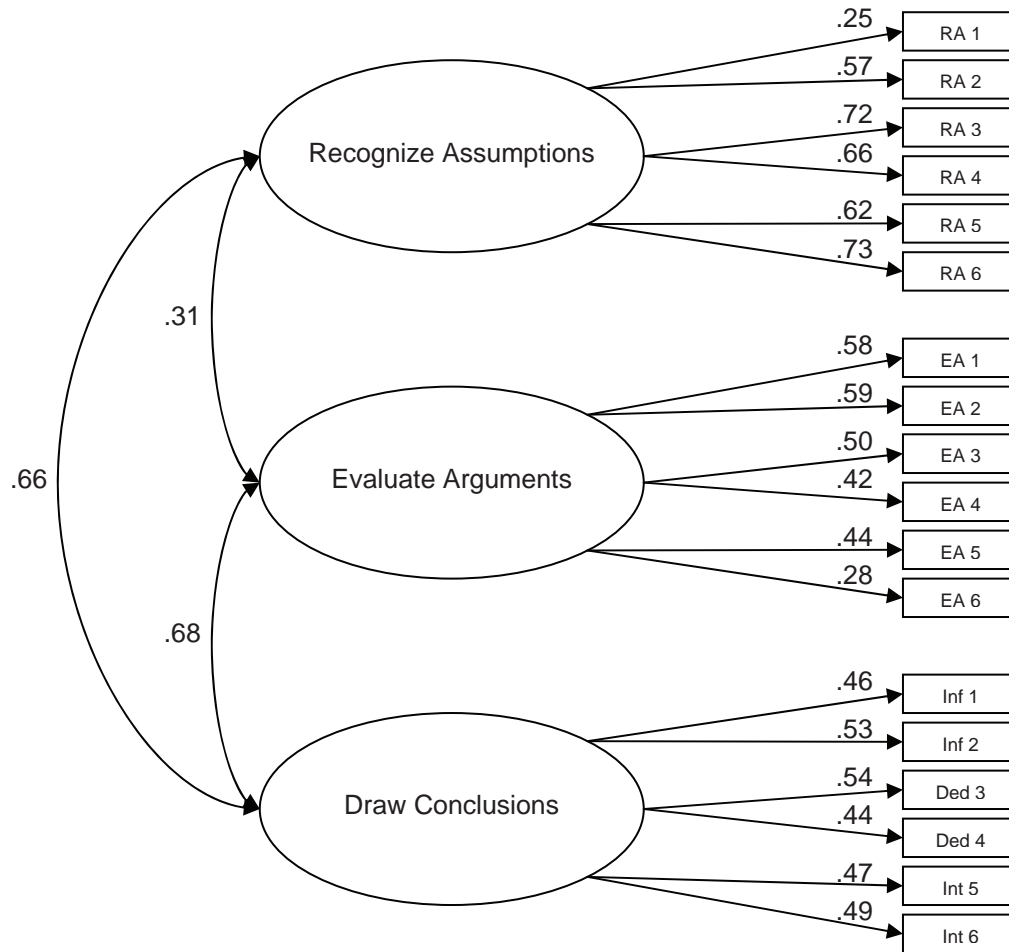
The results, which are presented in Table 7.1 and Figure 7.1, indicated that Model 1 did not fit the data as well as the other two models. Both Model 2 and model 3 fit the data, and there was no substantive difference between the two in terms of model fit. However, the phi coefficients in the five factor model were problematic and suggest that the constructs are not meaningfully separable. For example, the phi co-efficient was 1.18 between Inference and Deduction and .96 between Deduction and Interpretation. Given this evidence, the three factor model was confirmed as the optimal model for the Watson-Glaser II.

During standardization there was an opportunity to replicate the confirmatory factor analyses that were run during the tryout stage. A sample of 636 people participated in the validity studies. The general characteristics of this sample are provided in Table 7.5. Two hundred people did not provide all of the data needed for validation (e.g., job performance ratings), so this subgroup is not described in Table 7.5. The results of the confirmatory factor analysis supported the three factor model (GFI = .97; AGFI = .96; RMSEA = .03), providing further evidence for the three scales of the Watson-Glaser II.

**Table 7.1** Confirmatory Factor Analyses of Watson-Glaser II Form D (*N* = 306)

<b>Model</b>	<b>Chi-square</b>	<b>df</b>	<b>GFI</b>	<b>AGFI</b>	<b>RMSEA</b>
Model 1	367.16	135	0.85	0.81	0.08
Model 2	175.66	132	0.94	0.92	0.03
Model 3	159.39	125	0.95	0.93	0.03

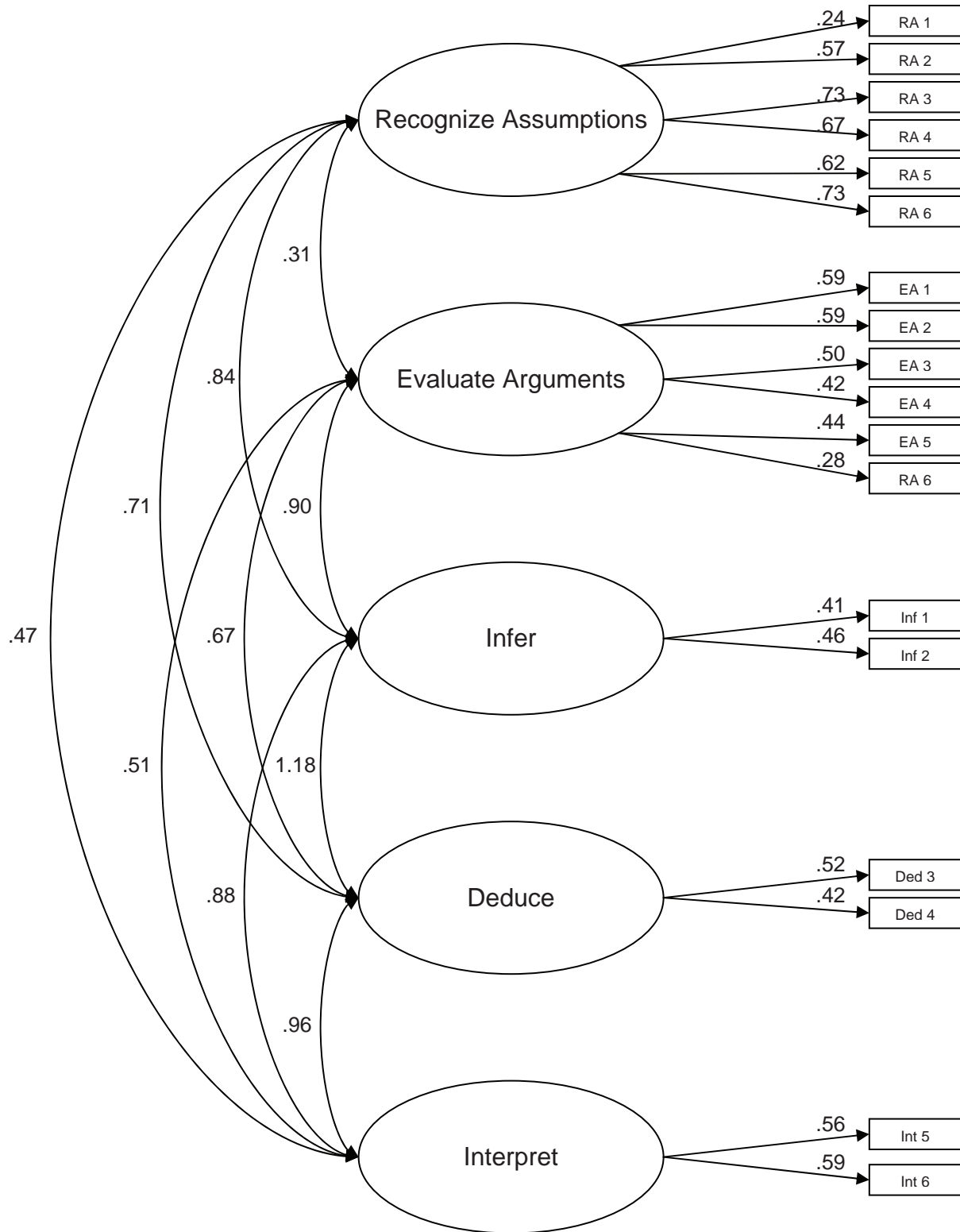
Note. The Chi-Square values are Maximum Likelihood Estimation Chi-Squares from SAS 9.0. See text for explanation of fit indices.



Note: Testlet scores were used as the unit of analysis.

RA = Recognize Assumptions; EA = Evaluate Arguments;  
 Inf = Infer; Ded = Deduce; Int = Interpret

Figure 7.1 Three Factor Model (Model 2) For Subtests and Testlets ( $N = 306$ )



Note: Testlet scores were used as the unit of analysis.

RA = Recognize Assumptions; EA = Evaluate Arguments;  
 Inf = Infer; Ded = Deduce; Int = Interpret

Figure 7.2 Five Factor Model (Model 3) For Subtests and Testlets (N = 306)

## Subscale Intercorrelations

Correlations among the Watson-Glaser II Form D subscales are presented in Table 7.2. The correlations were low to moderate, with Draw Conclusions and Recognize Assumptions correlating highest (.47) and Recognize Assumptions and Evaluate Arguments correlating lowest (.26). These correlations indicate that there is a reasonable level of independence and non-redundancy among the three subscales.

**Table 7.2 Intercorrelations Among Watson-Glaser II Form D Subscale Scores ( $N = 636$ )**

Scale	Mean	SD	1	2	3	4
1. Total	27.1	6.5	1.00			
2. Recognize Assumptions	8.2	3.1	.79	1.00		
3. Evaluate Arguments	7.7	2.2	.66	.26	1.00	
4. Draw Conclusions	11.2	3.1	.84	.47	.41	1.00

## Convergent Validity: Cognitive Ability Measures

Convergent evidence is provided when scores on a test relate to scores on other tests that measure similar traits or constructs. Over the years a number of studies have demonstrated that the Watson-Glaser correlates with other cognitive ability measures, including nonverbal reasoning ability, verbal reasoning, numerical reasoning, achievement (both ACT and SAT), and critical thinking. A summary of these studies is presented in Table 7.3. Correlations with measures of reasoning are particularly strong (e.g., .70 with *Miller Analogies Test for Professional Selection*, .68 with *Advanced Numerical Reasoning Appraisal*, and .53 with *Raven's Advanced Progressive Matrices*).

