



Watson-Glaser Critical Thinking Appraisal®

**Short Form
Manual**

Goodwin Watson & Edward M. Glaser





Copyright © 2008 by NCS Pearson, Inc.

All rights reserved. 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.

The **Pearson** and **TalentLens** logos, and **Watson-Glaser Critical Thinking Appraisal** 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.

Table of Contents

Chapter 1

Introduction	1
---------------------------	---

Chapter 2

Critical Thinking Ability and the Development of the Original Watson-Glaser Forms	3
The Watson-Glaser Short Form	4

Chapter 3

Directions for Paper-and-Pencil Administration and Scoring	5
Preparing for Administration	5
Testing Conditions	5
Materials Needed to Administer the Test	5
Answering Questions	6
Administering the Test	6
Timed Administration	7
Untimed Administration	8
Concluding Administration	8
Scoring	8
Scoring with the Hand-Scoring Key	8
Machine Scoring	8
Test Security	9
Accommodating Examinees with Disabilities	9

Chapter 4

Directions for Computer-Based Administration	11
Preparing for Administration	11
Testing Conditions	11
Answering Questions	11
Administering the Test	12
Scoring and Reporting	12
Test Security	12
Accommodating Examinees with Disabilities	13

Chapter 5

Norms	15
Using Norms Tables to Interpret Scores.....	15
Converting Raw Scores to Percentile Ranks.....	16

Chapter 6

Development of the Short Form	19
Test Assembly Data Set.....	19
Criteria for Item Selection.....	20
Maintenance of Reading Level.....	21
Updates to the Test.....	21
Test Administration Time.....	21

Chapter 7

Equivalence of Forms	23
Equivalence of Short Form to Form A.....	23
Equivalent Raw Scores.....	24
Equivalence of Computer-Based and Paper-and-Pencil Versions of the Short Form.....	25

Chapter 8

Evidence of Reliability	27
Historical Reliability.....	28
Previous Studies of Internal Consistency Reliability.....	28
Previous Studies of Test-Retest Reliability.....	29
Current Reliability Studies.....	29
Evidence of Internal Consistency Reliability.....	29
Evidence of Test-Retest Reliability.....	30

Chapter 9

Evidence of Validity	33
Evidence of Validity Based on Content.....	33
Evidence of Criterion-Related Validity.....	34
Previous Studies of Evidence of Criterion-Related Validity.....	35
Current Studies of Evidence of Criterion-Related Validity.....	35
Evidence of Convergent and Discriminant Validity.....	39
Previous Studies of Evidence of Convergent and Discriminant Validity.....	39
Studies of the Relationship Between the Watson-Glaser and General Intelligence.....	40
Current Studies of Evidence of Convergent and Discriminant Validity.....	40

Chapter 10

Using the Watson-Glaser as an Employment Selection Tool	43
Employment Selection.....	43
Fairness in Selection Testing.....	44
Legal Considerations.....	44

Group Differences/Adverse Impact.....	44
Monitoring the Selection System.....	44
Research	45
Appendix A	
Description of the Normative Sample and Percentile Ranks	47
Appendix B	
Final Item Statistics for the Watson-Glaser–Short Form Three-Parameter IRT Model	63
References	65
Research Bibliography	69
Glossary of Measurement Terms.....	79
Tables	
6.1 Distribution of Item Development Sample Form A Scores ($N = 1,608$)	19
6.2 Grade Levels of Words on the Watson-Glaser–Short Form	21
6.3 Frequency Distribution of Testing Time in Test-Retest Sample ($n = 42$)	22
7.1 Part-Whole Correlations (r_{pw}) of the Short Form and Form A	24
7.2 Raw Score Equivalencies Between the Short Form and Form A	25
7.3 Equivalency of Paper and Online Modes of Administration	26
8.1 Means, Standard Deviations (SD), Standard Errors of Measurement (SEM) and Internal Consistency Reliability Coefficients (r_{α}) for the Short Form Based on Previous Studies	28
8.2 Means, Standard Deviations (SD), Standard Errors of Measurement (SEM) and Internal Consistency Reliability Coefficients (r_{α}) for the Current Short Form Norm Groups	30
8.3 Test-Retest Reliability of the Short Form	31
9.1 Studies Showing Evidence of Criterion-Related Validity.....	37
9.2 Watson-Glaser Convergent Evidence of Validity.....	41
A.1 Description of the Normative Sample by Industry	47
A.2 Description of the Normative Sample by Occupation	51
A.3 Description of the Normative Sample by Position Type/Level	54
A.4 Percentile Ranks of Total Raw Scores for Industry Groups.....	57
A.5 Percentile Ranks of Total Raw Scores for Occupations	59
A.6 Percentile Ranks of Total Raw Scores for Position Type/Level.....	60
A.7 Percentile Ranks of Total Raw Scores for Position Type/ Occupation Within Industry	61
B.1 Final Item Statistics for the Watson-Glaser Short Form Three-Parameter IRT Model (reprinted from Watson & Glaser, 1994)	63

Acknowledgements

The development and publication of updated information on a test like the *Watson-Glaser Critical Thinking Appraisal*®–*Short Form* inevitably involves the helpful participation of many people in several phases of the project—design, data collection, statistical data analyses, editing, and publication. The Harcourt Assessment Talent Assessment team is indebted to the numerous professionals and organizations that provided assistance.

The Talent Assessment team thanks Julia Kearney, Sampling Special Projects Coordinator; Terri Garrard, Study Manager; and Victoria N. Locke, Director, Catalog Sampling Department, for coordinating the data collection phase of this project. David Quintero, Clinical Handscoring Supervisor, ensured accurate scoring of the paper-administered test data.

We thank Zhiming Yang, PhD, Psychometrician, and Jianjun Zhu, PhD, Manager, Data Analysis Operations. Zhiming's technical expertise in analyzing the data and Jianjun's psychometric leadership ensured the high level of analytical rigor and psychometric integrity of the results reported.

Our thanks also go to Troy Beehler and Peter Schill, Project Managers, for skillfully managing the logistics of this project. Troy and Peter worked with several team members from the Technology Products Group, Harcourt Assessment, Inc. to ensure the high quality and accuracy of the computer interface. These dedicated individuals included Paula Oles, Manager, Software Quality Assurance; Matt Morris, Manager, System Development; Christina McCumber and Johnny Jackson, Software Quality Assurance Analysts; Terrill Freese, Requirements Analyst; and Maurya Duran, Technical Writer. Dawn Dunleavy, Senior Managing Editor and Konstantin Tikhonov, Project Editor, provided editorial guidance and support. Production assistance was provided by Stephanie Adams, Director, Production; Mark Cooley, Designer; Debbie Glaeser, Production Coordinator; and Robin Espiritu, Production Manager, Manufacturing.

