Portions of this work were published in previous editions.

Printed in the United States of America.

Published by Pearson
19500 Bulverde Rd.
San Antonio, TX 78259, USA
1-800-211-8378

ISBN 0158339304

1 2 3 4 5 6 7 8 9 10 11 12 A B C D E

Table of Contents

List of Tables

Acknowledgements

The development and publication of updated information on a test like the *Bennett Mechanical Comprehension Test* inevitably involves the helpful participation of many people in several phases of the project—design, data collection, statistical data analyses, editing, and publication. The Pearson Education 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; Colleen McAndrews, CAT Team Workflow Coordinator; 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.

Special thanks go to Mark Cooley, Designer, who coordinated the creation of the updated artwork and the overall design of the manual and question booklets. Our thanks also go to Troy Beehler, Toby Mahan, and Peter Schill, Project Managers, for skillfully managing the logistics of this project. Troy, Toby, and Peter worked with several team members from the Technology Products Group 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; and Maurya Duran, Technical Writer. Dawn Dunleavy, Senior Managing Editor, provided editorial guidance and support. Production assistance was provided by Stephanie Adams, Director, Production; Dione Farmer Anderson, 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: Jenifer Kihm, PhD, Talent Assessment Product Line Manager; Judy Chartrand, PhD, Director, Psychological Assessment Products Group; Gene Bowles, Vice President, Publishing and Technology; Larry Weiss, PhD, Vice President, Psychological Assessment Products Group; and Aurelio Prifitera, PhD, Publisher, Pearson Education, Inc.

John Trent, MS, Research Director

Mark Rose, PhD, Research Director

Kingsley C. Ejiogu, PhD, Research Director

Introduction 1

The *Bennett Mechanical Comprehension Test* (BMCT) is an assessment tool for measuring a candidate's ability to perceive and understand the relationship of physical forces and mechanical elements in practical situations. This aptitude is important in jobs and training programs that require the understanding and application of mechanical principles. The current BMCT Forms, S and T, have been used to predict performance in a variety of vocational and technical training settings and have been popular selection tools for mechanical, technical, engineering, and similar occupations for many years.

In response to customer needs, Pearson Education, Inc. has made the following enhancements to better serve users of the *Bennett Mechanical Comprehension Test.*

Guidelines for Administration—Chapter 3 provides administration and scoring directions for the traditional paper-and-pencil version. Chapter 4 provides directions for administering the new computer-based version.

New Normative Information (Norms) is Provided—Chapter 5 presents new norms based on data from 10 new norm groups, based on 1,232 cases collected between 2003 and 2005.

New Item Art—New art was rendered for the Form S and Form T Test Booklets. Chapter 6 provides information on the updated item art for Forms S and T Test Booklets.

Equivalency Study Results—Chapter 6 presents the results of the equivalency study on the computer-based and the paper-and-pencil versions.

Updated Evidence of Reliability and Validity—New studies describing internal consistency and test-retest reliability are presented in chapter 7. New studies describing convergent and criterion-related validity are presented in chapter 8.

The BMCT may be administered to an individual or to a small group. Examinees study the items in a Test Booklet (Form S or T) and mark their answers on an Answer Sheet that may be hand- or machine-scored. With the computer-based BMCT, the problems are presented on screen and the examinee clicks on his or her answer option choice.

Mechanical Comprehension

The BMCT is an *aptitude* test and functions differently from an *achievement* test. An aptitude test is used to measure a person's ability for future learning. An achievement test is used to measure a person's present or past accomplishments. The BMCT was developed to measure a person's aptitude for understanding and

applying mechanical principles, from which an employer may infer future performance in jobs that require these skills. This aptitude, known as *mechanical comprehension,* is regarded as one aspect of intelligence, as intelligence is broadly defined. The individual who scores high in mechanical comprehension tends to learn readily the principles of the operation and repair of complex devices.

Like those who take other aptitude tests, a person's performance on the BMCT may be influenced by environmental factors, but not to the extent that interpreting his or her performance is significantly affected. Although an individual's scores on the BMCT can generally be improved through training and experience, it is unlikely that improvement will be dramatic (for additional information, see "Effects of Training and Practice" in chapter 7). This situation is due in part to the presentation and composition of items that are simple, frequently encountered mechanisms, neither resembling textbook illustrations nor requiring special knowledge.