**Table 7.3 Watson-Glaser Convergent Validity Evidence**

Group	N	Watson-Glaser			Other Test			r
		Form	Mean	SD	Description	Mean	SD	
Job incumbents across occupations (Pearson, 2006)	41	Short	28.0	5.4	Raven's APM	22.3	5.8	.53**
Job incumbents across occupations and industries (Pearson, 2006)	452	Short	31.3	5.8	Advanced Numerical Reasoning Appraisal	20.7	6.3	.68**
Job incumbents across industries (Pearson, 2005)	63	Short	28.7	7.5	Miller Analogies Test for Professional Selection	21.3	11.4	.70**
Job incumbents from multiple occupations in UK (Rust, 2002)	1,546	C <sup>UK</sup>	—	—	Rust Advanced Numerical Reasoning Appraisal			.63**
Education majors (Taube, 1995)	147–194	80-item	54.9	8.1	SAT-Verbal	431.5	75.3	.43**
					SAT-Math	495.5	91.5	.39*
					Ennis-Weir Critical Thinking Essay Test	14.6	6.1	.37*
Baccalaureate Nursing Students (Adams, Stover, & Whitlow, 1999)	203	80-item	54.0	9.3	ACT Composite	21.0	—	.53**
Dispatchers at a Southern railroad company (Watson & Glaser, 1994)	180	Short	24.9	5.0	Industrial Reading Test	29.9	4.4	.53**
					Test of Learning Ability	73.7	11.4	.50**
Lower-level management applicants (Watson & Glaser, 1994)	219	Short	33.5	4.4	Wesman, Verbal	27.5	6.0	.51**
	217				EAS, Verbal Comp.	20.7	3.1	.54**
	217				EAS, Verbal Reasoning	16.7	4.6	.48**
Mid-level management applicants (Watson & Glaser, 1994)	209	Short	34.0	4.2	Wesman, Verbal	27.5	6.0	.66**
	209				EAS, Verbal Comp.	21.0	3.0	.50**
	208				EAS, Verbal Reasoning	16.6	4.9	.51**
Executive management applicants (Watson & Glaser, 1994)	440	Short	33.4	4.2	Wesman, Verbal	27.0	5.8	.54**
	437				EAS, Verbal Comp.	21.1	3.4	.42**
	436				EAS, Verbal Reasoning	16.2	4.2	.47**

\*p < .05. \*\*p < .01.

## Watson-Glaser II and WAIS-IV

The recent release of the WAIS-IV created an opportunity to examine the correlation between WAIS-IV and Watson-Glaser II scores. The Watson-Glaser II was administered to 62 individuals with a Bachelor's degree or higher (a group similar to individuals in the Watson-Glaser II target population) who had recently taken the WAIS-IV (within the prior 11 to 23 months). The sample is described in Table 7.5, which is at the end of this section.

The WAIS-IV consists of 15 subtests that measure cognitive ability across 4 domains: Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed. Due to moderate overlap in the constructs measured, it was expected that Watson-Glaser II total score would correlate in the range of .4 to .6 with WAIS-IV total score. Further it was hypothesized that Watson-Glaser II total score would correlate with the WAIS-IV Working Memory index and the Perceptual Reasoning Index, and to a lesser extent Verbal Comprehension index. Reasoning and working memory are needed to perform critical thinking tasks that involve maintaining information in conscious awareness (e.g., a conclusion, a premise) and mentally manipulating the information to arrive at an answer. The Watson-Glaser II is a verbally loaded test and as such, it should correlate with Verbal Comprehension. Conversely, Processing Speed is an important component of cognitive ability, but it is not viewed as core to critical thinking. Finally, the WAIS-IV has a composite, Fluid Reasoning, which measures the ability to manipulate abstractions, rules, generalizations, and logical relationships. It was hypothesized that this composite would be strongly correlated with the Watson-Glaser II total score because both scales measure reasoning.

Table 7.4 presents the means, standard deviations, and correlation coefficients. The results indicated that Watson-Glaser II total scores were significantly related to WAIS-IV Full Scale IQ score. As predicted, Watson-Glaser II total scores were also significantly related to Working Memory (.44), Perceptual Reasoning (.46), and Verbal Comprehension (.42), but not Processing Speed (.14).

**Table 7.4** Descriptive Statistics and Correlations for the Watson-Glaser II Form D Raw and WAIS-IV Scaled and Composite Scores

WAIS-IV Composite/ Subtest Score	Watson-Glaser II Form D Score				WAIS-IV		
	Total Score	Recognize Assumptions	Evaluate Arguments	Draw Conclusions	Mean	SD	<i>n</i>
Full-Scale IQ	.52	.31	.21	.62	110.8	14.6	56
Perceptual Reasoning Index	.46	.20	.25	.56	107.8	15.8	62
Block Design	.49	.27	.27	.54	11.2	3.3	62
*Figure Weights	.47	.31	.24	.50	12.0	2.8	49
*Matrix Reasoning	.47	.18	.26	.61	12.1	3.2	62
Picture Completion	.05	-.06	.14	.05	10.4	2.7	56
Visual Puzzles	.21	.06	.11	.28	10.9	3.2	62
Working Memory Index	.44	.24	.13	.59	108.0	14.5	62
Arithmetic	.45	.17	.23	.59	11.4	2.8	62
Digit Span	.33	.25	.00	.46	11.5	3.1	62
Letter-Number Sequencing	.37	.09	.21	.53	11.4	3.6	49
Verbal Comprehension Index	.42	.34	.10	.46	111.8	13.6	56
Comprehension	.23	.19	.04	.25	11.7	2.9	56
Information	.40	.34	.06	.45	12.4	2.7	56
*Similarities	.38	.29	.17	.38	11.9	2.8	62
Vocabulary	.31	.25	.08	.35	12.1	2.7	62
Processing Speed Index	.14	.09	-.01	.22	105.9	15.2	56
Cancellation	-.14	-.16	-.10	-.07	10.2	3.1	49
Coding	.15	.08	-.02	.24	11.1	3.4	56
Symbol Search	.15	.12	.06	.16	11.1	2.8	62
Fluid Reasoning Composite	.60	.32	.36	.67	35.6	7.2	49
WG-II Form D							
Mean	27.6	8.3	8.0	11.2			
SD	6.0	2.7	2.5	2.9			
<i>n</i>	62	62	62	62			

\* The Fluid Reasoning composite was calculated by summing the scaled scores for Figure Weights, Matrix Reasoning, and Similarities.

At the Watson-Glaser II subscale level, it was expected that the subscale Draw Conclusions would be more highly correlated with the WAIS-IV than the Recognize Assumptions or Evaluate Arguments scales. Relative to other subscales, Draw Conclusions requires mental manipulation of a larger number of logical relationships, resulting in greater complexity. Therefore it was predicted that, among the three subscales, Draw Conclusions would be more strongly correlated with the Perceptual Reasoning, Working Memory, Verbal Comprehension Indices and the Fluid Reasoning Composite than Recognize Assumptions and Evaluate Arguments.

As predicted, the Draw Conclusions subscale was more highly correlated with WAIS–IV Perceptual Reasoning ( $r = .56$  versus  $r = .20$  for Recognize Assumptions and  $r = .25$  for Evaluate Arguments) Working Memory ( $r = .59$  versus  $r = .24$  for Recognize Assumptions and  $r = .13$  for Evaluate Arguments), Verbal Comprehension (e.g.,  $r = .46$  versus  $r = .34$  for Recognize Assumptions and  $r = .10$  for Evaluate Arguments), and Fluid Reasoning ( $r = .62$  versus  $r = .31$  for Recognize Assumptions and  $r = .21$  for Evaluate Arguments). These results also suggest that the subscale Recognize Assumptions is more closely associated with working memory, verbal comprehension, and fluid reasoning than Evaluate Arguments.

## Convergent Validity: Open Mindedness

Several prior studies have found significant relationships between Watson-Glaser scores and personality characteristics. For example, the Watson-Glaser correlated .34 with an Openness scale on the Personality Characteristics Inventory (Impelman & Graham, 2009), .36 with an Openness to Experience composite (derived from the CPI Achievement via Independence and Flexibility scales; Spector, Schneider, Vance, & Hezlett, 2000), and .33 with the Checklist of Educational Views which measures preferences for contingent, relativistic thinking versus “black-white, right-wrong” thinking (Taube, 1995). These findings suggest that the Watson-Glaser measures an attitudinal component of critical thinking.

In light of this prior research, the Watson-Glaser II scales were studied in relation to several personality measures. It was predicted that Watson-Glaser II total score would be moderately correlated (e.g., .20–.30s) with personality traits that measure inquisitiveness, openness, and a flexible attitude toward understanding the position of others (Facione 1990, 2009). It was also predicted that the subscale Evaluate Arguments would correlate negatively with personality measures that tap emotionality.

### Big 5

The Workplace Big Five Profile (Howard and Howard, 2001) measures an individual’s personality using five supertraits that conform to the traditional five-factor model of personality, and 24 subtraits (4–6 per factor). Openness was expected to correlate with Watson-Glaser II total score. At the subscale level, Intensity (a tendency to express anger) was expected to correlate negatively with Evaluate Arguments. To test these hypotheses, the Watson-Glaser II Form D and the Workplace Big Five Profile were administered to 72 professionals assessed for placement purposes by a northeastern recruiting organization.

Counter to the hypotheses, the Watson-Glaser II total score was not significantly correlated with Openness ( $r = .09$ ). At the subscale level, Intensity was not significantly correlated with Evaluate Arguments ( $r = -.17$ ), but was significantly correlated with both Draw Conclusions ( $r = -.27$ ) and the Watson-Glaser II total score ( $r = -.27$ ). These correlations suggest that elements of critical thinking performance are associated with a calm and composed, rather than tense or agitated, disposition. Relationships that were not hypothesized, but were significant included Recognize Assumptions scores with Consolidation ( $r = -.27$ ), and the Consolidation subtraits Perfectionism ( $r = -.30$ ) and Organization ( $r = -.27$ ).

### Myers-Briggs Type Indicator (MBTI)

The Myers-Briggs Type Indicator (Myers & Myers, 2004) measures four separate personality preferences, with each having opposite poles: Extraversion/Introversion, Sensing/Intuition, Thinking/Feeling, and Judging/Perceiving. The Watson-Glaser II Form D and MBTI were administered to 60 medical professionals working in a northeastern hospital network.

It was hypothesized that the Evaluate Arguments subscale would correlate with the MBTI Thinking/Feeling scale. A Feeling preference is characterized by a more personal or emotional investment in issues, which could impede the evaluation of arguments, especially those that are controversial. Consistent with expectations, Evaluate Arguments scores correlated  $-.27$  with the MBTI Feeling preference. There were no other significant correlations between the MBTI and the Watson-Glaser II.

## Golden Personality Type Profiler

The Golden Personality Type Profiler (Golden, 2004) measures an individual's personality using five global dimensions (Extraverting/Introverting, Sensing/Intuiting, Thinking/Feeling, Adapting/Organizing, Tense/Calm) and 36 facet scales. The Watson-Glaser II Form D and the Golden were administered to 123 undergraduate students at a southeastern university.

Similar to the prior study, it was predicted that Thinking and Feeling scales would be correlated with the Evaluate Arguments subscale. This hypothesis was not supported ( $r = .08$  with Thinking and  $-.02$  with Feeling). However, there were significant relationships with the Draw Conclusions subscale, including significant correlations with Thinking ( $r = .26$ ) and Feeling ( $r = -.21$ ) at the global dimension level and Autonomous ( $r = .26$ ), Analytic ( $r = .28$ ), and Warm ( $r = -.29$ ) at the facet level.

## Discriminate Validity

Prior research on the Watson-Glaser indicated that critical thinking was not related to Social Skills and Neuroticism (Robertson & Molloy, 1982). The preceding Watson-Glaser II studies provide additional evidence that critical thinking is not related to most major dimensions of personality. Watson-Glaser II total score was not related to MBTI or Golden Personality Type Profiler scales, such as Introversion/Extraversion, Sensing/Intuition, or Judging/Perceiving. It was also not related to the Workplace Big 5 Profile supertraits of Need for Stability, Extraversion, Accommodation, or Consolidation.