Finally, we wish to acknowledge the leadership, guidance, support, and commitment of the following people through all the phases of this project: Gene Bowles, Vice President, Publishing and Technology, Larry Weiss, PhD, Vice President, Psychological Assessment Products Group, and Aurelio Prifitera, PhD, Publisher, Harcourt Assessment, Inc., and President, Harcourt Assessment International.

Kingsley C. Ejiogu, PhD, Research Director

Mark Rose, PhD, Research Director

John Trent, M.S., Senior Research Analyst

The *Watson-Glaser Critical Thinking Appraisal*[®] (subsequently referred to in this manual as the Watson-Glaser) is designed to measure important abilities involved in critical thinking. Critical thinking ability plays a vital role in academic instruction and occupations that require careful analytical thinking to perform essential job functions. The Watson-Glaser has been used to predict performance in a variety of educational settings and has been a popular selection tool for executive, managerial, supervisory, administrative, and technical occupations for many years. When used in conjunction with information from multiple sources about the examinee's skills, abilities, and potential for success, the Watson-Glaser can contribute significantly to the quality of an organization's selection program.

The Watson-Glaser–Short Form was published in 1994 to enhance the use of the Watson-Glaser in assessing adult employment applicants, candidates for employment-related training, career and vocational counselees, college students, and students in technical schools and adult education programs. As an abbreviated version of the Watson-Glaser–Form A, the Short Form uses a subset of Form A scenarios and items to measure the same critical thinking abilities.

This manual provides the following information about the Short Form:

- **Updated guidelines for administration.** Chapter 3 includes guidelines for administering and scoring the traditional paper-and-pencil version. Chapter 4 provides guidelines for administering the new computer-based version.
- **Updated normative information (norms).** Twenty-three new norm groups, based on 6,713 cases collected in 2004 and 2005, are presented in chapter 5.
- **Results of an equivalency study on the computer-based and paper-and-pencil versions.** To ensure equivalence between the newly designed computer-based and traditional paper-and-pencil based formats of the Watson-Glaser, a study was conducted comparing scores on the two versions. A full description of the study, which supported equivalence of the two versions, is presented in chapter 7, Equivalence of Forms.
- **Updated reliability and validity information.** New studies describing internal consistency and test-retest reliability are presented in chapter 8. New studies describing convergent and criterion-related validity are presented in chapter 9.

Information on Forms A and B was published in the 1994 Watson-Glaser manual.

Critical Thinking Ability and the Development of the Original Watson-Glaser Forms

2

Development of the Watson-Glaser was driven by the conceptualization of critical thinking as a combination of attitudes, knowledge, and skills. This conceptualization suggests that critical thinking includes:

- the 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.

The precursors of the Watson-Glaser were developed by Goodwin Watson in 1925 and Edward Glaser in 1937. These tests were developed with careful consideration given to the theoretical concept of critical thinking, as well as practical applications. In 1964, The Psychological Corporation (now Harcourt Assessment, Inc.) published Watson-Glaser Forms Ym and Zm. Each form contained 100 items and replaced an earlier version of the test, Form Am. In 1980, Form Ym and Form Zm were modified for clarity, current word usage, and the elimination of racial and sexual stereotypes. The revised instruments, each containing 80 items, were published as Form A and Form B.

The Watson-Glaser measures the extent to which examinees need training or have mastered certain critical thinking skills. The availability of comparable forms (i.e., the Short Form, Form A, and Form B) makes it possible to partially gauge the efficacy of instructional programs, and to measure developments of these skills over an extended period of time. The Watson-Glaser also has been a particularly popular tool for assessing the success of critical thinking instruction programs and courses, and for placing students in gifted and talented programs at the high school level, and in honors curriculum at the university level.

The Watson-Glaser is composed of a set of five tests. Each test is designed to tap a somewhat different aspect of critical thinking. A high level of competency in critical thinking, as measured by the Watson-Glaser, may be operationally defined as the ability to correctly perform the domain of tasks represented by the five tests.

1—Inference. Discriminating among degrees of truth or falsity of inferences drawn from given data.

2—Recognition of Assumptions. Recognizing unstated assumptions or presuppositions in given statements or assertions.

3—Deduction. Determining whether certain conclusions necessarily follow from information in given statements or premises.

4—Interpretation. Weighing evidence and deciding if generalizations or conclusions based on the given data are warranted.

5—Evaluation of Arguments. Distinguishing between arguments that are strong and relevant and those that are weak or irrelevant to a particular issue.

Each test 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 examinee responds. There are two types of 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 and scientific facts or experiments. Scenarios and items having controversial content refer to political, economic, and social issues that frequently provoke strong emotional responses. As noted in the research literature about critical thinking, strong attitudes, opinions, and biases affect the ability of some people to think critically (Jaeger & Freijo, 1975; Jones & Cook, 1975; Mitchell & Byrne, 1973; Sherif, Sherif, & Nebergall, 1965).

Though the Watson-Glaser comprises five tests, it is the total score of these tests that yields a reliable measure of critical thinking ability. Individually, the tests are composed of relatively few items and lack sufficient reliability to measure specific aspects of critical thinking ability. Therefore, individual test scores should not be relied upon for most applications of the Watson-Glaser.

The Watson-Glaser Short Form

The Short Form was designed to offer a brief version of the Watson-Glaser without changing the essential nature of the constructs measured. The length of time required to administer Form A or Form B of the Watson-Glaser is approximately one hour, making both forms well suited for administration during a single classroom period in a school setting. However, such lengthy administration time increases the cost and decreases the practicality of using the Watson-Glaser in adult assessment, particularly in the employment selection context.

The Short Form is composed of 16 scenarios and 40 items selected from the 80-item Form A. The Short Form takes about 30 minutes to complete in a paper-and-pencil or computer-based format. It takes an additional five to ten minutes to read the directions and sample questions. At one-half the length of Form A, the Short Form presents a more practical measure of critical thinking ability, yet retains an equivalent nature (see chapter 7, Equivalence of Forms). Organizations requiring an alternative to the Short Form for retesting or other purposes may use the full length Form B.