The various forms of the BMCT have been independently evaluated by a number of authors. For reviews of the test, see Bechtoldt (1972); Ciechalski (2005); Dagenais (1992); Hambleton (1972); Roberts (1972); Spangler (2005); and Wing (1992). See Hegarty, Just, and Morrison (1988) for a theoretical discussion of mechanical comprehension and the BMCT.

Description of the BMCT

Each form of the BMCT is composed of 68 items that are illustrations of simple, frequently encountered mechanisms. For each item, the examinee reads a question about an illustration, examines the illustration, and chooses the best answer to the question from among three options. Though there is a 30-minute time limit for completing the test, the BMCT is not truly a *speeded* test. Speeded tests are composed of relatively easy items and rely on the number of correct responses within restrictive time limits to differentiate performance among examinees. Because BMCT items represent a wide range of item-level difficulty and the test has a 30-minute time limit, it is considered a *timed power* test. Professionals who are responsible for talent assessment can set different BMCT cut scores appropriate for different job requirements.

Reading Level

The reading levels of BMCT Forms S and T were compared to the *EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies* (Taylor et al., 1989) and *Basic Reading Vocabularies* (Harris & Jacobson, 1982). Approximately 99.5% of the words in Form S directions and exercises were at or below the sixth-grade reading level, and 98.6% of words in Form T directions and exercises were at or below the sixth-grade level. Table 1.1 is a summary of word distribution by grade level for both forms. This comparison showed that both the directions and the test questions fell within the "fairly easy" range, similar to reading levels in popular fiction books or magazines.

Table 1.1 Grade Level of Words on BMCT Forms S and T

	Form S		Form T	
Grade Level	**Freq.**	**Percent**	**Freq.**	**Percent**
Preprimer	187	18.9	169	18.2
1	212	21.5	190	20.4
2	273	27.6	242	26.1
3	200	20.3	185	19.9
4	75	7.6	91	9.8
5	23	2.3	28	3.0
6	13	1.3	11	1.2
7	1	0.1	1	0.1
8	4	0.4	10	1.1
9	0	0.0	0	0.0
10	0	0.0	2	0.2
11	0	0.0	0	0.0
Total	988	100.0	929	100.0

General Directions for Administration and Scoring

2

Testing Conditions

It is important to administer the test 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.

For Paper-and-Pencil Administration: Each examinee needs an adequate flat surface on which to work. The surface should be large enough to accommodate an open Test Booklet and an Answer Sheet.

Not Allowed in the Testing Session: Handbags, briefcases, and other personal materials are not allowed on or near the work surface. Examinees may not use reference materials, books, or notes to take the test, and such materials must be placed out of reach.

Preparing for Administration

The person responsible for administering the BMCT does not need special training, but he or she must be able to carry out standard assessment procedures. To ensure accurate and reliable results, the administrator must become thoroughly familiar with the Directions for Administration and the test materials before attempting to administer the test. Being thoroughly prepared with the testing materials before the examinee's arrival will result in a more efficient testing session, whether you administer the paper-and-pencil or online version of the test. The best way for test administrators to familiarize themselves with the BMCT is to take the test, complying with the 30-minute time limit, prior to administration.

Although the BMCT is intended as a measure of power rather than speed, there is a 30-minute time limit for completing the test. As the test administrator, you should have a regular watch with second hand, a wall clock with a sweep second hand, or any other accurate timer to time the administration. To facilitate accurate timing, write the starting time immediately after you have given the signal to begin. Write the finishing time immediately after you have given the signal to stop.

Allow at least 50 minutes for total administration time—this includes giving directions, distributing test materials, actual testing, and collecting materials.

Answering Questions

Be sure that the examinees understand how to mark their answers correctly on the Answer Sheet. Although the Answer Sheet is not difficult to mark properly, some examinees may not be familiar with this type of form. It is the responsibility of the test administrator to ensure that examinees understand the correct way to indicate their answers on the Answer Sheet.

Examinees may ask questions about the test before you give the signal for the test to begin. To maintain standardized testing conditions, answer questions by rereading the appropriate section of the directions. Do not volunteer new explanations or examples, but make sure the examinees understand what is required of them. Do not rush or omit the question period before the test begins.

All directions that the test administrator reads aloud to examinees appear in bold type in this manual and in the Directions for Administration. 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.