**Table 7.5 Demographic Data for Convergent and Criterion-Related Validity Studies**

	WAIS-IV	WBFP	MBTI	Golden	Insurance Company	Financial Services Company
<i>N</i>	62	75	60	123	68	35
Industry						
Education	22	—	1	9	—	—
Health Care	12	4	58	—	—	—
Manufacturing & Production	5	16	—	3	—	—
Financial Services, Banking, Insurance	4	—	—	5	—	33
Government, Public Service, Defense	4	—	—	2	—	—
Publishing, Printing	3	—	—	—	—	—
Transportation Warehousing	2	1	—	—	—	—
Advertising, Marketing, Public Relations	1	—	—	—	—	—
Aerospace, Aviation	1	—	—	1	—	—
Construction	1	—	—	6	—	—
Information Technology, High-Tech, Telecommunications	1	—	—	1	—	—
Retail & Wholesale	—	19	—	28	—	—
Pharmaceuticals, Biotechnology	—	—	1	—	—	—
Energy, Utilities	—	—	—	—	—	—
Hospitality, Tourism	—	—	—	22	—	—
Professional, Business Services	—	—	—	2	—	1
Real Estate	—	—	—	2	—	—
Other	5	32	—	26	—	—
Not Applicable	1	3	—	16	—	1
Position						
Executive	1	1	1	1	—	5
Director	4	4	8	1	—	3
Manager	3	56	33	2	34	7
Professional/Individual Contributor	33	5	6	4	—	12
Supervisor	1	—	5	1	34	7
Public Safety	2	—	—	—	—	—
Self-Employed/Business Owner	4	—	—	3	—	—

**Table 7.5 Demographic Data for Convergent and Criterion-Related Validity Studies**

	WAIS-IV	WBFP	MBTI	Golden	Insurance Company	Financial Services Company
Administrative/Clerical	2	—	5	13	—	—
Skilled Trades	2	—	—	4	—	—
General Labor	1	—	—	7	—	—
Customer Service/Retail Sales	—	1	—	58	—	—
Not Applicable	9	8	2	29	—	—
<b>Ethnicity</b>						
White non-Hispanic	44	62	52	84	57	30
Hispanic, Latino/a	8	1	2	17	2	—
Black, African American	4	2	—	5	2	1
Asian/Pacific Islander	1	—	—	4	5	—
Native American	—	—	—	—	—	—
Other	—	1	1	1	1	1
Multiracial	—	—	—	3	1	—
No Response	5	9	5	9	—	3
<b>Age</b>						
16-20	—	—	—	56	1	—
21-24	—	1	—	41	—	1
25-29	5	7	2	14	12	4
30-34	9	12	7	1	16	—
35-39	6	12	9	1	15	11
40-49	10	21	18	1	19	9
50-59	12	9	17	—	5	6
60-69	6	4	1	—	—	—
70+	11	—	—	—	—	—
No Response	3	9	6	9	—	4
<b>Sex</b>						
Female	32	9	38	54	22	20
Male	27	57	17	61	46	12
No Response	3	9	5	8	—	3
<b>Education Level</b>						
HS/GED	—	3	5	7	1	—
1–2 yrs college	—	6	12	48	3	5
Assoc.	—	3	8	18	1	—
3–4 yrs college	1	3	2	40	1	1
Bachelor's	27	44	10	2	56	17
Master's	25	7	14	—	6	9
Doctorate	4	—	2	—	—	—
No Response	5	9	7	8	—	3

## Criterion-Related Validity

One of the primary reasons tests are used is to predict an examinee's potential for future success. Criterion-related validity evidence occurs when a statistical relationship exists between scores on the test and one or more criteria, or between scores on the tests and measures of performance. By collecting test scores and criterion scores (e.g., job performance ratings, grades in a training course, supervisor ratings), one can determine how much confidence may be placed in using test scores to predict job success. Provided that the conditions for a meaningful validity study have been met (sufficient sample size, adequate criteria, etc.), these correlation coefficients are important indices of the utility of the test.

Cronbach (1970) characterized validity coefficients of .30 or better as having "definite practical value." The U.S. Department of Labor (1999) provides the following general guidelines for interpreting validity coefficients: above .35 are considered "very beneficial," .21–.35 are considered "likely to be useful," .11–.20 "depends on the circumstances," and below .11 "unlikely to be useful." It is important to point out that even relatively lower

validities (e.g., .20) may justify the use of a test in a selection program (Anastasi & Urbina, 1997). The practical value of the test depends not only on the validity, but also other factors, such as the base rate for success (i.e., the proportion of people who would be successful in the absence of any selection procedure). If the base rate for success is low (i.e., few people would be successful on the job), tests of low validity can have considerable utility or value. When the base rate is high (i.e., selected at random, most people would succeed on the job), even highly valid tests may not contribute significantly to the selection process.

### Prior Evidence of Criterion-Related Validity

Previous studies of the Watson-Glaser have demonstrated a positive relationship between Watson-Glaser scores and various job and academic success criteria. A complete summary of these studies is provided in the manuals for the Watson-Glaser Short Form (2006) and Forms A and B (1980), respectively. A few selected findings are highlighted here.

The Watson-Glaser is correlated with organizational success. For example, Pearson (2006) found that for 2,303 job incumbents across 9 industry categories, Watson-Glaser scores correlated .33 with job success as indicated by organizational level achieved.

The Watson-Glaser is also correlated with potential to advance, job performance, and specific job performance capabilities related to thinking, problem solving, analysis, and judgment. Spector et al., (2000) evaluated the relationship between Watson-Glaser scores and assessment center exercise performance for managerial and executive level assessment center participants. They found that Watson-Glaser scores significantly correlated with six of eight assessment center exercises, and related more strongly to exercises involving cognitive problem-solving skills (e.g.,  $r = .26$  with in-basket scores) than exercises involving interpersonal skills (e.g.,  $r = .16$  with in-basket coaching exercise). Scores also correlated .28 with "Total Performance," a sum of ratings on 19 job performance behaviors, and .24 with ratings on a single-item measure of Overall Potential.

Using a sample of 71 leadership assessment center participants, Kudish and Hoffman (2002) reported that Watson-Glaser scores correlated .58 with ratings on Analysis and .43 with ratings on Judgment. Ratings on Analysis and Judgment were based on participants' performance across assessment center exercises including a coaching meeting, in-basket exercise or simulation, and a leaderless group discussion. Using a sample of 142 job incumbents, Pearson (2006) found that Watson-Glaser scores correlated .33 with supervisory ratings of Analysis and Problem Solving behaviors, and .23 with supervisory ratings on a dimension made up of Judgment and Decision Making behaviors. Using a sample of 64 analysts from a government agency, Pearson (2006) found that Watson-Glaser scores correlated .40 with supervisory ratings on each of two dimensions composed of (a) Analysis and Problem Solving behaviors and, (b) Judgment and Decision Making behaviors, and correlated .37 with supervisory ratings on a dimension composed of behaviors dealing with Professional/Technical Knowledge and Expertise. The Watson-Glaser scores correlated .39 with "Total Performance" and .25 with Overall Potential.

In the educational domain, Behrens (1996) found that Watson-Glaser scores correlated .59, .53, and .51 respectively, with semester GPA for three freshmen classes in a Pennsylvania nursing program. Similarly, Gadzella, Baloglu, & Stephens (2002) found Watson-Glaser subscale scores explained 17% of the total variance in GPA (equivalent to a multiple correlation of .41) for 114 Education students. Williams (2003), in a study of 428 educational psychology students, found Watson-Glaser total scores correlated .42 and .57 with mid-term and final exam scores, respectively. Gadzella, Ginther, and Bryant (1996) found in a study of 98 college freshmen that Watson-Glaser scores were significantly higher for A students than B and C students, and significantly higher for B students relative to C students.

Studies have also shown significant relationships between Watson-Glaser scores and clinical decision making effectiveness (Shin, 1998), educational experience and level (Duchesne, 1996; Shin, 1998; Yang & Lin, 2004), educational level of parents (Yang & Lin, 2004), and academic performance during pre-clinical years of medical education (Scott & Markert, 1994).

**Table 7.6 Previous Studies Showing Evidence of Criterion-Related Validity**

Group	N	Watson-Glaser			Criterion				
		Form	Mean	SD	Description	Mean	SD	r	
Job incumbents across multiple industries (Pearson, 2006)	142	Short	30.0	6.0	Supervisory Ratings:				
					Analysis and Problem Solving	37.5	6.7	.33**	
					Judgment and Decision Making	31.9	6.1	.23**	
					Total Performance Potential	101.8	16.9	.28**	
					3.1	1.2	.24**		
Job applicants and incumbents across multiple industries (Pearson, 2006)	2,303	Short	31.0	5.6	Org. Level	3.1	1.2	.33*	
Analysts from a government agency (Pearson, 2006)	64	Short	32.9	4.4	Supervisory Ratings:				
					Analysis and Problem Solving	38.3	6.6	.40**	
					Judgment and Decision Making	32.8	5.8	.40**	
					Professional / Technical Knowledge & Expertise	17.1	2.4	.37**	
					Total Performance Potential	100.4	14.3	.39**	
					3.2	1.2	.25*		
Leadership assessment center participants from a national retail chain and a utility service (Kudish & Hoffman, 2002)	71	80-item	—	—	Assessor Ratings:				
					Analysis	—	—	.58*	
					Judgment	—	—	.43*	
Middle-management assessment center participants (Spector, Schneider, Vance, & Hezlett, 2000)	189–407	—	66.5	7.3	Assessor Ratings:				
					In-basket	2.9	0.7	.26*	
					In-basket Coaching	3.1	0.7	.16*	
					Leaderless Group	3.0	0.6	.19*	
					Project Presentation	3.0	0.7	.25*	
					Project Discussion	2.9	0.6	.16*	
					Team Presentation	3.1	0.6	.28*	
					41.8	6.4	.36*		
Freshmen classes in a Pennsylvania nursing program (Behrens, 1996)	41	80-item	50.5	—	Semester 1 GPA	2.5	—	.59**	
	31		52.1	—	Semester 1 GPA	2.5	—	.53**	
	37		52.1	—	Semester 1 GPA	2.4	—	.51**	
Education majors (Gadzella, Baloglu, & Stephens, 2002)	114	80-item	51.4	9.8	GPA	3.1	.51	.41**	
Educational psychology students (Williams, 2003)	158–164	Short	—	—	Exam 1 Score	—	—	.42**	
					Exam 2 Score	—	—	.57**	
Job applicants and incumbents across multiple industries (Pearson, 2006)	919	Short	31.1	5.6	Education Level	8.0	1.8	.40**	
Education majors (Taube, 1995)	147–194	80-item	54.9	8.1	GPA	2.8	.51	.30*	
Educational psychology students (Gadzella, Stephens, & Stacks, 2004)	139	80-item	—	—	Course Grades	—	—	.42**	
					GPA	—	—	.28**	

\* $p < .05$ . \*\* $p < .01$

## Studies Using the Watson-Glaser II

The Watson-Glaser II is correlated with occupational and educational attainment. In the standardization sample, 432 people provided job level information and 581 indicated their level of educational attainment. The correlations between Watson-Glaser II and occupational and educational attainment were .28 and .33, respectively.