Like Form A and Form B, the Short Form is appropriate for use with persons who have at least the equivalent of a ninth-grade education (see chapter 6, Development of the Short Form, for more information on the reading level of the Watson-Glaser).

Directions for Paper-and-Pencil Administration and Scoring

3

Preparing for Administration

The person responsible for administering the Watson-Glaser does not need special training, but must be able to carry out standard examination procedures. To ensure accurate and reliable results, the administrator must become thoroughly familiar with the administration instructions and the test materials before attempting to administer the test. It is recommended for test administrators to take the Watson-Glaser prior to administration, being sure to comply with the directions and any time requirement.

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.

Materials Needed to Administer the Test

- This manual or the Directions for Administration booklet
- 1 Test Booklet for each examinee
- 1 Answer Document 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 machine-scored)

Intended as a test of critical thinking power rather than speed, the Watson-Glaser may be given in either timed or untimed administrations. In timed administrations, the time limit is based on the amount of time required to finish the test by the majority of examinees in test tryouts. The administrator should have a regular watch with a second hand, a wall clock with sweep-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 and answer questions.

Answering Questions

Examinees may ask questions about the test before the signal to begin is given. 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 Document 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, questions about the test directions should be handled by telling the examinee to do his or her best.

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 again.

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

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

Point to the *Correct Mark* and *Incorrect Marks* samples on the Answer Document.

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 Short Form Answer Document so that examinees do not skip anything or put the correct information in the wrong place.

Say **In the upper left corner of the Answer Document, 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 Document 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. If optional information is collected, the test administrator should explain to the examinees 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 Document only. Do not make any additional marks on the Answer Document until I tell you to do so.**

Distribute the Test Booklets.

Say **In this test, all the questions are in the Test Booklets. There are five separate tests 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 Document. 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 Document. 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. If you finish all the tests before time is up, you may go back and check your answers.

Timed Administration

Say, **You will have 30 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 30 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 Document 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 Documents, and pencils. Place the completed Answer Documents 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 Answer Document may be hand scored with the Hand Scoring Key or machine scored.

Scoring With the Hand-Scoring Key

First, cross out multiple responses to the same item with a heavy red mark that will show through the key. (Note: Red marks are only suitable for hand-scored documents.) Check for any answer spaces that were only partially erased by the examinee in changing an answer; partial erasures should be completely erased.

Next, place the scoring key over the Answer Document so that the edges are neatly aligned and the two stars appear through the two holes that are the closest to the bottom of the key. Count the number of correctly marked spaces (other than those through which a red line has been drawn) appearing through the holes in the stencil. Record the total in the "Score" box on the Answer Document. The maximum raw score for the Short Form is 40. The percentile score corresponding to the raw score may be recorded in the "Percentile" space on the Answer Document, and the norm group used to determine that percentile may be recorded in the space labeled "Norms Used."

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 Documents available for the Short Form may be processed by any reflective scanning device programmed to your specifications.

Test Security

Watson-Glaser scores 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 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.

Accommodating Examinees with Disabilities

You will need to routinely provide reasonable accommodations that make it possible for candidates with particular needs to comfortably take the test, such as left-handed desks for some candidates and adequate and comfortable seating for all individuals.

On occasion, a special administration may be required for an examinee with an impairment that affects his or her ability to take a test in the standard manner. Harcourt Assessment, Inc. recommends that reasonable accommodations for these examinees be made in accordance with the Americans with Disabilities Act (ADA) of 1990. The ADA has established basic legal rights for individuals with physical or mental disabilities that substantially limit one or more major life activities. Reasonable accommodations may include, but are not limited to, modifications to the testing environment (e.g., high desks), medium (e.g., having a reader read questions to the examinee), time limit, and/or content (Society for Industrial and Organizational Psychology, 2003).

If an examinee's disability is not likely to impair job performance, but may hinder his or her performance on the Watson-Glaser, you may want to consider waiving administration of the Watson-Glaser or de-emphasizing the test score in lieu of other application criteria.

Directions for Computer-Based Administration **4**

The computer-based Watson-Glaser is administered through eAssessTalent.com, the Internet-based testing system designed by Harcourt Assessment, Inc., for the administration, scoring, and reporting of professional assessments. Instructions for administrators on how to order and access the test online are provided at eAssessTalent.com. Instructions for accessing the Watson-Glaser interpretive reports are provided on the website. After a candidate has taken the Watson-Glaser on eAssessTalent.com, you can review the candidate's results in an interpretive report, using the link that Harcourt Assessment provides.

Preparing for Administration

Being thoroughly prepared before the examinee's arrival will result in a more efficient online administration session. It is recommended for test administrators to take the computer-based Watson-Glaser prior to administering the test, being sure to comply with the directions and any time requirement. Examinees will not need pencils or scratch paper for this computer-based test. In addition, examinees should not 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

Examinees may ask questions about the test before the signal to begin is given. To maintain standard testing conditions, answer such questions by rereading the appropriate section of these directions. Do not volunteer new explanations or examples. As the test administrator, it is your responsibility to ensure that examinees understand the correct way to indicate their answers 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, questions about the test directions should be handled by telling the examinee to do his or her best.

Administering the Test

After the initial instruction screen for the Watson-Glaser 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 clicks the “Start Your Test” button, test administration begins with the first page of test questions. The examinee 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 30 minutes.

If an examinee’s computer develops technical problems during testing, move the examinee to another suitable computer location. If the technical problems cannot be solved by moving to another computer location, contact Harcourt Assessment, Inc. Technical Support for assistance. The contact information, including phone and fax numbers, can be found at the eAssessTalent.com website.

Scoring and Reporting

Scoring is automatic, and the report is available a few seconds after the test is completed. A link to the report will be available on eAssessTalent.com. Adobe® Acrobat Reader® is necessary to open the report. You may view, print, or save the candidate’s report.

Test Security

Watson-Glaser scores 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 access to individuals who do not have a legitimate need for the information. Storing test scores in a locked cabinet or password protected file that can only be accessed by designated test administrators will help ensure their security. The security of testing materials (e.g., access to online tests) and protection of copyright must also be maintained by authorized individuals. Avoid disclosure of test access information such as usernames or passwords and only administer the Watson-Glaser in proctored environments. All the computer stations used in administering the Watson-Glaser must be in locations that can be easily supervised with the same level of security as with the paper-and-pencil administration.