Test Security

Keep all test materials secure before, during, and after testing. Store materials in a secure place (for example, a locked cabinet). Allow only designated test administrators to have access to the BMCT materials.

BMCT 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 individuals access to score information if they do not legitimately need it. Storing test scores in a locked cabinet or password protected file that can be accessed only by designated test administrators will help ensure security. The security of testing materials and protection of copyright must be maintained at all times by authorized individuals.

Do not disclose test access information, such as usernames or passwords, and administer the test only in proctored sessions. All the computer stations used in administering the BMCT must be in locations that can be supervised easily with the same level of security as with paper-and-pencil administration.

Accommodating Examinees With Disabilities

You must provide reasonable accommodations for candidates with special needs to take the test comfortably. Reasonable accommodations may include, but are not limited to, modifications to the test environment (e.g., desk height) and delivery (e.g., having a reader read questions to the examinee) (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 BMCT, 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 BMCT 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 test administrator must provide reasonable accommodation for the candidate to take the test on paper, and then have the candidate's responses 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.

Administering the Paper-and-Pencil BMCT 3

Paper-and-Pencil Administration Materials

- The BMCT Directions for Administration (DFA)
- A watch or clock with a sweep second hand
- 1 Test Booklet and 1 Answer Sheet per examinee. (The Answer Sheet can be used with printed Form S or T.)
- 2 sharpened No. 2 pencils with erasers per examinee

When all examinees are seated, give each person two pencils and then distribute the Answer Sheets. After you distribute the Answer Sheets, 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 blacken the circle properly.

Point to the *Correct Mark* and *Incorrect Mark* examples on the Answer Sheet.

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

Note: If text or numbers should be aligned in a particular way for subsequent processing of Answer Sheets (i.e., all right-aligned or all left-aligned), instruct examinees to align their entries accordingly.

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

The Answer Sheet provides space for a nine-digit identification number. If you would like the examinees to use this space, 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 examinees do not need to record information in the Identification box, tell them to skip box B and go on to box C.

Say **Find box C, labeled DATE. Write 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 any additional information you would like to obtain from the examinees (e.g., job title, department, years of education). Let the examinees know what information, if any, they should provide in this box.

Note: If optional information is collected, tell examinees how the information will be used.

Say **After you receive a 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 **Below box D on the Answer Sheet is a box labeled "E, Form Used." Look on the cover of your Test Booklet where you will see either Form S or Form T printed in dark type. Now, on the Answer Sheet, blacken the circle of the letter of the form that you are taking.**

If you are only using one form, simply say, **Fill in the *S* circle** or **Fill in the *T* circle.** After the examinees have filled in the circle, say,

Are there any questions?

Answer any questions the examinees have about the test. When you are sure that all questions have been answered, say

Open your Test Booklet to page 2 and fold back the rest of the booklet. Read the directions silently at the top of page 2 while I read them aloud. Look at Sample X on this page. It shows two men carrying a weighted object on a plank, and it asks, "Which man carries more weight?" Because the object is closer to man B than to man A, man B is shouldering more weight; so blacken the circle with the B on your Answer Sheet. Now look at Sample Y and answer it yourself. Blacken the circle with the letter that represents your answer on the Answer Sheet.

Allow sufficient time for the examinees to complete the sample.

Say **Study the drawing. Which letter shows the seat where a passenger will get the smoothest ride? Note that the letter C is between the two supporting wheels, so the passenger sitting at C will get the smoothest ride. You should have blackened the circle with the C for Sample Question Y.**

Now read the instructions silently at the bottom of Page 2 while I read them aloud. On the following pages, there are more pictures and more questions. Read each question carefully, look at the picture, and blacken the circle with the letter that represents your answer on the Answer Sheet. Make sure that your marks are heavy and black. Completely erase any answer you wish to change. Do not make any marks in this booklet.

You will have 30 minutes for the entire test. Start with Page 3 and continue until you reach the last page marked "End of Test." If you need a pencil, raise your hand and I will give you one. Are there any questions? If you have any questions, you must ask them now because questions are not allowed once the test has begun.

Answer any questions, preferably by rereading the appropriate section of the directions. Then, say,

Remember, do not write in the Test Booklet. Blacken the correct circle that corresponds to the answer option you have chosen and make sure the item number on the Answer Sheet matches the item you are answering in the Test Booklet. Make sure that you blacken the entire circle, and that you keep your marks within the circle. Try to answer every question. Now turn over the Test Booklet to page 3 and begin the test.