The relationship between the Watson-Glaser II and job performance was examined using a sample of 68 managers and their supervisors from the claims division of a national insurance company. Managers completed the Watson-Glaser II and supervisors of these participants rated the participants' job performance across thinking domains (e.g., Creativity, Analysis, Critical Thinking, Job Knowledge) and Overall Performance and Potential.

Table 7.7 presents means, standard deviations, and correlations. Results showed that Watson-Glaser II total score correlated .28 with supervisory ratings on a scale of Core Critical Thinking Behaviors and .25 with ratings of Overall Potential.

The pattern of relationships at the subscale level indicated that Draw Conclusions correlated significantly with all performance ratings except Job Knowledge ( $r = .23$ , *ns*). Evaluate Arguments was significantly related only to Job Knowledge ( $r = .26$ ), and Recognize Assumptions was not significantly related to any of the performance dimensions.

**Table 7.7** Descriptive Statistics and Correlations for Watson-Glaser II Scores and Performance Ratings

Supervisory Performance Criteria	Watson-Glaser II Form D Score				Performance Ratings		
	Total Score	Recognize Assumptions	Evaluate Arguments	Draw Conclusions	Mean	SD	<i>n</i>
Core Critical Thinking Behaviors	0.28	0.06	0.17	0.37	162.0	18.0	68
Evaluating Quality of Reasoning and Evidence	0.27	0.05	0.17	0.35	81.4	7.2	68
Bias Avoidance	0.22	0.04	0.19	0.25	22.5	4.2	68
Creativity	0.24	0.06	0.12	0.33	23.3	4.7	68
Job Knowledge	0.24	0.03	0.26	0.23	22.1	3.5	68
Overall Performance	0.09	-0.15	0.04	0.31	4.9	0.7	68
Overall Potential	0.25	0.03	0.14	0.37	3.2	0.9	68
<b>WG-II Form D</b>							
Mean	28.3	8.5	7.9	11.9			
SD	5.3	2.6	2.3	2.6			
<i>n</i>	68	68	68	68			

A second study examined the relationship between Watson-Glaser II scores and job performance using 35 professionals at a large financial services company. Incumbents were ranked by the human resource staff familiar with their performance. The ranking involved categorizing incumbents into a "top" and "contrast" group based on critical thinking effectiveness demonstrated over time. Pearson provided the human resources staff with a list of behaviors typically exhibited by strong and weak critical thinkers, respectively, to help guide rankings.

Table 7.8 presents a comparison of average Watson-Glaser II total and subscale scores achieved for each group. As expected, the group of top ranked critical thinkers achieved higher Watson-Glaser II total and subscale scores than the contrast group.

**Table 7.8 Mean Performance of Highly Ranked Critical Thinkers and a Contrast Group**

Watson-Glaser II Score	Top Critical Thinkers ( <i>n</i> = 23)		Contrast Group ( <i>n</i> = 12)		<i>n</i>	Group Mean Comparison			
	Mean	<i>SD</i>	Mean	<i>SD</i>		Difference	<i>F</i> value	<i>p</i> value	Cohen's <i>d</i>
Total Score	31.8	3.9	25.5	6.7	35	6.33	12.66	<.01	1.27
Recognize Assumptions	9.5	1.3	7.6	3.0	35	2.05	7.05	.01	.95
Evaluate Arguments	9.2	1.6	6.8	1.9	35	1.94	15.01	<.01	1.38
Draw Conclusions	13.1	2.5	11.1	2.9	35	2.34	4.59	.04	.76

## Summary

Collectively, the evidence of content, construct, and criterion-related validity for the Watson-Glaser II is solid and the relationships are of a magnitude similar to those found with prior versions of the instrument. The Watson-Glaser II measures the cognitive abilities that underlie critical thinking skills. There is some, albeit limited, evidence that its components are also related to attitudes or personality preferences that can foster critical thinking performance. Finally, the Watson-Glaser II is associated with occupational and educational attainment and job performance, especially performance related to thinking and problem solving competencies.

## Directions for Administration

### Computer-Based Administration

The computer-based Watson-Glaser II is administered through TalentLens.com, an Internet-based testing system designed by Pearson for the administration, scoring, and reporting of professional assessments. Because examinee data is instantly captured for processing through this online system, you can immediately access scores and interpretive reports.

### Preparing for Administration

If you are not familiar with the Watson-Glaser II, we recommend that you take the computer-based test prior to administering the test, being sure to comply with the directions. Doing so will help you anticipate questions or issues that may arise during test administration.

Being thoroughly prepared before an examinee's arrival will result in a more efficient online administration session. Examinees will not need pencils or scratch paper for this computer-based test. In addition, do not allow examinees to have access to any reference materials (e.g., dictionaries or calculators).

### Testing Conditions

It is important to ensure that the test is administered in a quiet, well-lit room. The following conditions are necessary for accurate scores and for maintaining the cooperation of the examinee: good lighting, comfortable seating, adequate desk or table space, comfortable positioning of the computer screen, keyboard, and mouse, and freedom from noise and other distractions.

### Answering Questions

Though the instructions for completing the assessment are presented on-screen, it is important to develop and maintain rapport with participants. You should ensure that participants understand all requirements and how to interact with the assessment interface appropriately.

Examinees may ask questions about the assessment before they begin taking it. Clarification of what is required of examinees and confirmation that they understand these requirements are appropriate.

If examinees have routine questions after the testing has started, try to answer them without disturbing the other examinees. However, avoid explaining the meaning of words or items to examinees, as this could lead to inappropriate prompting of examinees toward certain responses. If examinees have questions about the interpretation of an item, you should encourage them to respond to the item as they best understand it.

### Administering the Test

After the initial instruction screen for the Watson-Glaser II has been accessed and the examinee is seated at the computer, say,

**The on-screen directions will take you through the entire process which begins with some demographic questions. After you have completed these questions, the test will begin. You will have as much time as you need to complete the test items. The test ends with a few additional demographic questions. Do you have any questions before starting the test?**

Answer any questions and say, **Please begin the test.**

Once the examinee completes the demographic questions and clicks the NEXT button, test administration begins with the first page of test questions. During the test, examinees have the option of skipping items and returning to them later. The examinee also may review test items at the end of the test. Examinees have as much time as they need to complete the exam, but they typically finish within 35 minutes.

## Technical Issues

If an examinee's computer develops technical problems during testing, you should move the examinee to another suitable computer location. If the technical problems cannot be solved by moving to another computer location, contact Pearson Technical Support for assistance. The contact information, including phone numbers, can be found at the TalentLens.com website.

## Scoring and Reporting

The score report is available in the administrator's TalentLens.com account for viewing on screen, printing, or saving as a .pdf file.

## Paper-and-Pencil Administration

The majority of our customers administer the Watson-Glaser II online. However, if you do need to use paper administration, the following administration practices apply.

## Preparing for Administration

You do not need special training to administer the Watson-Glaser II, but you must be able to carry out standard examination procedures. To ensure accurate and reliable results, you should become thoroughly familiar with the administration instructions and the test materials before attempting to administer the test. If you are not familiar with the Watson-Glaser II, you should take the test prior to administration, being sure to comply with the directions and any time requirement. Doing so will help anticipate questions or issues that may arise during test administration.

## Testing Conditions

Generally accepted conditions of good test administration should be observed: good lighting, comfortable seating, adequate desk or table space, and freedom from noise and other distractions. Examinees should have sufficient seating space to minimize cheating.

Each examinee needs an adequate flat surface on which to work. Personal materials should be removed from the work surface. Do not allow examinees to have access to any reference materials (e.g., dictionaries or calculators).

## Materials Needed to Administer the Test

- This Manual
- 1 Test Booklet for each examinee
- 1 Answer Sheet for each examinee
- 2 No. 2 pencils with erasers for each examinee
- A clock or stopwatch if the test is timed
- 1 Hand-Scoring Key (if the test will be hand-scored rather than scanned)

Intended as a test of critical thinking power rather than speed, the Watson-Glaser II may be given in either timed or untimed administrations. For timed administrations, a 40-minute time limit is recommended. Forty minutes should allow the vast majority of examinees to complete the test while working at a reasonably comfortable pace.

You should have a stopwatch, a watch with a second hand, a wall clock with a second hand, or any other accurate device to time the test administration. To facilitate accurate timing, the starting time and the finishing time should be written down immediately after the signal to begin has been given. In addition to testing time, allow 5–10 minutes to read the directions on the cover of the test booklet and answer questions.

## Answering Questions

Examinees may ask questions about the test before you give the signal to begin. To maintain standard testing conditions, answer such questions by rereading the appropriate section of the directions. Do not volunteer new explanations or examples. It is the responsibility of the test administrator to ensure that examinees understand the correct way to indicate their answers on the Answer Sheet and what is required of them. The question period should never be rushed or omitted.

If any examinees have routine questions after the testing has started, try to answer them without disturbing the other examinees. However, explaining the meaning of words or items to candidates must be avoided, as this could lead to inappropriate prompting of candidate responses. If candidates have questions about the interpretation of an item, they should be encouraged to respond to the item as they best understand it.

## Administering the Test

All directions that the test administrator reads aloud to examinees are in **bold type**. Read the directions exactly as they are written, using a natural tone and manner. Do not shorten the directions or change them in any way. If you make a mistake in reading a direction, say,

**No that is wrong. Listen again.**

Then read the direction correctly.

When all examinees are seated, give each examinee two pencils and an Answer Sheet.

Say **Please make sure that you do not fold, tear, or otherwise damage the Answer Sheets in any way. Notice that your Answer Sheet has an example of how to properly blacken the circle.**

Point to the “Correct Mark” and “Incorrect Marks” samples on the Answer Sheet.

Say **Make sure that the circle is completely filled in as shown.**

**Note.** You may want to point out how the test items are ordered on the front page of the Watson-Glaser II Answer Sheet so that examinees do not skip anything or put information in the wrong place.

Say **In the upper left corner of the Answer Sheet, you will find box “A” labeled NAME. Neatly print your Last Name, First Name, and Middle Initial here. Fill in the appropriate circle under each letter of your name.**

The Answer Sheet provides space for a nine-digit identification number. If you want the examinees to use this space for an employee identification number, provide them with specific instructions for completing the information at this time. For example, say, **In box “B” labeled IDENTIFICATION NUMBER, enter your employee number in the last four spaces provided. Fill in the appropriate circle under each digit of the number.** If no information is to be recorded in the space, tell examinees that they should not write anything in box B.

Say **Find box “C”, labeled DATE. Write down today’s Month, Day, and Year here. (Tell examinees today’s date.) Blacken the appropriate circle under each digit of the date.**

Box D, labeled OPTIONAL INFORMATION, provides space for additional information you would like to obtain from the examinees. Let examinees know what information, if any, they should provide in this box.

**Note.** It is recommended that if optional information is collected, the test administrator inform examinees of the purpose of collecting this information (i.e., how it will be used).

Say **Are there any questions?**

Answer any questions.