Accommodating Examinees with Disabilities

As noted in chapter 3 above under the section dealing with examinees with disability, the test administrator should provide reasonable accommodations to enable candidates with special needs to comfortably take the test. Reasonable accommodations may include, but are not limited to, modifications to the test environment (e.g., high desks) and medium (e.g., having a reader read questions to the examinee, or increasing the font size of questions) (Society for Industrial and Organizational Psychology, 2003). In situations where an examinee's disability is not likely to impair his or her job performance, but may hinder the examinee's performance on the Watson-Glaser, the organization may want to consider waiving the test or de-emphasizing the score in lieu of other application criteria. Interpretive data as to whether scores on the Watson-Glaser are comparable for examinees who are provided reasonable accommodations are not available at this time due to the small number of examinees who have requested such accommodations.

If, due to some particular impairment, a candidate cannot take the computer-administered test but can take the test on paper, the administrator could provide reasonable accommodation for the candidate to take the test on paper, and then have the candidate's certified responses and results entered into the computer system. 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.

The raw score on the Watson-Glaser–Short Form is calculated by adding the total number of correct responses. The maximum raw score is 40. 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 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. 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* or *standardization sample*.

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 selecting from among applicants to fill a particular job, normative information derived from a specific, relevant, well-defined group may be most useful. However, the composition of a sample of job applicants is influenced by a variety of situational factors, including job demands and local labor market conditions. Because such factors can vary across jobs, locations, and over time, the limitations on the usefulness of any set of published norms should be acknowledged.

When a test is used to help make human resource decisions, the most appropriate 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, or employees. One of the factors that must be considered in preparing norms is sample size. With large samples, all possible scores can be converted to percentile ranks. Data from smaller samples tend to be unstable and the presentation of percentile ranks for all possible scores presents an unwarranted impression of precision. Until a sufficient and representative number of cases has been collected (preferably 100 or more), the norms presented in Appendix A should be used to guide the interpretation of test scores.

Using Norms Tables to Interpret Scores

The Short Form norms in Appendix A were derived from new data, collected June 2004 through March 2005, from 6,713 adults in a variety of employment settings. Please note that the distributions of occupational levels across industry

samples vary. Therefore, it is not appropriate to compare industry means presented in Appendix A to each other. The tables in Appendix A show the total raw scores on the Watson-Glaser with their corresponding percentile ranks for identified norm groups.

When using the norms tables in Appendix A, look for a group that is similar to the individual or group tested. For example, you would compare the test score of a person who applied for an engineer's position with norms derived from the scores of other engineers. If a person applied for a management position, you would compare the candidate's test score with norms for managers, or norms for managers in manufacturing. When using the norms tables in Appendix A to interpret candidates' scores, keep in mind that norms are affected by the composition of the groups that participated in the normative study. Therefore, it is important to examine specific industry and occupational characteristics of a norm group.

By comparing an individual's raw score to the data in a norms table, it is possible to determine the percentile rank corresponding to that score. The percentile rank indicates an individual's relative position in the norm group. Percentiles should not be confused with percentage scores which represent the percentage of correct items. Percentiles are derived scores which are expressed in terms of the percent of people in the norm group scoring equal to or below a given raw score.

Although percentiles 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. These factors should be taken into consideration when interpreting percentiles.

Converting Raw Scores to Percentile Ranks

To find the percentile rank of a candidate's raw score, locate the raw score from either the extreme right- or left-hand column in Tables A.4–A.7. The corresponding percentile rank is read from the selected norm group column. For example, if a person applying for a job as an engineer had a score of 35 on the Watson-Glaser–Short Form, it is appropriate to use the Engineer norms in Appendix A (Table A.5) for comparison. In this case, the percentile rank corresponding to a raw score of 35 is 63. This percentile rank indicates that about 63% of the people in the norm group scored lower than or equal to a score of 35 on the Watson-Glaser–Short Form, and about 37% scored higher than a score of 35 on the Watson-Glaser–Short Form.

Each group's size (N), mean, and standard deviation (SD) are shown at the bottom of the norms tables. The group mean or average is calculated by summing the raw scores and dividing the sum by the total number of examinees. The standard deviation indicates the amount of variation in a group of scores. In a normal distribution, approximately two-thirds (68.26%) of the scores are within the range of $-1 SD$ (below the mean) to $+1 SD$ (above the mean). These statistics are often used in describing a study sample and setting cut scores. For example, a cut score may be set as one SD below the mean.

In accordance with the Civil Rights Act of 1991, Title 1, Section 106, the norms provided in Appendix A combine data for males and females, and for white and minority examinees. The use of combined group norms can exacerbate adverse impact if there are expected differences in scores due to differences in group membership. Previous investigations conducted during the development of Watson-Glaser Form A and Form B found no consistent differences between the scores of male examinees and the scores of female examinees. Other studies of earlier Forms Ym and Zm also found no consistent differences based on the sex of the examinee in critical thinking ability as measured by the Watson-Glaser (e.g., Burns, 1974; Gurfein, 1977; Simon & Ward, 1974).

The Watson-Glaser–Short Form is a shortened version of Form A. Historical and test development information for Form A may be found in the Watson-Glaser, Forms A and B Manual, 1980 edition.

Test Assembly Data Set

Two overlapping sets of data were used in the development of the Short Form. The first data set consisted of item-level responses to Form A from 1,608 applicants and employees. These data were obtained from eight sources between 1989 and 1992. This data set was used to generate item statistics and to make decisions about item selection for inclusion in the Short Form. The average Form A score for this sample was 61.78 ($SD = 9.30$) and the internal consistency (i.e., KR-20) coefficient was .87. Table 6.1 presents a frequency distribution of Form A scores for the sample.