Start timing immediately after you give the signal to begin. If any of the examinees finish before the end of the test period, tell them to sit quietly until everyone has finished, or collect their materials and dismiss them quietly.

At the end of 30 minutes, say,

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

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. Test booklets with marks that cannot be completely erased should not be reused.

Scoring the BMCT Answer Sheet

The BMCT Answer Sheet, used for both Form S and Form T, may be hand-scored with the corresponding scoring template or machine-scored.

Using the Hand-Scoring Templates

Make sure you use the correct scoring template for the form that was administered. Before you apply the Scoring Template:

- ❑ Check the Answer Sheet for items that have multiple responses and draw a heavy red mark through multiple responses to the same item. (*Note:* Using a red line is suitable only when the sheets are to be hand-scored.)
- ❑ Then, check for any answer spaces that were only partially erased by the examinee in changing an answer and erase it completely.
- ❑ Next, place the scoring template on top of the Answer Sheet so that the edges are aligned and the two stars appear through the two holes closest to the top of the template.
- ❑ Count the number of correctly marked spaces (other than those through which a red line has been drawn) that appear through the holes in the stencil, and record the total in the Raw Score space on the Answer Sheet. The maximum raw score for Form S or T is 68. You may record the percentile score that corresponds to the raw score in the space labeled Percentile. You may record the norm group used to determine that percentile in the space labeled "Norms Used" (e.g., Automotive Mechanic).

Preparing Answer Sheets for Machine-Scoring

You may program any reflective scanning device to process the machine-scoreable Answer Sheet. Before you can scan an Answer Sheet for scoring, you must completely erase multiple responses to the same item or configure the scanning program to treat multiple responses as incorrect answers. Any answer options that were only partially erased by the examinee in changing an answer must be completely erased.

Directions for Computer-Based Administration

4

The computer-based *Bennett Mechanical Comprehension Test* is administered through eAssessTalent.com, an Internet-based testing system designed by Pearson Education, 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 BMCT interpretive reports also are provided on the website. After a candidate has taken the BMCT on eAssessTalent.com, the test administrator can review the candidate's results in an interpretive report, using the link that Pearson Education, Inc., provides.

Administering the Test

Examinees do not need pencils or scratch paper for this computer-based test. After you have accessed eAssessTalent.com and the initial instruction screen for the BMCT appears, seat the examinee at the computer and 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 30 minutes to work on this test. The test ends with a few additional demographic questions. Do you have any questions before starting the test?

Answer any questions the examinee may have and say,

Please begin the test.

Once the examinee clicks the "Start Your Test" button, the first page of test questions appears. During the test, the examinee has 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 30 minutes to complete the test.

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 Pearson Education 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.

Comparing Scores to Normative Data 5

The BMCT raw score is the total number of correct responses. Raw scores may be used to rank examinees in order of performance, but little can be inferred from raw scores alone. To make the test results meaningful, it is necessary to relate the scores to specifically defined normative groups.

Normative information (norms) provides a basis for evaluating an individual's score relative to the scores of other individuals who took the same test. Norms provide a means by which raw scores can be converted 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; norms represent the performance of the group.

The characteristics of the sample used for preparing norms are critical in determining the usefulness of the norms. For some purposes, such as intelligence testing, norms that are representative of the general population are essential. For other purposes, such as selection 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 and locations, and over time, the limitations on the usefulness of any set of published norms should be acknowledged.

When test results are used in making employment selection 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 would give an unwarranted impression of precision. Until a sufficient and representative number of cases has been collected (preferably 100 or more), use the normative data in Tables A.4 and A.5 to guide interpretation of test scores.

Development of Scaled Scores

Using a group of examinees who took the BMCT between 2003 and 2005 (N = 4,606), the BMCT raw scores were converted to estimated theta values (i.e., ability scores) using Item Response Theory (IRT—the 1-parameter logistic model, also known as the Rasch model; Rasch, 1980). The theta values were then used to derive scaled scores (Yang, Wang, Trent, & Rose, 2006). The scaled scores created for the BMCT range from 10 to 90, with a mean of 50 and a standard deviation of 10. The scaled scores provide a basis for comparisons between examinees and the general population. However, making employment selection decisions based on small differences in scaled scores is not recommended. The consideration of test scores for selection decisions should be made using a comparison group that is most appropriate and relevant for the job. Raw scores, corresponding scaled scores, and percentiles are presented in appendix A, Tables A.4 and A.5.