Say **After you receive your Test Booklet, please keep it closed. You will do all your writing on the Answer Sheet only. Do not make any additional marks on the Answer Sheet until I tell you to do so.**

Distribute the Test Booklets.

Say **In this test, all the questions are in the Test Booklet. There are five separate sections in the booklet, and each one is preceded by its own directions. For each question, decide what you think is the best answer. Because your score will be the number of items you answered correctly, try to answer each question even if you are not sure that your answer is correct.**

**Record your choice by making a black mark in the appropriate space on the Answer Sheet. Always be sure that the answer space has the same number as the question in the booklet and that your marks stay within the circles. Do not make any other marks on the Answer Sheet. If you change your mind about an answer, be sure to erase the first mark completely.**

**Do not spend too much time on any one question. When you finish a page, go right on to the next one. When you finish all the sections, you may go back and check your answers.**

### Timed Administration

Say **You will have 40 minutes to work on this test. Now read the directions on the cover of your Test Booklet.**

After allowing time for the examinees to read the directions, say,

**Are there any questions?**

Answer any questions, preferably by rereading the appropriate section of the directions, then say, **Ready? Please begin the test.**

Start timing immediately. If any of the examinees finish before the end of the test period, either tell them to sit quietly until everyone has finished, or collect their materials and dismiss them. At the end of 40 minutes, say,

**Stop! Put your pencils down. This is the end of the test.**

Intervene if examinees continue to work on the test after the time signal is given.

### Untimed Administration

Say **You will have as much time as you need to work on this test. Now read the directions on the cover of your Test Booklet.**

After allowing time for the examinees to read the directions, say,

**Are there any questions?**

Answer any questions, preferably by rereading the appropriate section of the directions, then instruct examinees regarding what they are to do upon completing the test (e.g., remain seated until everyone has finished, bring Test Booklet and Answer Sheet to the test administrator).

Say **Ready? Please begin the test.**

Allow the group to work until everyone is finished.

### Concluding Administration

At the end of the testing session, collect all Test Booklets, Answer Sheets, and pencils. Place the completed Answer Sheets in one pile and the Test Booklets in another. The Test Booklets may be reused, but they will need to be inspected for marks. Marked booklets should not be reused, unless the marks can be completely erased.

## Scoring

The Watson-Glaser II Answer Sheet may be hand scored with the Hand-Scoring Key or machine scored. The directions for hand-scoring are also included on the Hand-Scoring Key.

### Scoring With the Hand-Scoring Key

Before you start scoring, review each Answer Sheet for multiple responses to the same item. Draw a heavy red mark through such responses. These items receive no credit. If you find any partially erased responses, erase them completely.

To score responses, place the correct Scoring Key (for Form D or Form E) over the Answer Sheet and align the two stars with the two holes that are closest to the bottom of the key. Make sure the correct circle is blackened in Box E, FORM USED on the Answer Sheet, and that it shows through the hole for Form Used on your Scoring Key stencil.

There are three subscale raw scores to be recorded: Recognize Assumptions (Test 2), Evaluate Arguments (Test 5), and Draw Conclusions (Tests 1, 3, and 4). Each subscale's set of answers is bordered by a dashed-line on the Scoring Key. Follow the arrows on the Scoring Key as you count the number of correctly marked spaces through the holes on the stencil. Record the RA, EA, and DC raw scores in the Score box on the Answer Sheet. Then add the three subscale raw scores to get the critical thinking Total Raw score. Record it on the Answer Sheet.

Use the Watson-Glaser II Form D or E norms tables to convert the subscale scores to ranges (Low, Average, or High), and the Total Raw score to a percentile rank, which can be recorded in row two of the "Score" box. You may record the name of the norm group that you used in row three.

### Machine Scoring

First, completely erase multiple responses to the same item or configure the scanning program to treat multiple responses as incorrect answers. If you find any answer spaces that were only partially erased by the examinee, finish completely erasing them.

The machine-scorable Answer Sheets available for the Watson-Glaser II may be processed with any reflective scanning device programmed to your specifications. Pearson does not offer scanning services for the Watson-Glaser II.

## Additional Considerations for Administration

### Test Security

Watson-Glaser II scores and reports are confidential and should be stored in a secure location accessible only to authorized individuals. It is unethical and poor test practice to allow test score/report access to individuals who do not have a legitimate need for the information. The security of testing materials and protection of copyright must also be maintained by authorized individuals. Storing test scores and materials in a locked cabinet (or password-protected file in the case of scores maintained electronically) that can only be accessed by designated test administrators is an effective means to ensure their security. Avoid disclosure of test access information such as usernames and passwords and only administer the Watson-Glaser II in proctored environments. All the computer stations used in administering the computer-based Watson-Glaser II must be in locations that can be easily supervised.

## Differences in Reading Ability—English as a Second Language

Watson-Glaser II directions and items were written at or below the 9th grade reading level. Because a level of reading proficiency in the English language is assumed, reasonable precautions must be taken when assessing candidates whose first language is not English. When possible, the Watson-Glaser II should be administered in the examinee's first language. Contact your Pearson Account Manager for information on language versions available. If a version is not available in the examinee's first language and he or she has difficulty with the language or the reading level of the items, note this and consider it when interpreting the scores.

## Accommodating Examinees with Disabilities

The Americans with Disabilities Act (ADA) of 1990 requires an employer to reasonably accommodate the known disability of a qualified applicant, provided such accommodation would not cause an "undue hardship" to the operation of the employer's business. Therefore, you should provide reasonable accommodations to enable candidates with special needs to comfortably take the test. Reasonable accommodations may include, but are not limited to, modification of the assessment format and procedure, such as live assistance, in which an intermediary reads the test content to a visually impaired candidate and marks their answers for them (Society for Industrial and Organizational Psychology, 2003). Consult with your qualified legal advisor or human resource professional for additional guidance on providing appropriate reasonable accommodations.

# Using the Watson-Glaser II as an Employment Selection Tool 9

## Employment Selection

Many organizations use testing as a component of their employment selection process. Typical selection test programs make use of cognitive ability tests such as the Watson-Glaser II, aptitude tests, personality tests, and basic skills tests, to name a few. Tests are used as part of a larger battery (e.g., resumes, interviews) to screen out unqualified candidates or to categorize prospective employees according to their probability of success on the job.

The Watson-Glaser II is designed to assist in the selection of employees for jobs that require careful analysis and problem solving. Many executive, managerial, and other professional positions require the type of critical thinking abilities and skills measured by the Watson-Glaser II. The test can be used to assess applicants for a wide variety of professional jobs, including individual contributor positions (e.g., non-retail sales, nurse, accountant), and lower-to-upper level management jobs across industries, occupations, and education levels.

Organizations using the Watson-Glaser II are encouraged to conduct a local validation study that establishes the relationship between examinees' scores and their subsequent performance on the job. Local validation can be conducted with a concurrent study in which incumbents' test scores are correlated with measures of performance. A more resource- and time-consuming approach is to conduct a predictive study in which applicants' scores are not initially used for selection decisions, but are correlated with job performance at some designated point in the future (e.g., 6 months performance ratings). This information will inform score interpretation and will most effectively enable a Watson-Glaser II user to set cut scores to differentiate examinees who are likely to be successful from those who are not.

It is ultimately your responsibility to determine how you will use the Watson-Glaser II scores. If you establish a cut score, examinees' scores should be considered in the context of appropriate measurement data for the test, such as the standard error of measurement and data regarding the predictive validity of the test. In addition, selection decisions should always be based on multiple job-relevant measures rather than relying on any single measure (e.g., using only Watson-Glaser II scores to make decisions).

## Fairness in Selection Testing

Fair employment regulations and their interpretation are continuously subject to changes in the legal, social, and political environments. It therefore is advised that you consult with qualified legal advisors and human resources professionals as appropriate.

## Legal Considerations

There are governmental and professional regulations that cover the use of all personnel selection procedures. Relevant source documents that the user may wish to consult include the *Standards for Educational and Psychological Testing* (AERA et al., 1999); the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 2003); and the federal *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, 1978). For an overview of the statutes and types of legal proceedings which influence an organization's equal employment opportunity obligations, the user is referred to Cascio and Aguinis (2005) or the *U.S. Department of Labor's (2000) Testing and Assessment: An Employer's Guide to Good Practices*.

## Group Differences/Adverse Impact

According to the *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, 1978), adverse impact is normally indicated when the selection rate for one group is less than 80% (or 4 out of 5) that of another. Adverse impact is likely to occur with cognitive ability tests such as the Watson-

Glaser II. A test with adverse impact can be used for selection (Equal Employment Opportunity Commission, 1978), but the testing organization must demonstrate that the selection test is job-related, predicts performance, and is consistent with business necessity. A local validation study, in which scores on the Watson-Glaser II are correlated with indicators of on-the-job performance, provides evidence to support the use of the test in a particular job context. In addition, a local study that demonstrates that the Watson-Glaser II is equally predictive for protected subgroups, as outlined by the Equal Employment Opportunity Commission, will help establish test fairness. Additional guidance on monitoring your selection system for fairness is provided in the following section.

## Monitoring the Selection System

To evaluate selection strategies and to implement fair employment practices, an organization needs to know the demographic characteristics of applicants and incumbents. Monitoring these characteristics and accumulating test score data are necessary for establishing legal defensibility of a selection system, including those systems that incorporate the Watson-Glaser II. The most effective use of the Watson-Glaser II will be achieved where the following best practices are incorporated over time:

- At least once every 5 years, conduct a job analysis of the position for which you are administering the Watson-Glaser II. A job analysis will help you determine if the job has changed in a way that requires adjustments to your assessment system.
- Periodically (e.g., once every 5 years) reassess the criterion-related validity of the selection system through local validation studies.
- Carefully monitor assessment scores for evidence of adverse impact. Adverse impact is typically evaluated by comparing the rate of selection for individuals from EEOC protected subgroups (e.g., gender or ethnicity) with selection rates of historically advantaged groups. Information that should be recorded to facilitate these analyses include applicant demographics (e.g., voluntary information on gender, race/ethnicity, and age), assessment scores, and employment status (e.g., hired/not hired).
- Periodically reexamine cut scores considering recent validity results, adverse impact, market data, and other factors (e.g., projected workload) and make adjustments as necessary.
- When sufficient samples of employees and candidates have been obtained (e.g., >25 per demographic group), conduct a study to observe whether the selection procedure predicts equally for the majority group and EEOC protected subgroups. The Cleary model is a commonly-used approach to evaluate the fairness of selection tools (Guion, 1998). This model utilizes regression analysis to determine if a test demonstrates differential validity or prediction among subgroups of applicants. That is, it determines if a test over- or under-predicts job performance based on subgroup membership.

Pearson offers professional services to facilitate local validation and test fairness research in your organization. Contact your Account Manager for more information.

## The Watson-Glaser II Profile Report

The Profile Report provides an overview of the examinee's performance on the Watson-Glaser II. You can use this report to inform selection decisions, as well as build awareness of an examinee's strengths and development needs. To help you better understand the constructs measured by the assessment, the report provides detailed definitions for each of the skills (i.e., Recognize Assumptions, Evaluate Arguments, and Draw Conclusions). To facilitate your interpretation of an examinee's results, the report includes both numerical and graphical presentations of the overall score and the subscale score ranges. To help you envision how the scores might translate to actual behaviors, brief interpretive summaries are offered at the overall and subscale level. For selection purposes, use only the overall score rather than the subscale scores—the overall score provides a more precise and complete picture of a candidate's critical thinking skills. Review the next section for guidance on the interpretation of test results.