Table 6.1 Distribution of Item Development Sample Form A Scores ($N = 1,608$)

Form A Score	Frequency	Percent	Form A Score	Frequency	Percent ¹
1	0	0.0	41	3	0.2
2	0	0.0	42	8	0.5
3	1	0.1	43	8	0.5
4	0	0.0	44	13	0.8
5	0	0.0	45	19	1.2
6	0	0.0	46	15	0.9
7	0	0.0	47	14	0.9
8	0	0.0	48	17	1.1
9	0	0.0	49	31	1.9
10	0	0.0	50	30	1.9
11	0	0.0	51	25	1.6
12	0	0.0	52	37	2.3
13	1	0.1	53	43	2.7
14	0	0.0	54	45	2.8
15	0	0.0	55	38	2.4
16	0	0.0	56	36	2.2
17	0	0.0	57	46	2.9
18	0	0.0	58	46	2.9
19	0	0.0	59	67	4.2
20	0	0.0	60	56	3.5

(continued)

Table 6.1 Distribution of Item Development Sample Form A Scores ($N = 1,608$) (continued)

Form A Score	Frequency	Percent	Form A Score	Frequency	Percent ¹
21	1	0.1	61	57	3.5
22	0	0.0	62	66	4.1
23	1	0.1	63	59	3.7
24	0	0.0	64	79	4.9
25	0	0.0	65	63	3.9
26	1	0.1	66	74	4.6
27	0	0.0	67	76	4.7
28	0	0.0	68	76	4.7
29	2	0.1	69	68	4.2
30	4	0.2	70	88	5.5
31	1	0.1	71	60	3.7
32	2	0.1	72	62	3.9
33	0	0.0	73	39	2.4
34	1	0.1	74	37	2.3
35	3	0.2	75	33	2.1
36	2	0.1	76	22	1.4
37	5	0.3	77	8	0.5
38	3	0.2	78	6	0.4
39	3	0.2	79	2	0.1
40	5	0.3	80	0	0.0

¹ The total percent equals 100.4 due to rounding.

A second set of data was created by combining the first dataset with item-level data obtained in 1993 from three additional sources of Form A data ($N = 2,119$). The combined data set ($N = 3,727$) was used to evaluate the psychometric properties of the Short Form, including reliability, and to examine the equivalency between the Short Form and Form A.

Criteria for Item Selection

In assembling the Short Form, the primary goal was to significantly reduce the time limit required for Form A without changing the essential nature of the constructs measured. For additional information regarding the selection of items for the Short Form, please refer to the 1994 edition of the Watson-Glaser manual. The following criteria were used to select Short Form items:

- Maintenance of the Watson-Glaser five sub-test structure and the scenario-based format

- Items represent psychometrically sound scenarios and items

- Maintenance of test reliability

- Maintenance of reading level

- Update of test currency

Maintenance of Reading Level

The reading level of the shortened test form was assessed using *EDL Core Vocabulary in Reading, Mathematics, Science, and Social Studies* (Taylor, et al., 1989) and *Basic Reading Vocabularies* (Harris & Jacobson, 1982). Approximately 98.2% of words appearing in directions and exercises were at or below ninth-grade reading level. A summary of word distribution by grade level is presented in Table 6.2.

Table 6.2 Grade Levels of Words on the Watson-Glaser–Short Form

Grade Level	Frequency	Percent
Preprimer	106	5.8
1	224	12.2
2	310	16.9
3	379	20.7
4	222	12.1
5	191	10.4
6	229	12.5
7	80	4.4
8	53	2.9
9	5	0.3
10	5	0.3
11	25	1.5
Total	1829	100.0

Updates to the Test

During the test assembly process, attention was given to the currency of Form A scenarios and items. Some scenarios deal with dated subject matter, such as Russia prior to the dissolution of the USSR. The item selection process removed such dated scenarios, thereby making the composition of the Short Form more contemporary.

Test Administration Time

The optional 30-minute time limit for the Short Form was established during a study investigating the instrument's test-retest reliability. A sample of 42 employees (92.9% non-minority; 54.8% female) at a large publishing company completed the Short Form twice (two-week testing interval). The participants worked in a variety of positions ranging from Secretary to Project Director. During the first testing session, participants were given as much time as they required to complete the test. A frequency distribution of time taken (see Table 6.3) revealed that approximately 90% of the respondents completed the Short Form in 30 minutes or less. Consistent with the method used to establish testing time limits for previous forms of the Watson-Glaser, these results were used to set the time limit for

completing the Short Form at 30 minutes. The fact that the majority of respondents complete the Short Form within the allotted 30-minute time supports the point that the Watson-Glaser is a test of critical thinking power, rather than speed. Furthermore, normative data gathered in both timed and untimed administration may be used to interpret Short Form results, as the variability in scores is derived from test items rather than testing time limits.

Table 6.3 Frequency Distribution of Testing Time in Test-Retest Sample ($n = 42$)

Completion Time	Frequency	Percent	Cumulative Percent
20 minutes or less	2	4.8	4.8
21 to 25 minutes	14	33.3	38.1
26 to 30 minutes	22	52.4	90.5
31 minutes or more	4	9.5	100.0

Equivalence of Short Form to Form A

To support the equivalence of the Short Form and Form A, test item contents were not changed and the new form was assembled from Form A test items. As a result, the Short Form may be considered to measure the same abilities as Form A.

Following assembly of the Short Form, correlation coefficients were computed between raw scores on the Short Form and those on Form A. Because the constituent test items of the Short Form are completely contained in the longer Form A, the coefficients are considered part-whole correlations (r_{pw}). Part-whole correlations are known to overstate the relationship between independently measured variables and cannot be interpreted as alternate form correlations. However, they can be used to support the equivalence of a Short Form to a longer one, because examinees are expected to respond the same way to the same items, regardless of the form.

The overall correlation coefficient was calculated by using data from a sample of 3,727 adults who were administered Form A. To compute the correlations, each Watson-Glaser was scored twice. First, the Form A raw score was computed, then the Short Form raw score was computed by ignoring responses to the Form A items not used in the Short Form. For the entire sample, the resulting coefficient was .96. Correlations between Form A and the Short Form scores were also computed separately for each of 21 sources providing data, some of which were not included in the Short Form developmental analysis. The resulting coefficients are presented in Table 7.1. A description of the sample group is followed by the group size (N) and the part-whole correlation (r_{pw}) between the Short Form and Form A. The coefficients presented in Table 7.1 indicate that raw scores on the Short Form correlated very highly with Form A raw scores in a variety of groups.

Table 7.1 Part-Whole Correlations (r_{pw}) of the Short Form and Form A

Group	<i>N</i>	r_{pw}
Lower-level management applicants	219	.93
Lower to upper-level management applicants	501	.94
Mid-level management applicants	211	.94
Upper-level management applicants at Board of County Commissioners	215	.94
Construction management applicants	322	.94
Executive management applicants	453	.93
Supervisory and managerial applicants in the corrugated container industry	149	.94
Sales applicants	473	.94
Mid-level marketing applicants	909	.94
Bank employees	95	.95
Bank management associates	131	.94
Candidates for the ministry	126	.95
Clergy	99	.91
Railroad dispatchers	199	.92
Nurse managers and educators	111	.95
Police officers	225	.95
Administrative applicants in city government	23	.97
Security applicants	42	.89
Candidates for police captain	41	.89
Police department executives	55	.94
Various occupations	133	.97

Equivalent Raw Scores

Table 7.2 presents raw score equivalents for the Short Form and Form A. For every possible score on the Short Form, this table contains an equivalent raw score on Form A. To convert a Form A raw score, find that score in the Form A column. Then, look in the Short Form raw score column on the left. To convert a Short Form raw score, simply reverse the procedure. Table 7.2 was prepared with data obtained from 3,727 adults comprising the item selection sample. To establish equivalent raw scores, raw-score-to-ability estimates for both the Short Form and Form A were generated using Rasch-model difficulty parameters. Then, using interpolation when necessary, the ability estimates were calibrated for all possible scores on each form. Form A and Short Form raw scores corresponding to the same ability estimate were considered equivalent (i.e., represent the same ability level).