Development of Norms

The norms in appendix A, Tables A.4 and A.5 were derived from data collected between 2003 and 2005 from a group of 1,232 adults representing a variety of employment settings. Table A.4 shows BMCT raw scores with the corresponding scaled scores and percentiles based on occupation. Table A.5 provides BMCT raw scores with the corresponding scaled scores and percentiles based on industry.

Before looking up scaled scores or percentiles in Tables A.4 and A.5, look for a group in Tables A.1–A.3 that is similar to the individual or group that was tested. Tables A.1–A.3 provide relevant information about each norm group to assist in selecting the most appropriate norm group. For example, test scores of a person applying for a position as a mechanic may be compared with norms derived from the scores of other mechanics. If a person has applied for an electrician position, his or her test scores should be compared with those of others in similar positions (e.g., skilled tradesperson norms).

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 position relative to the norm group. Percentile ranks are not the same as percentage scores, which represent the percentage of correct items. Percentile ranks are derived scores, which are expressed in terms of the percentage of people in the norm group who scored below or equal to a given raw score.

Although percentiles are useful for explaining an examinee's performance relative to others, there are 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 percentile ranks.

Converting Raw Scores to Percentile Ranks

To find the percentile rank that corresponds to a raw score, locate the raw score in the right or left column in Tables A.4–A.5. Then, look for the appropriate norm group column. Look down the column until you find the percentile rank for the raw score. For example, if a person applying for a job as an engineer had a BMCT score of 50, it is appropriate to use the Engineer norms in Table A.4 for comparison. In this case, a raw score of 50 has a corresponding percentile rank of 35. This percentile rank indicates that about 35% of the people in the norm group scored lower than or equal to a score of 50 on the BMCT, and about 65% scored higher than a score of 50 on the BMCT.

Each group's size *(N* count*),* raw score mean, and standard deviation *(SD)* are listed at the bottom of each norm table. 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.

History of the BMCT 6

Forms S and T were originally published in 1969. Prior to this, a number of BMCT forms had been published in an effort to provide tests of appropriate content and difficulty for different applications and testing situations. In 1940, The Psychological Corporation (now Pearson Education, Inc.) published the first of these, Form AA. A second form, BB, was copyrighted in 1941 and was appreciably more difficult than Form AA. Form W1, developed using more generic situations and fewer technical terms, appeared in 1942. In addition, a number of restricted versions of the test were prepared under contract with the military during mobilization for World War II. The Psychological Corporation published a variant of the BMCT, Form CC, in 1949. This form, developed by W. A. Owens and the Bureau of Naval Personnel, had illustrations more nearly resembling mechanical drawings and presented examinees with five, rather than three, response alternatives.

When the *Differential Aptitude Tests* (DAT; Bennett, Seashore, & Wesman) were prepared in 1947, a test of mechanical comprehension was included in the battery. The DAT Mechanical Reasoning Test was created to parallel BMCT Form AA and, though the content differs, the item type and general subject matter clearly identify it as an alternate form of the BMCT. Data on the Mechanical Reasoning test can be found in the *Differential Aptitude Tests for Personnel and Career Assessment Technical Manual* (Bennett, Seashore, & Wesman, 1991).

Development of Forms S and T

The complete revision of the BMCT began in 1966. Four primary objectives guided this revision effort:

- develop a pair of alternate forms to replace Forms AA, BB, and W1;
- increase the range of item difficulty and maintain test reliability;
- replace dated illustrations with current ones; and
- simplify scoring by eliminating the "correction for guessing."

Development of Forms S and T began with careful review of items in Forms AA, BB, and W1. Ninety-five items were retained in their original form, 43 items were changed in either depiction or wording, and 42 new items were developed, totaling 180 items. These items were distributed among three tryout forms administered to 706 male students in grades 11 and 12. Approximately one-third of the sample took any given form. Item analyses conducted on the resulting data indicated that 136 items had acceptable psychometric properties. These items were distributed across 18 categories of mechanical principles (see Table 6.1).