### Interpreting Test Results Using Norms and Percentiles

A raw score is the total number of correct responses. The maximum overall raw score for the Watson-Glaser II is 40, comprised of 12 Recognize Assumptions items, 12 Evaluate Arguments items, and 16 Draw Conclusions items. Raw scores may be used to rank examinees in order of performance, but little can be inferred from raw scores alone. It is important to relate the scores to specifically defined normative groups (i.e., "norms") to make the test results meaningful. Norms provide a basis for evaluating an individual's score relative to the scores of other individuals who took the same test. They are typically constructed from the scores of a large sample of individuals who took a test. This group of individuals is referred to as the normative (norm) group.

### Using Local Norms

The most appropriate norm group is one that is most representative of those who will be taking the test in the local situation (i.e., local norms). You can develop local norms by accumulating the test scores of applicants, trainees, employees for a given position, or students. However, you should consider whether your sample size will be sufficient to create a local norm. With large samples, the distribution of scores is more stable and all possible test scores can be converted to percentile ranks. Data from smaller samples tend to be unstable, and the presentation of percentile ranks for each score presents an unwarranted impression of precision. Until you can collect a sufficient and representative number of cases (preferably 100 or more) to create your own local norms, you can use the norms established by Pearson to help you interpret test scores.

### Using Pearson Norms

When selecting norms developed by Pearson, look for a group that is similar to the individual or group being tested. For example, you would compare the test score of a candidate who applied for an engineer's position with norms derived from the scores of other engineers. If a candidate applied for a management position, you would compare his or her test score with norms for managers, or norms for managers in manufacturing if that was more accurate. Keep in mind that norms are affected by the composition of the groups that participated in the normative study. Therefore, be sure to examine the specific characteristics (e.g., industry, occupation, etc.) of any norm groups you are considering. Some examples of norms that Pearson currently offers for the Watson-Glaser II include Executive, Director, Manager, Supervisor, and Professional/Individual Contributor. Details of the normative samples' occupational composition and other demographics are available through your Account Manager. The norm groups available and their composition characteristics are updated frequently, so contact your Account Manager for the most current offerings.

## Interpreting Percentiles

The percentile rank indicates an examinee's relative position in the norm group. Percentile ranks should not be confused with percentage scores, which represent the percentage of correct items. Percentile ranks are derived scores that are expressed as the percent of people in the norm group scoring equal to or below a given raw score.

Although percentile ranks are useful for explaining an examinee's performance relative to others, they have limitations. Percentile ranks do not have equal intervals. In a normal distribution of scores, percentile ranks tend to cluster around the 50th percentile. This clustering affects scores in the average range the most because a difference of one or two raw score points may change the percentile rank. Extreme scores are less affected; a change in one or two raw score points typically does not produce a large change in percentile ranks. Be sure to consider these factors when interpreting percentiles.

*Example.* If a person is applying for a position as a manager, it is appropriate to use the Manager norm group for comparison. If this candidate achieved a raw score of 35 on the Watson-Glaser II Form D, the percentile rank corresponding to a raw score of 35 is 86 for the Manager norm group. This percentile rank indicates that about 86% of the people in the Manager norm group scored lower than or equal to a raw score of 35 on the Watson-Glaser, and therefore about 14% scored higher than a score of 35 on the Watson-Glaser. If a less representative norm group were used, the percentile rank could be much higher or lower, and would inappropriately compare the candidate to a group of people less like his or her peers.

## Score Ranges Used for Reports

Each of the Watson-Glaser II reports—Profile, Interview, and Development—offers customized information based on an examinee's score ranges. The score ranges were derived empirically using data on overall and subscale score distributions, as well as criterion performance levels for each range. For the Profile report, the overall test score uses the examinee's percentile rank within a given norm group to determine his or her score range. Examinees with scores equal to or less than the 30th percentile are described as “below average” in applying the critical thinking necessary for effective analysis and decision making; examinees with scores between the 31st and 70th percentiles are described as being “moderately skilled and consistent;” examinees with scores equal to or greater than the 71st percentile are described as being “highly skilled and consistent.” A similar approach was employed at the subscale level; however the raw subscale scores are converted to stanine scores, rather than percentile scores, within a given norm group. Given the smaller number of items for each subscale, the use of percentiles would result in large jumps in percentile ranks and would give an inappropriate impression of precision at the subscale level. By comparison, stanine scores convert a distribution of raw scores into nine distinct categories (instead of 100 categories, which is essentially what percentiles do), which is a more appropriate form of categorization at the subscale level. The score ranges for all three subscales were defined as follows: scores falling into stanine 1 were described as being in the “low range”; scores falling into stanines 2–5 were described as being in the “average range”; and scores falling into stanines 6–9 were described as being in the “high range.” The customized content for both the Interview and Development reports is based on the examinee's subscale score ranges as defined above. The next sections provide additional detail on the design and content of those reports.

## The Watson-Glaser II Interview Report

The Watson-Glaser II Interview Report is a behavioral interview guide designed as a companion to the Watson-Glaser II for further evaluation of examinees' critical thinking. Probing critical thinking skills in an interview, in conjunction with the assessment results, provides you with a richer picture of how the examinee's critical thinking skills are likely to appear day-to-day. The report was designed to facilitate interviewing for selection purposes, although it could reasonably be used as a means of assessing an examinee's developmental needs for critical thinking as well (e.g., as the first step in a coaching process).

A primary feature of the report is the structure it provides around interviewing an examinee. The literature clearly

demonstrates that providing more structure to an interview process increases the reliability of the interview and its validity for predicting future performance criteria (McDaniel et al., 1994). Within that structure, the content and layout of the report incorporate a number of best practices from the behavioral interviewing literature. At a broader level, the Watson-Glaser II Interview Report fits with multiple different interviewing workflows because organizations approach their selection processes uniquely. For example, interview questions can be selected to fit within a single interview or across multiple interviews with different interviewers. Each page includes key information to facilitate effective interviewing:

- To help you better understand what is being assessed, the report provides detailed definitions of each dimension and guidance on what to look for in an examinee's response.
- To help you customize the interview to your needs, the report offers a choice of six different interview questions per dimension, with 18 total questions provided per report. Nine questions are standard across reports and can be asked of all examinees to enable straightforward comparisons. The other nine questions are customized based on the subscale score ranges (i.e., low, average, high as described previously) to provide you with a deeper understanding of the examinee's critical thinking skills based on his or her level of critical thinking.
- Probing questions are provided so you can gather more detailed information about the critical elements of an examinee's response (i.e., the situation, his or her behavior, and the results of that behavior).
- To help you document and score each response consistently, 5-point rating scales and note-taking space are provided for each question, and a Total Score Matrix is provided at the end of the report to facilitate aggregation across all interview questions.
- Because interviewing is a challenging skill with significant legal, organizational, and personal implications, guidance is also provided regarding the proper use of the report, tips for effective interviewing, and best practices for conducting a fair and legally-defensible interview (note that this report is subject to the same legal precautions and should incorporate the same best practices recommended for use of the Watson-Glaser II assessment in employment selection).

The Watson-Glaser II Interview Report incorporates a host of best practices to help you obtain a more reliable and valid picture of how an examinee's critical thinking skills are likely to appear on the job.

## The Watson-Glaser II Development Report

The Watson-Glaser II Development Report provides insight and specific guidance to strengthen an individual's critical thinking knowledge and skills. The report is primarily directed toward the individual (e.g., "You scored higher than most of your peers."), but managers, coaches, or other development professionals can also use it to identify areas of focus for building that individual's critical thinking skills. Additionally, the report can be used on its own or as part of a broader development process. Best practices from the training and development literature suggest that the report will be more effective when combined with other interventions such as coaching, classroom training, e-learning, and/or structured self-study (Goldstein & Ford, 2002).

It is important to note that critical thinking is, in part, a cognitive ability. As such, certain elements that facilitate effective critical thinking (e.g., working memory; reading ability) are unlikely to change through a developmental intervention. Still, the academic literature clearly demonstrates that critical thinking can be improved when development efforts focus on improving knowledge and behavioral skills (Halpern, 1998; 2003). To that end, the report includes a number of features that facilitate growth:

- The report begins with an applied example and an organizational framework (i.e., the RED Model) to build an understanding of the components of critical thinking that were assessed. In addition, it provides detailed definitions for each dimension, as well as behavioral examples of skilled and unskilled critical thinking.

- To promote an awareness of where the individual stands on each of the three dimensions, the report provides an in-depth review of the individual's assessment results for each of the three subscales, including interpretations of how his or her scores would translate into actual behaviors.
- To help individuals build their critical thinking skills, the report offers multiple, customized development suggestions grounded in the academic literature. The suggestions are based on the individual's subscale score ranges (as described previously), meaning they receive a different set of suggestions depending on whether their scores were in the "high range" (Strength to Leverage), in the "average range" (Further Exploration), or in the "low range" (Opportunity for Development).
- To enable individuals to translate the results into their day-to-day experiences, structured space is provided for them to reflect on the meaning of their results and the development suggestions that seem most useful to them. The report also provides guidance to help individuals apply knowledge of their critical thinking skills to key workplace competencies (e.g., decision making).
- To facilitate a strong development plan, the report offers guidance on how to create a realistic plan for building the individual's critical thinking skills based on best practices for development.
- The report concludes with suggestions for next steps that individuals should take to continue growing.

In total, the Watson-Glaser II Development Report offers individuals key insights, suggestions, and structured guidance to promote the growth of their critical thinking knowledge and skills.