Organizations requiring an alternative to the Short Form for retesting or other purposes may use Form B. Form A and Form B are equivalent, alternate forms. Raw scores on one form (*A* or *B*) may be interpreted as having the same meaning as identical raw scores on the other form (*A* or *B*). Therefore, scores from either Form A or Form B may be equated to Form S using Table 7.2.

Table 7.2 Raw Score Equivalencies Between the Short Form and Form A

Short Form Raw Score	Form A Raw Score	Short Form Raw Score	Form A Raw Score
40	78–80	20	40–41
39	77	19	38–39
38	75–76	18	36–37
37	73–74	17	34–35
36	71–72	16	33
35	69–70	15	31–32
34	67–68	14	29–30
33	65–66	13	27–28
32	63–64	12	25–26
31	61–62	11	23–24
30	59–60	10	21–22
29	57–58	9	19–20
28	55–56	8	17–18
27	53–54	7	15–16
26	51–52	6	13–14
25	49–50	5	10–12
24	48	4	8–9
23	46–47	3	6–7
22	44–45	2	4–5
21	42–43	1	1–3

Equivalence of Computer-Based and Paper-and-Pencil Versions of the Short Form

Studies of the effect of the medium of test administration 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, Harcourt Assessment conducted an equivalency study using paper-and-pencil and computer-administered versions of the Short Form.

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 the other participants ($n = 108$) completed the tests in the reverse order. Table 7.3 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 coefficients indicate that paper-and-pencil raw scores correlate very highly with online administration raw scores (.86 and .88, respectively). 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 7.3 Equivalency of Paper and Online Modes of Administration

Administration Order	N	Paper		Online		r
		Mean	SD	Mean	SD	
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

The *reliability* of a measurement instrument refers to the accuracy and precision of test results and is a widely used indicator of the confidence that may be placed in those results. 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. A perfectly reliable test would have a reliability coefficient of 1.00 and no measurement error. A completely unreliable test would have a reliability coefficient of .00. 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), alternate forms (the consistency of scores across alternate forms of a test), and internal consistency of the test items (e.g., *Cronbach’s alpha coefficient*, Cronbach 1970).

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. A true score is a theoretical score entirely free of error. Since the true score is a hypothetical value that can never be obtained because testing always involves some measurement error, the score obtained by an examinee on any test will vary somewhat from administration to administration. As a result, 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.

Historical Reliability

Previous Studies of Internal Consistency Reliability

For the sample used in the initial 1994 development of the Watson-Glaser Short Form ($N = 1,608$), Cronbach's alpha coefficient (r) was .81. Cronbach's alpha and the SEM were also calculated for a number of groups separately, including some groups that were in the development sample and some that were not in the development sample (see Table 8.1).

Table 8.1 Means, Standard Deviations (SD), Standard Errors of Measurement (SEM) and Internal Consistency Reliability Coefficients (r_{α}), Based on Previous Studies

Group	<i>N</i>	Mean	<i>SD</i>	SEM	r_{α}
Lower-level management applicants	219	33.50	4.40	2.17	.76
Lower to upper-level management applicants	501	32.29	4.63	2.31	.75
Mid-level management applicants	211	33.99	4.20	2.12	.74
Upper-level management applicants at a Board of County Commissioners	215	31.80	5.20	2.35	.80
Executive management applicants	453	33.42	4.21	2.18	.73
Construction management applicants	322	32.05	4.87	2.32	.77
Supervisory and managerial applicants in the corrugated container industry	149	31.48	5.00	2.39	.77
Sales applicants	473	30.88	4.98	2.43	.76
Mid-level marketing applicants	909	31.02	5.08	2.42	.77
Bank employees	95	32.75	4.58	2.25	.76
Bank management associates	131	31.61	4.69	2.39	.74
Candidates for the ministry	126	34.10	4.71	2.08	.80
Clergy	99	34.56	3.79	2.05	.71
Railroad dispatchers	199	25.15	5.00	2.78	.69
Nurse managers and educators	111	30.52	4.86	2.46	.74
Police officers	225	28.00	6.03	2.64	.81
Various occupations	133	30.68	6.65	2.40	.87
Administrative applicants in city government ¹	23	30.43	5.82	2.42	.83
Security applicants	42	25.00	4.79	2.77	.67
Candidates for police captain ²	41	27.95	4.60	2.69	.66
Police department executives ³	55	32.56	4.14	2.32	.69

¹ D.O.T. Code 169.167-010

² D.O.T. Code 375.167-034

³ Includes Commander (D.O.T. Codes 375.167-034 and 375.267-026) Chief (D.O.T. Code 375.117-010), Deputy Chief (D.O.T. Code 375.267-026) and Warden (D.O.T. Code 187, 117-018)

Using the SEM means that scores are interpreted as bands or ranges of scores, rather than as precise points. Thinking in terms of score ranges serves as a check against overemphasizing small differences between scores. The SEM may be used to determine whether an individual's score is significantly different from a cut score, or whether the scores of two individuals differ significantly. An example of one general rule of thumb is that the difference between two scores on the same test should not be interpreted as significant unless the difference is equal to at least twice the SEM of the test (Aiken, 1979; as reported in Cascio, 1982).

The internal consistency estimates calculated for the Short Form tests were moderately low, consistent with research involving previous forms of the Watson-Glaser; for this reason, individual test scores should not be used.

Previous Studies of Test-Retest Reliability

In 1994, a study investigating the test-retest reliability and required completion time of the Watson-Glaser–Short Form was conducted at a large publishing company. A sample of 42 employees (92.9% non-minority; 54.8% female) completed the Short Form two weeks apart. The participants worked in a variety of positions ranging from Secretary to Project Director. The mean score at the first testing was 30.5 ($SD = 5.6$) and at the second testing 31.4 ($SD = 5.9$), while the test-retest correlation was .81 ($p < .001$). Scores for females ($Mean = 31.0$, $SD = 6.1$) and male ($Mean = 31.8$, $SD = 5.7$) respondents were not significantly different ($t = 0.11$, $df = 40$).