Table 6.1 Categorization of Form S and Form T Items

Category	Number of Items
Acoustics	3
Belt Drive	2
Center of Gravity	7
Centrifugal Force	5
Electricity	6
Gears	10
Gravity and Velocity	9
Heat	8
Hydraulics	16
Inertia	5
Levers	8
Optics	6
Planes and Slopes	2
Pulley Systems	10
Resolution of Forces	6
Shape and Volume	7
Structures	12
Miscellaneous	14
Total	136

The standardization versions of the forms were constructed by distributing the number of items as evenly as possible across categories. To maintain the equivalency of item difficulty across the two forms, the cumulative sum of item difficulty values was kept the same. The final item difficulty and discrimination indices used for Forms S and T are presented in Table 6.2.

Table 6.2 Summary of Item Difficulty and Discrimination in Forms S and T (Bennett,1969)

	Difficulty Index (*p* value)			Discrimination Index (item-test point biserial coefficient)		
Form	Minimum	Maximum	Mean	Minimum	Maximum	Mean
S	.17	.96	.62	.20	.51	.32
T	.19	.95	.62	.21	.55	.33

Updated Artwork

The artwork for Forms S and T was updated in 2005. The goal of this effort was to increase face validity of the items and to ensure that the items were contemporary in appearance, that diversity was better represented in the depiction of people, and that the concepts and difficulty level of the original items were maintained. For most items, relatively minor cosmetic changes were made (e.g., changing the picture of an older-style car to a newer-style SUV); however, items 30, 44, and 68 in Form S were substantially changed based on user feedback.

Item 30 originally presented a depiction of a cannon being fired. Current test users indicated that the item had inappropriate connotations for an employment test. The picture was replaced with one of lightning and thunder to present

the same concept. Similarly, users found the depiction of a bomb being released from a plane in Item 44 as violent and inappropriate. A picture of a cargo crate being dropped from the plane was substituted for the bomb. Item 68 originally depicted an outdated refrigerator and required the examinee to make a judgment regarding the air temperature in different areas in the refrigerator. Test users suggested that modern refrigerators are more efficient at keeping a uniform temperature throughout the unit. Therefore, the item was changed to a picture of a kitchen to present the concept that was originally intended.

Evidence of reliability and validity of the updated forms provides support that the revised art did not significantly change the psychometric properties of the instruments. Additional support for the equivalence of the previous depictions versus the updated art is provided later in this chapter.

Equivalence of Forms

Previous Studies of Equivalence of Form S to Form T

To support the equivalence of Form S to Form T, a study was conducted with 302 applicants for skilled trade jobs at an automobile company (Bennett, 1969). First, all applicants completed Form BB of the BMCT, and then completed either Form S or Form T. Frequency distributions of Form BB raw scores were generated separately for those applicants who completed Form S and for those who took Form T. Cases from each distribution were then selected so that distributions of Form BB scores for both groups were the same. For these comparable groups, distributions of raw scores on Form S and Form T were compared. The results of this comparison indicated that Forms S and T may be considered equivalent alternate forms; while the means differed by 0.5 raw score units and the standard deviations differed by 1.4 units, the differences were not statistically significant (see Table 6.3).

Table 6.3 Comparison of Forms S and T Scores

Form	*N*[1]	Mean	*SD*
S	151	45.2	8.6
T	151	45.7	10.0

[1] Applicants for skilled trade jobs at an automobile company.

Current Studies of Equivalence of Form S to Form T

To provide additional evidence of the equivalence of Form S to Form T, Pearson conducted an equivalency study in 2006. The study consisted of 225 adult participants from a variety of occupations, and test administration was counter-balanced. Approximately half of the group (n = 111) completed Form T followed by Form S, while the other participants (n = 103) completed the tests in the reverse order. Table 6.4 presents means, standard deviations, and correlations obtained from an analysis of the resulting data. As indicated in the table, mean score differences between forms were less than two points (1.5 and 1.6). The variability of scores also was similar, with standard deviations ranging from 9.1 to 11.1. The coefficients indicate that scores on the two forms are highly correlated (.88 and .81). The high correlations provide further support that the raw scores from one form may be interpreted as having the same meaning as identical raw scores on the other form.

Overall, the difference in mean scores between the forms was statistically small (d' = .00). This difference (d'), proposed by Cohen (1988), is useful as an index to measure the magnitude of the actual difference between two means. Difference (d') values below .20 indicate a small effect size (Cohen, 1988). 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 (1996) Formula 10.4.