- Adams, M.H., Stover, L.M., & Whitlow, J.F. (1999). A longitudinal evaluation of baccalaureate nursing students' critical thinking abilities. *Journal of Nursing Education, 38*, 139–141.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Americans With Disabilities Act of 1990, Titles I & V (Pub. L. 101-336). *United States Code, Volume 42*, Sections 12101–12213.
- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, N.J.: Prentice Hall.
- Behrens, P. J. (1996). The Watson-Glaser Critical Thinking Appraisal and academic performance of diploma school students. *Journal of Nursing Education, 35*, 34–36.
- Cascio, W. F., & Aguinis, H. (2005). *Applied psychology in human resource management* (6th ed.). Upper Saddle River, NJ: Prentice Hall.
- Cronbach, L. J. (1970). *Essentials of psychological testing*, third edition, New York: Harper & Row
- Deary, I. J., Whalley, L. J., Lemmon, H., Crawford, J. R., & Starr, J. M. (2000). The stability of individual differences in mental ability from childhood to old age: Follow-up of the 1932 Scottish mental survey. *Intelligence, 28*, 49–55.
- Duchesne, R. E., Jr. (1996). Critical thinking, developmental learning, and adaptive flexibility in organizational leaders (Doctoral dissertation, University of Connecticut, 1996). *Dissertation Abstracts International, 57*, 2121.
- Equal Employment Opportunity Commission. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43* (166), 38295–38309.
- Facione, P. A. (1990). *Executive summary: The Delphi report*. Millbrae, CA: California Academic Press.
- Facione, P. A. (2009). *Critical thinking: What it is and why it counts*. Millbrae, CA: Insight Assessment.
- Fischer, S. C., & Spiker, V. A. (2000). *A model of critical thinking*. Report prepared for the U.S. Army Research Institute.
- Gadzella, B. M., Baloglu, M., & Stephens, R. (2002). Prediction of GPA with educational psychology grades and critical thinking scores. *Education, 122*(3), 618–623.
- Gadzella, B. M., Ginther, D. W., & Bryant, G. W. (1996, August). *Teaching and learning critical thinking skills*. Paper presented at the XXVI International Congress of Psychology, Montreal, Quebec.
- Gadzella, B. M., Stephens, R., & Stacks, J. (2004). *Assessment of critical thinking scores in relation with psychology and GPA for education majors*. Paper presented at the Texas A & M University Assessment Conference, College Station, TX.
- Geisinger, K. F. (1998). Review of Watson-Glaser Critical Thinking Appraisal. In J. C. Impara & B. S. Plake (Eds.), *The thirteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.
- Glaser, E. M. (1937). An experiment in the development of critical thinking. *Contributions to Education, No. 843*. New York: Bureau of Publications, Teachers College, Columbia University.
- Goldstein, I. L., & Ford, J. K. (2002). *Training in organizations* (4th ed.). Belmont, CA: Wadsworth.
- Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum.
- Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist, 53*, 449-455.
- Halpern, D. F. (2003). *Thought and knowledge: An introduction to critical thinking*. Mahwah, N.J. Lawrence Erlbaum.
- Howard, P. J., & Howard, J. M. (2001). *Professional manual for the Workplace Big Five Profile (WB5P)*. Charlotte: Center for Applied Cognitive Studies.
- Impelman, K., & Graham, H. (2009). *Interactive effects of openness to experience and cognitive ability*. Paper presented at the 24th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.
- Kelloway, E. K. (1998). *Using LISREL for structural equation modeling: A researcher's guide*. Thousand Oaks, CA: Sage Publications.
- Klaczynski, P. A., Gordon, D. H., & Fauth, J. (1997). Goal-oriented critical reasoning and individual differences in critical reasoning biases. *Journal of Educational Psychology, 89*, 470–485.
- Kudish, J. D., & Hoffman, B. J. (2002, October). *Examining the relationship between assessment center final dimension ratings and external measures of cognitive ability and personality*. Paper presented at the 30th International Congress on Assessment Center Methods, Pittsburgh, PA.

- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563–575.
- McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S.D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*, 599–616.
- Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449–458.
- Myers, P. B., & Myers, K. D. (2004). *Myers-Briggs Type Indicator Profile*. Mountain View, CA: CPP, Inc.
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175–220.
- Paul, R. W., & Elder, L. (2002). *Critical thinking: Tools for taking charge of your professional and personal life*. Upper Saddle River, NJ: Financial Times Prentice Hall.
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.
- Robertson, I. T., & Molloy, K. J. (1982). Cognitive complexity, neuroticism, and research ability. *British Journal of Educational Psychology, 52*, 113–118.
- Rust, J. (2002). *Rust Advanced Numerical Reasoning Appraisal Manual*. London: The Psychological Corporation.
- Sa, W. C., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology, 91*, 497–510.
- Scott, J. N., & Markert, R. J. (1994). Relationship between critical thinking skills and success in preclinical courses. *Academic Medicine, 69*(11), 920–924.
- Shin, K. R. (1998). Critical thinking ability and clinical decision-making skills among senior nursing students in associate and baccalaureate programs in Korea. *Journal of Advanced Nursing, 27*(2), 414–418.
- Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.
- Spector, P. A., Schneider, J. R., Vance, C. A., & Hezlett, S. A. (2000). The relation of cognitive ability and personality traits to assessment center performance. *Journal of Applied Social Psychology, 30*(7), 1474–1491.
- Stanovich, K.E., & West, R.F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89*, 342–357.
- Stanovich, K.E., & West, R.F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94*, 672–695.
- Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*, 173–180.
- Taube, K. T. (1995, April). *Critical thinking ability and disposition as factors of performance on a written critical thinking test*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.
- U.S. Department of Labor. (1999). *Testing and assessment: An employer's guide to good practices*. Washington, DC: Author.
- Watson, G., & Glaser, E. M. (1952). *Watson-Glaser Critical Thinking Appraisal manual*. New York: Harcourt, Brace, & World.
- Watson, G., & Glaser, E. M. (1980). *Watson-Glaser Critical Thinking Appraisal, Forms A and B manual*. San Antonio, TX: The Psychological Corporation.
- Watson, G., & Glaser, E. M. (1994). *Watson-Glaser Critical Thinking Appraisal, Form S manual*. San Antonio, TX: The Psychological Corporation.
- Watson, G., & Glaser, E. M. (2006). *Watson-Glaser Critical Thinking Appraisal, Short Form manual*. San Antonio, TX: Pearson.
- West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology, 100*, 930–941.
- Williams, R. L. (2003). *Critical thinking as a predictor and outcome measure in a large undergraduate educational psychology course*. (Report No. TM-035-016). Knoxville, TN: University of Tennessee. (ERIC Document Reproduction Service No. ED478075)
- Yang, S. C., & Lin, W. C. (2004). The relationship among creative, critical thinking and thinking styles in Taiwan High School Students. *Journal of Instructional Psychology, 31*(1). 33–45.

## Glossary of Measurement Terms

This glossary is intended to aid in the interpretation of statistical information presented in this manual, as well as other manuals published by Pearson. The terms defined are basic. In the definitions, certain technicalities have been sacrificed for the sake of succinctness and clarity.

**average**—A general term applied to the various measures of central tendency. The three most widely used averages are the arithmetic mean (mean), the median, and the mode. When the “average” is used without designation as to type, the most likely assumption is that it is the mean. See CENTRAL TENDENCY, MEAN, MEDIAN.

**central tendency**—A measure of central tendency provides a single most typical score as representative of a group of scores. The “trend” of a group of measures is indicated by some type of average, usually the mean or the median.

**Classical Test Theory (also known as True Score Theory)**—The earliest theory of psychological measurement which is based on the idea that the observed score a person gets on a test is composed of the person's theoretical “true score” and an “error score” due to unreliability (or imperfection) in the test. In Classical Test Theory (CTT), item difficulty is indicated by the proportion ( $p$ ) of examinees that answer a given item correctly. Note that in CTT, the more difficult an item is, the lower  $p$  is for that item.

**Coefficient Alpha ( $r_{\text{alpha}}$ )**—An index to measure the internal consistency of a test by providing the mean of all possible half-splits. The extent that test items are highly intercorrelated, the coefficient alpha will yield a higher estimate of reliability. The coefficient alpha is considered a measure of the internal consistency only and is not an indication of stability over time.

**Cohen's  $d$** —An index to measure the magnitude of the actual difference between two means. The difference ( $d$ ) is calculated by dividing the difference of the two test means by the square root of the pooled variance, using Cohen's Formula 10.4.

**composite score**—A score which combines several scores, usually by addition; often different weights are applied to the contributing scores to increase or decrease their importance in the composite. Most commonly, such scores are used for predictive purposes and the weights are derived through multiple regression procedures.

**correlation**—Relationship or “going-togetherness” between two sets of scores or measures; tendency of one score to vary concomitantly with the other, as the tendency of students of high IQ to be above the average in reading ability. The existence of a strong relationship (i.e., a high correlation) between two variables does not necessarily indicate that one has any causal influence on the other. Correlations are usually denoted by a coefficient; the correlation coefficient most frequently used in test development and educational research is the Pearson or product-moment  $r$ . Unless otherwise specified, “correlation” usually refers to this coefficient. Correlation coefficients range from  $-1.00$  to  $+1.00$ ; a coefficient of  $0.0$  (zero) denotes a complete absence of relationship. Coefficients of  $-1.00$  or  $+1.00$  indicate perfect negative or positive relationships, respectively.

**criterion**—A standard by which a test may be judged or evaluated; a set of other test scores, job performance rating, etc., with which a test is designed to measure, to predict, or to correlate. See VALIDITY.

**cutoff point (cut score)**—A specified point on a score scale at or above which applicants pass the test and below which applicants fail the test.

**deviation**—The amount by which a score differs from some reference value, such as the mean, the norm, or the score on some other test.

**difficulty index ( $p$  or  $b$ )**—The proportion of examinees correctly answering an item. The greater the proportion of correct responses, the easier the item.

**discrimination index ( $d$  or  $a$ )**—The difference between the proportion of high-scoring examinees who correctly answer an item and the proportion of low-scoring examinees who correctly answer the item. The greater the difference, the more information the item has regarding the examinee's level of performance.

**distribution (frequency distribution)**—A tabulation of the scores (or other attributes) of a group of individuals to show the number (frequency) of each score, or of those within the range of each interval.

**equivalence**—Occurs when test forms measure the same construct and every level of the construct is measured with equal accuracy by the forms. Statistically equivalent test forms may be used interchangeably.

- factor analysis**—A term that represents a large number of different mathematical procedures for summarizing the interrelationships among a set of variables or items in terms of a reduced number of hypothetical variables, called factors. Factors are used to summarize scores on multiple variables in terms of a single score, and to select items that are homogeneous.
- factor loading**—An index, similar to the correlation coefficient in size and meaning, of the degree to which a variable is associated with a factor; in test construction, a number that represents the degree to which an item is related to a set of homogeneous items.
- Fit to the model**—No model can be expected to represent complex human behavior or ability perfectly. As a reasonable approximation, however, such a model can provide many practical benefits. Item-difficulty and person-ability values are initially estimated on the assumption that the model is correct. An examination of the data reveals whether or not the model satisfactorily predicts each person's actual pattern of item passes and failures. The model-fit statistic, based on discrepancies between predicted and observed item responses, identifies items that "fit the model" better. Such items are then retained in a shorter version of a long test.
- internal consistency**—Degree of relationship among the items of a test; consistency in content sampling.
- Item Response Theory (IRT)**—Refers to a variety of techniques based on the assumption that performance on an item is related to the estimated amount of the "latent trait" that the examinee possesses. IRT techniques show the measurement efficiency of an item at different ability levels. In addition to yielding mathematically refined indices of item difficulty ( $b$ ) and item discrimination ( $a$ ), IRT models may contain additional parameters (i.e., Guessing).
- mean ( $M$ )**—A kind of average usually referred to as the "mean." It is obtained by dividing the sum of a set of scores by the number of scores. See CENTRAL TENDENCY.
- median ( $Md$ )**—The middle score in a distribution or set of ranked scores; the point (score) that divides the group into two equal parts; the 50th percentile. Half of the scored are below the median and half above it, except when the median itself is one of the obtained scores. See CENTRAL TENDENCY.
- normal distribution**—A distribution of scores or measures that in graphic form has a distinctive bell-shaped appearance. In a perfect normal distribution, scores or measures are distributed symmetrically around the mean, with as many cases up to various distances above the mean as down to equal distances below it. Cases are concentrated near the mean and decrease in frequency, according to a precise mathematical equation, the farther one departs from the mean. Mean, median, and mode are identical. The assumption that mental and psychological characteristics are distributed normally has been very useful in test development work.
- normative data (norms)**—Statistics that supply a frame of reference by which meaning may be given to obtained test scores. Norms are based upon the actual performance of individuals in the norm group(s) for the test. Since they represent average or typical performance, they should not be regarded as standards or as universally desirable levels of attainment. The most common types of norms are deviation IQ, percentile rank, grade equivalent, and stanine. Reference groups are usually those of specified occupations, age, grade, gender, or ethnicity.
- norm group**—A large, representative sample that has taken the test and has similar characteristics (e.g., occupation, position level) to the examinee whose test score is being interpreted. This group is used to establish percentiles or other standardized scores to be used as a comparison for interpreting individual test scores.
- percentile (P)**—A point (score) in a distribution at or below which fall the percent of cases indicated by the percentile. Thus a score coinciding with the 35th percentile is regarded as equaling or surpassing 35% of the persons in the group, such that 65% of the performances exceed this score. "Percentile" does not mean the percent of correct answers on a test.
- Use of percentiles in interpreting scores offers a number of advantages: percentiles are easy to compute and understand, can be used with any type of examinee, and are suitable for any type of test. The primary drawback of using a raw score-to-percentile conversion is the resulting inequality of units, especially at the extremes of the distribution of scores. For example, in a normal distribution, scores cluster near the mean and decrease in frequency the farther one departs from the mean. In the transformation to percentiles, raw score differences near the center of the distribution are exaggerated—small raw score differences may lead to large percentile differences. This is especially the case when a large proportion of examinees receive same or similar scores, causing a one- or two-point raw score difference to result in a 10- or 15-unit percentile