Current Reliability Studies

Evidence of Internal Consistency Reliability

Cronbach's alpha and the standard error of measurement (SEM) were calculated for the samples used for the current norm groups (see Table 8.2). Reliability estimates for these samples were similar to those found in previous studies and ranged from .76 to .85. Consistent with previous research, these values indicate that the total score possesses adequate reliability. The test scores obtained lower estimates of internal consistency reliability, thereby suggesting that the test scores alone should not be used.

Table 8.2 Means, Standard Deviations (SD), Standard Errors of Measurement (SEM) and Internal Consistency Reliability Coefficients (r_{α}) for the Current Short Form Norm Groups

Group	<i>N</i>	Mean	<i>SD</i>	SEM	r_{α}
Industry					
Advertising/Marketing/Public Relations	101	28.7	6.1	2.52	.83
Education	119	30.2	5.4	2.47	.79
Financial Services/Banking/Insurance	228	31.2	5.7	2.42	.82
Government/Public Service/Defense	130	30.0	6.3	2.44	.85
Health Care	195	28.3	6.5	2.60	.84
Information Technology/Telecommunications	295	31.2	5.5	2.40	.81
Manufacturing/Production	561	32.0	5.3	2.31	.81
Professional Business Services (e.g., Consulting, Legal)	153	31.9	5.6	2.31	.83
Retail/Wholesale	307	30.8	5.3	2.43	.79
Occupation					
Accountant/Auditor/Bookkeeper	118	30.2	5.8	2.46	.82
Consultant	139	33.3	4.8	2.20	.79
Engineer	225	32.8	4.8	2.25	.78
Human Resource Professional	140	30.0	5.7	2.48	.81
Information Technology Professional	222	31.4	5.9	2.36	.84
Sales Representative—Non-Retail	353	29.8	5.1	2.50	.76
Position Type/Level					
Executive	409	33.4	4.5	2.20	.76
Director	387	32.9	4.7	2.20	.78
Manager	973	30.7	5.4	2.41	.80
Supervisor	202	28.8	6.2	2.63	.82
Professional/Individual Contributor	842	30.6	5.6	2.44	.81
Hourly/Entry-Level	332	27.7	5.9	2.64	.80
Norms by Occupation Within Specific Industry					
Manager in Manufacturing/Production	170	31.9	5.3	2.31	.81
Engineer in Manufacturing/Production	112	32.9	4.7	2.25	.77

Evidence of Test-Retest Reliability

Test-Retest reliability was evaluated for the total score and for the individual test scores in a sample of job incumbents representing various organizational levels and industries. ($N = 57$). The test-retest intervals ranged from 4 to 26 days, with a mean interval of 11 days. As the data in Table 8.3 indicate, the Watson-Glaser Total score demonstrates acceptable test-retest reliability ($r_{12} = .89$). The difference in mean scores between the first testing and the second testing is statistically small ($d' = 0.17$). This difference (d'), proposed by Cohen (1988), is useful as an index to measure the magnitude of the actual difference between two means. The difference (d') is calculated from dividing the difference of the two test means by the square root of the pooled variance, using Cohen's (1996) Formula 10.4.

The test-retest reliability of the Watson-Glaser Total score has a small difference index ($d' = 0.17$), indicating that the magnitude of the difference in mean scores between first testing and the retesting is statistically small. In other words, the Watson-Glaser Total score is stable over the test-retest period. The test-retest reliability coefficients of the test scores are somewhat lower, suggesting that the Total score is more reliable than the test scores as a measure of critical thinking.

Table 8.3 Test-Retest Reliability of the Short Form

Score	First Testing		Second Testing		r_{12}	Difference (d')
	Mean	<i>SD</i>	Mean	<i>SD</i>		
Total	29.5	7.0	30.7	7.0	.89	.17
Inference	4.3	1.9	4.6	1.9	.70	.16
Recognition of Assumptions	6.1	2.2	6.5	2.0	.83	.19
Deduction	7.1	1.8	7.0	2.0	.55	-.05
Interpretation	5.0	1.6	5.1	1.6	.78	.06
Evaluation of Arguments	7.2	1.6	7.5	1.3	.67	.21

The validity of a test refers to the degree to which specific data, research, or theory support that the test measures what it is intended to measure. Validity is a unitary concept. It is the extent to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose (AERA, APA, & NCME, 1999). “Validity is high if a test gives the information the decision maker needs” (Cronbach, 1970). Data from the Short Form sample was analyzed for evidence of validity based on content, test-criterion relationships, and evidence of convergent and discriminant validity.

Evidence of Validity Based on Content

Evidence based on the content of a test exists when a test includes a representative sample of tasks, behaviors, knowledge, skills, abilities, or other characteristics necessary to perform the job. Evidence of content validity is usually gathered through job analysis and is most appropriate for evaluating knowledge and skills tests.

Evaluation of content-related evidence is usually a rational, judgmental process (Cascio & Aguinis, 2005). 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 the job performance universe 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 the job performance universe (e.g., a typing test administered to a secretary applies to typing, one job performance domain in the job performance universe of a secretary). 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 measured by the Watson-Glaser. Content-related validity evidence of the Watson-Glaser in classroom and instructional settings may be examined by noting the extent to which the Watson-Glaser measures a sample of the specified objectives of such instructional programs.

Evidence of Criterion-Related Validity

One of the primary reasons tests are used is to provide an educated guess about an examinee's potential for future success. For example, selection tests are used to hire or promote those individuals most likely to be productive employees. The rationale behind selection tests is this: the better an individual performs on a test, the better this individual will perform as an employee.

Criterion-related validity evidence addresses the inference that individuals who score better on tests will be successful on some criterion of interest. Criterion-related validity evidence indicates the statistical relationship (e.g., for a given sample of job applicants or incumbents) between scores on the test and one or more criteria, or between scores on the tests and independently obtained measures of subsequent job 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. Typically, correlations between criterion measures and scores on the test serve as indices of criterion-related validity evidence. 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.