Table 6.4 Equivalency of Form S and Form T

		Form T		Form S		
Administration Order	***N***	**Mean**	**SD**	**Mean**	**SD**	***r***
Form T Followed by Form S	111	44.4	11.1	46.0	10.0	.88
Form S Followed by Form T	103	47.7	9.1	46.1	9.5	.81

Equivalence of Computer-Based and Paper-and-Pencil Versions of the BMCT

Studies of the effect of the administration medium have generally supported the equivalence of paper-and-pencil and computerized versions of non-speeded cognitive ability tests (Mead & Drasgow, 1993). To ensure that these findings held true for the BMCT, Pearson conducted an equivalency study using paper-and-pencil and computer-administered versions of Form S.

In this study, a group of 225 adult participants from a variety of occupations was administered Form S with paper-and-pencil and on computer. Approximately half of the group (n = 105) completed the paper Form S followed by the online version of Form S, while the other participants (n = 120) completed the tests in the reverse order. Table 6.5 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 approximately one point (1.1 and 1.2). The difference in mean scores between the paper version and the online version was statistically small (d' = 0.17). The variability of scores also was similar, with standard deviations ranging from 8.8 to 12.0.

The results shown in Table 6.5 indicate that paper-and-pencil raw scores correlate very highly with online administration raw scores (.95 and .90). The high correlations provide further evidence that raw scores from one mode of administration (paper or online) may be interpreted as having the same meaning as identical raw scores from the other mode of administration. Because the study included the previous artwork on the paper version and the updated artwork on the online version, this study provides additional support for the equivalency of the previous artwork and the updated artwork.

Table 6.5 Equivalency of Paper and Online Modes of Administration

		Paper		Online		
Administration Order	*N*	Mean	SD	Mean	SD	*r*
Paper Followed by Online	105	43.5	11.3	42.4	12.0	.95
Online Followed by Paper	120	47.6	9.3	48.8	8.8	.90

Differences in Mechanical Comprehension Test Scores Based on Sex

It is consistently reported in the literature on cognitive abilities that females perform better than males on tests of verbal ability, though males perform better than females on visual-spatial and mathematical ability tests (Halpern, 1986). The results of studies investigating differences in mechanical comprehension test scores based on the examinee's sex have been consistent with these findings (Bennett & Cruikshank, 1942; deWolf, 1981; Lunneborg & Lunneborg, 1985; McCall, 1973). Although many studies indicate that sex-related differences on aptitude tests are diminishing, differences in mechanical aptitude are still commonly found to be of sufficient magnitude to be statistically significant (Feingold, 1988; Sapitula & Shartzer, 2001).

In a recent review of the literature on gender differences in cognitive ability, Spelke (2005) stated that the gender difference in standardized test performance "likely reflects a complex mix of social, cultural, and biological factors" (p. 956). In a study conducted by Fortson (1991), a multiple regression analysis was used to explain score variance between sexes in BMCT scores. When education factors, work experience factors, and leisure factors were entered into the equation, gender explained only 2% additional unique variance. Regardless of the explanations for group differences in performance on cognitive ability tests, the literature indicates that these differences are consistent across tests and situations. The origins of observed differences in cognitive ability test scores based on sex are external to the test, assuming that the test is reliable and the construct valid.

Fairness of Using the BMCT With Female Applicants

"A lower mean test score in one group compared to another is not by itself evidence of bias" (Guion, 1998, p. 436). Group differences in mean test scores can represent actual differences in the ability being measured, as is the case with the BMCT. A more appropriate indicator of whether an ability test is biased is to evaluate whether the test is unequally predictive of job performance based on group membership (i.e., differential prediction).

In a study conducted for the revision of this manual in 2006, Pearson evaluated evidence of validity based on the relationship between the BMCT scores and on-the-job performance for 53 female job incumbents in various occupations and industries that require mechanical comprehension. *Job performance* was defined as the sum of supervisory ratings on 20 behaviors determined to be important to most jobs that require mechanical aptitude, as well as ratings on single-item measures of "Overall Performance" and "Overall Potential." Results of this study indicated that the criterion-related validity coefficients for the female sample

were comparable to, if not higher than, those obtained for the predominantly male samples provided in chapter 8. A strong relationship between BMCT scores and job performance was observed in the sample of females, as evidenced by high correlations with the sum of job performance ratings (uncorrected $r = .38$, $p < .05$, $n = 29$), ratings of overall job performance (uncorrected $r = .36$, $p < .05$, $n = 53$), and ratings of overall potential (uncorrected $r = .41$, $p < .05$, $n = 52$).