difference. Short tests with a limited number of possible raw scores often result in a clustering of scores. The resulting effect on tables of selected percentiles is “gaps” in the table corresponding to points in the distribution where scores cluster most closely together.

percentile band—An interpretation of a test score which takes into account the measurement error that is involved. The range of such bands, most useful in portraying significant differences in battery profiles, is usually from one standard error of measurement below the obtained score to one standard error of measurement above the score.

percentile rank (PR)—The expression of an obtained test score in terms of its position within a group of 100 scores; the percentile rank of a score is the percent of scores equal to or lower than the given score in its own or some external reference group.

point-biserial correlation ( $r_{pbis}$ )—A type of correlation coefficient calculated when one variable represents a dichotomy (e.g., 0 and 1) and the other represents a continuous or multi-step scale. In test construction, the dichotomous variable is typically the score (i.e., correct or incorrect) and the other is typically the number correct for the entire test; good test items will have moderate to high positive point-biserial correlations (i.e., more high-scoring examinees answer the item correctly than low-scoring examinees).

practice effect—The influence of previous experience with a test on a later administration of the same or similar test; usually an increased familiarity with the directions, kinds of questions, etc. Practice effect is greatest when the interval between test events is short, when the content of the two tests is identical or very similar, and when the initial test-taking represents a relatively novel experience for the subjects.

profile—A graphic representation of the results on several tests or subscales, for either an individual or a group, when the results have been expressed in some uniform or comparable terms (standard scores, percentile ranks, grade equivalents, etc.). The profile method of presentation permits identification of area of strength or weakness.

$r$ —See Correlation.

range—For some specified group, the difference between the highest and the lowest obtained score on a test; thus a very rough measure of spread or variability, since it is based upon only two extreme scores. Range is also used in reference to the possible range of scores on a test, which in most instances is the number of items in the test.

Rasch model—A technique in Item Response Theory (IRT) using only the item difficulty parameter. This model assumes that both guessing and item differences in discrimination are negligible.

raw score—The first quantitative result obtained in scoring a test. Examples include the number of right answers, number right minus some fraction of number wrong, time required for performance, number of errors, or similar direct, unconverted measures.

reliability—The extent to which a test is consistent in measuring whatever it does measure; dependability, stability, trustworthiness, relative freedom from errors of measurement. Reliability is usually expressed by some form of reliability coefficient or by the standard error of measurement derived from it.

reliability coefficient—The coefficient of correlation between two forms of a test, between scores on two administrations of the same test, or between halves of a test, properly corrected. The reliability coefficient is a measure of the stability or internal consistency of a test.

representative sample—subset that corresponds to or matches the population of which it is a sample with respect to characteristics important for the purposes under investigation. In a clerical aptitude test norm sample, such significant aspects might be the level of clerical training and work experience of those in the sample, the type of job they hold, and the geographic location of the sample.

skewness—The extent to which the curve of a frequency distribution departs from perfect symmetry. Skewness is described as positive when the tail of the distribution extends to the right, and negative when the tail of the distribution extends to the left.

split-half reliability coefficient ( $r_{split}$ )—A coefficient of reliability obtained by correlating scores on one half of a test with scores on the other half, and applying the Spearman-Brown formula to adjust for the double length of the total test. Generally, but not necessarily, the two halves consist of the odd-numbered and the even-numbered items. Split-half reliability coefficients are sometimes referred to as measures of the internal consistency of a test; they involve content sampling only, not stability over time.

standard deviation ( $SD$ )—A measure of the variability or dispersion of a distribution of scores. The more the scores cluster around the mean, the smaller the standard deviation. For a normal distribution, approximately two thirds (68.25%) of the scores are within the range from one  $SD$  below the mean to one  $SD$  above the mean. Computation of the  $SD$  is based upon the square of the deviation of each score from the mean.

standard error (*SE*)—A statistic providing an estimate of the possible magnitude of “error” present in some obtained measure, whether (1) an individual score or (2) some group measure, as a mean or a correlation coefficient.

(1) standard error of measurement (*SEM*)—As applied to a single obtained score, the amount by which the score may differ from the hypothetical true score due to errors of measurement. The larger the *SEM*, the less reliable the measurement and the less reliable the score. The *SEM* is an amount such that in about two-thirds of the cases, the obtained score would not differ by more than one *SEM* from the true score. (Theoretically, then, it can be said that the chances are 2:1 that the actual score is within a band extending from the true score minus one *SEM* to the true score plus one *SEM*; but since the true score can never be known, actual practice must reverse the true-obtained relation for an interpretation.) Other probabilities are noted under (2) below. See TRUE SCORE.

(2) standard error—When applied to sample estimates (e.g., group averages, standard deviation, correlation coefficients), the *SE* provides an estimate of the “error” which may be involved. The sample or group size and the *SD* are the factors on which standard errors are based. The same probability interpretation is made for the *SEs* of group measures as is made for the *SEM*; that is, 2 out of 3 sample estimates will lie within 1.0 *SE* of the “true” value, 95 out of 100 within 1.96 *SE*, and 99 out of 100 within 2.6 *SE*.

standard score—A general term referring to any of a variety of “transformed” scores, in terms of which raw scores may be expressed for reasons of convenience, comparability, ease of interpretation, etc. The simplest type of standard score, known as a *z* score, is an expression of the deviation of a score from the mean score of the group in relation to the standard deviation of the scores of the group. Thus,

$$\text{Standard Score} = (\text{Score} - \text{Mean}) / \text{Standard Deviation}$$

Adjustments may be made in this ratio so that a system of standard scores having any desired mean and standard deviation may be set up. The use of such standard scores does not affect the relative standing of the individuals in the group or change the shape of the original distribution.

Standard scores are useful in expressing the raw score of two forms of a test in comparable terms in instances where tryouts have shown that the two forms are not identical in difficulty. Also, successive levels of a test may be linked to Form A continuous standard-score scale, making across-battery comparisons possible.

standardized test—A test designed to provide a systematic sample of individual performance, administered according to prescribed directions, scored in conformance with definite rules, and interpreted in reference to certain normative information. Some would further restrict the usage of the term “standardized” to those tests for which the items have been chosen on the basis of experimental evaluation, and for which data on reliability and validity are provided.

testlet—A single test scenario that has a number of test questions based directly on the scenario. A testlet score is generated by summing the responses for all items in the testlet.

test-retest reliability coefficient—A type of reliability coefficient obtained by administering the same test a second time, after a short interval, and correlating the two sets of scores. “Same test” was originally understood to mean identical content, i.e., the same form. Currently, however, the term “test-retest” is also used to describe the administration of different forms of the same test, in which case this reliability coefficient becomes the same as the alternate-form coefficient. In either type, the correlation may be affected by fluctuations over time, differences in testing situations, and practice. When the time interval between the two testings is considerable (i.e., several months), a test-retest reliability coefficient reflects not only the consistency of measurement provided by the test, but also the stability of the trait being measured.

true score—A score entirely free of error; hence, a hypothetical value that can never be obtained by psychological testing, because testing always involves some measurement error. A “true” score may be thought of as the average score from an infinite number of measurements from the same or exactly equivalent tests, assuming no practice effect or change in the examinee during the test events. The standard deviation of this infinite number of “samplings” is known as the standard error of measurement.

validity—The extent to which a test does the job for which it is used. This definition is more satisfactory than the traditional “extent to which a test measures what it is supposed to measure,” since the validity of a test is always specific to the purposes for which the test is used.

(1) content validity. For achievement tests, validity is the extent to which the content of the test represents a balanced and adequate sampling of the outcomes (knowledge, skills, etc.) of the job, course, or instructional program it is intended to cover. It is best evidenced by a comparison of the test content with job descriptions, courses of study, instructional materials, and statements of educational goals; and often by analysis of the process required in making correct responses to the items. Face validity, referring to an observation of what a test appears to measure, is a non-technical type of evidence; apparent relevancy is, however, quite desirable.

(2) construct validity. The extent to which the test measures the construct intended to be measured. No one study can determine construct validity. Rather, it is supported by multiple sources of evidence, including results from convergent and discriminate studies.

(3) criterion-related validity. The extent to which scores on the test are in agreement with (concurrent validity) or predict (predictive validity) some given criterion measure. Predictive validity refers to the accuracy with which an aptitude, prognostic, or readiness test indicates future success in some area, as evidenced by correlations between scores on the test and future criterion measures of such success (e.g., the relation of the score on a clerical aptitude test administered at the application phase to job performance ratings obtained after a year of employment). In concurrent validity, no significant time interval elapses between administration of the test and collection of the criterion measure. Such validity might be evidenced by concurrent measures of academic ability and of achievement, by the relation of a new test to one generally accepted as or known to be valid, or by the correlation between scores on a test and criteria measures which are valid but are less objective and more time-consuming to obtain than a test score.

(4) evidence based on internal structure. The extent to which a test measures some relatively abstract psychological trait or construct; applicable in evaluating the validity of tests that have been constructed on the basis of analysis (often factor analysis) of the nature of the trait and its manifestations.

(5) convergent and discriminate validity. Tests of personality, verbal ability, mechanical aptitude, critical thinking, etc., are validated in terms of the relation of their scores to pertinent external data. Convergent evidence refers to the relationship between a test score and other measures that have been demonstrated to measure similar constructs. Discriminate evidence refers to the relationship between a test score and other measures demonstrated to measure dissimilar constructs.

variability—The spread or dispersion of test scores, best indicated by their standard deviation.

variance—For a distribution, the variance is the average of the squared deviations from the mean. Thus, the variance is the square of the standard deviation.