Unfortunately, the conditions for evaluating criterion-related validity evidence are often difficult to fulfill in the ordinary employment setting. Studies of test-criterion relationships should involve a sufficiently large number of persons hired for the same job and evaluated for success using a uniform criterion measure. The criterion itself should be reliable and job-relevant, and should provide a wide range of scores. In order to evaluate the quality of studies of test-criterion relationships, it is essential to know at least the size of the sample and the nature of the criterion.

Assuming that the conditions for a meaningful evaluation of criterion-related validity evidence have been met, 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). This suggestion is because the practical value of the test depends not only on the validity, but also other factors, such as the base rate for success on the job (i.e., the proportion of people who would be successful in the absence of any selection procedure). If the base rate for success on the job 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.

In addition to the practical value of validity coefficients, the *statistical significance* of coefficients should be noted. Statistical significance refers to the odds that a non-zero correlation could have occurred by chance. If the odds are 1 in 20 that a non-zero correlation could have occurred by chance, then the correlation is considered statistically significant. Some experts prefer even more stringent odds, such as 1 in 100, although the generally accepted odds are 1 in 20. In statistical analyses, these odds are designated by the lower case *p* (probability) to signify whether a non-zero correlation is statistically significant. When *p* is less than or

equal to .05, the odds are presumed to be 1 in 20 (or less) that a non-zero correlation of that size could have occurred by chance. When p is less than or equal to .01, the odds are presumed to be 1 in 100 (or less) that a non-zero correlation of that size occurred by chance.

Previous Studies of Evidence of Criterion-Related Validity

Previous studies have shown evidence of the relationship between Watson-Glaser scores and various job and academic success criteria. Gaston (1993), in a study of law enforcement personnel, found a relationship between Watson-Glaser scores and organizational level. Among his findings were that executives scored in the 7–9 decile range more often than non-executives, and non-executives scored in the 1–3 decile range more often than executives. Holmgren and Covin (1984), in a study of students majoring in education, found Watson-Glaser scores correlated .50 with grade-point average (GPA) and .46 with English Proficiency Test scores. In studies of nursing students, Watson-Glaser scores correlated .50 with National Council Licensure Exam (NCLEX) scores (Bauwens and Gerhard, 1987) and .38 with state licensing exam scores (Gross, Takazawa, & Rose). Additional studies of the relationship between Watson-Glaser scores and various criteria are reported in the previous version of the manual (1994).

Current Studies of Evidence of Criterion-Related Validity

Studies continue to provide strong criterion-related validity evidence for the Watson-Glaser. Kudisch and Hoffman (2002) reported that, in a sample of 71 leadership assessment center participants, Watson-Glaser scores correlated with ratings on Analysis, .58, and with ratings on Judgment, .43. Ratings on Analysis and Judgment were based on participants' performance on assessment center exercises, including a coaching meeting, in-basket exercise or simulation, and a leaderless group discussion.

Spector, Schneider, Vance, and Hezlett (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 overall scores on six of eight assessment center exercises, and related more strongly to exercises involving primarily cognitive problem-solving skills (e.g., $r = .26$, $p < .05$, with in-basket scores) than exercises involving a greater level of interpersonal skills (e.g., $r = .16$, $p < .05$, with in-basket coaching exercise).

In a study we conducted for this revision of the manual in 2005, we examined the relationship between Watson-Glaser scores and on-the-job performance of 142 job incumbents in various industries. Job performance was defined as supervisory ratings on behaviors determined through research to be important to most professional, managerial, and executive jobs. The study found that Watson-Glaser scores correlated .33 with supervisory ratings on a dimension made up of Analysis and Problem Solving behaviors, and .23 with supervisory ratings on a dimension made up of Judgment and Decision Making behaviors. Supervisory ratings from the sum of ratings on 19 job performance behaviors ("Total Performance"), as well as ratings on a single-item measure of "Overall Potential" were also obtained. The Watson-Glaser scores correlated .28 with "Total Performance" and .24 with ratings of Overall Potential.

In an analysis of a sub-group of the 2005 study mentioned above, we examined the relationship between the Watson-Glaser scores and on-the-job performance of 64 analysts from a government agency. The results showed that Watson-Glaser

scores correlated .40 with supervisory ratings on each of the 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. In the sample of 64 analysts mentioned above, the Watson-Glaser scores correlated .39 with "Total Performance" and .25 with Overall Potential.

Another part of the study we conducted in 2005 for this revision of the manual examined the relationship between Watson-Glaser scores and job success as indicated by organizational level achieved, for 2,303 job incumbents across 9 industry categories. Results indicated that Watson-Glaser scores correlated .33 with organizational level.

Other studies of job-relevant criteria have found significant correlations between Watson-Glaser scores and creativity (Gadzella & Penland, 1995), facilitator effectiveness (Offner, 2000), positive attitudes toward women (Loo & Thorpe, 2005), and openness to experience (Spector, et al., 2000).

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), in a study of 98 college freshmen, found 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), academic performance during pre-clinical years of medical education (Scott & Markert, 1994), and preference for contingent, relativistic thinking versus "black-white, right-wrong" thinking (Taube, 1995).

Table 9.1 presents a summary of studies that evaluated criterion-related validity evidence for the Watson-Glaser since 1994 when the previous manual was published. Only studies that reported validity coefficients are shown. Additional studies are reported in this chapter as well as the previous version of manual (1994).

In Table 9.1, the column entitled *N* details the number of cases in the sample. The criterion measures include job performance and grade point average, among others. Means and standard deviations, for studies in which they were available, are shown for both the test and criterion measures. The validity coefficient for the sample appears in the last column. Validity coefficients such as those reported in Table 9.1 apply to the specific samples listed.

Table 9.1 Studies Showing Evidence of Criterion-Related Validity

Group	N	Watson-Glaser			Criterion			r
		Form	Mean	SD	Description	Mean	SD	
Leadership assessment center participants from a national retail chain and a utility service (Kudisch & Hoffman, 2002)	71	80-item	–	–	Assessor Ratings:			
					Analysis	–	–	.58*
Middle-management assessment center participants (Spector, Schneider, Vance, & Hezlett, 2000)	189–407	80-item	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*
					CPI Score: Openness to Experience	41.8	6.4	.36*
Job incumbents across multiple industries (Harcourt Assessment, Inc., 2005)	142	Short	Supervisory Ratings:					
			Analysis and Problem Solving				.33**	
			Judgment and Decision Making				.23**	
			Total Performance				.28**	
Job applicants and incumbents across multiple industries (Harcourt Assessment, Inc., 2005)	2,303	Short				Potential		.24**
						Org. Level		.33*

(continued)