Adverse Impact

Users of cognitive ability tests need to recognize that adverse impact may occur. This situation is not desirable; however, assessment instruments measuring constructs that are job-related and consistent with business necessity may be used legally in employee selection (Equal Employment Opportunity Commission, 1978). The most stringent method of establishing job-relatedness is criterion-related validation; that is, statistically establishing that scores on a selection instrument predict successful performance of the job in question. With a measure such as the BMCT, a content validity argument may also be used to establish job-relatedness. Because the BMCT measures a specific aptitude (i.e., mechanical comprehension), it can be linked logically to essential functions of a job that have been identified with a job analysis instrument.

Evidence of Reliability 7

The *reliability* of a measurement instrument refers to the accuracy, consistency, and precision of test scores across situations (Anastasi & Urbina, 1997) and the confidence that may be placed in those results. Test theory posits that a test score is an estimate of an individual's hypothetical *true score,* or the score an individual would receive if the test were perfectly reliable. The reliability of a test is expressed as a *correlation coefficient,* which represents the consistency of scores that would be obtained if a test could be given an infinite number of times. In actual practice, administrators do not have the luxury of administering a test an infinite number of times, so some measurement error is to be expected. A reliable test has relatively small measurement error.

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 to .89 is "good," .70 to .79 is considered "adequate," and below .70 "may have limited applicability."

The methods most commonly used to estimate test reliability are internal consistency of the test items (e.g., *Cronbach's alpha coefficient,* Cronbach 1970), test-retest (the stability of test scores over time), and alternate forms (the consistency of scores across alternate forms of a test). The reliability of a test should always be considered in the interpretation of obtained test scores.

Standard Error of Measurement

The greater the reliability of a test, the smaller the standard error of measurement. Because the true score is a hypothetical value that can never be obtained because testing always involves some measurement error, an examinee's score 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.

Repeated testing always results in some variation, so no single test event ever measures an examinee's actual ability with complete accuracy. Therefore, it is necessary to estimate the 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 estimate of error is the *standard error of measurement* (*SEM*).

The *SEM* is a quantity that is added to and subtracted from an examinee's test score, creating a confidence interval or band of scores around the obtained score. The confidence interval is a score *range* within which the examinee's hypothetical true score lies, and represents the examinee's actual ability. A true score is a theoretical score entirely free of error. A small *SEM* denotes higher reliability of a test, and a large *SEM* denotes less reliable measurement and less reliable scores.

The standard error of measurement is calculated with the formula:

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

In this formula, *SEM* represents the standard error of measurement, *SD* represents the standard deviation unit, and r_{xx} represents the reliability coefficient of the scale. Approximately 68% of the time, the observed score lies within +1.0 and –1.0 *SEM* of the true score; 95% of the time, the observed score lies within +1.96 and –1.96 *SEM* of the true score; and 99% of the time, the observed score lies within +2.58 and –2.58 *SEM* of the true score. For example, with a *SEM* of 2, if a candidate obtains a BMCT raw score of 20, you can be 68% confident that his or her true score lies within the range of 18 to 22. Thinking in terms of score ranges serves as a check against overemphasizing small differences between scores. One general rule 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).

Internal Consistency

Previous Studies of Internal Consistency Reliability

The internal consistency of Forms S and T was evaluated with split-half coefficients corrected for the full length of the test, using the Spearman Brown formula (Bennett, 1969). Table 7.1 presents reliability coefficients and standard errors of measurement (*SEM*) for the BMCT. In Table 7.1, a description of the group for which data are reported is followed by the test form administered, the size (*N*) of the group, the average (Mean) score, the standard deviation (*SD*) of scores for the group, the *SEM* for the group, and the split-half reliability coefficient (r_{xx}) for the group. The data for Forms S and T indicate a range of reliability coefficients from .81 to .93 with a median value of .86, while the *SEM* ranges from 3.0 to 3.8.

Table 7.1 Means, Standard Deviations (*SD*), Standard Error of Measurement (*SEM*), and Split-Half Reliability Coefficients (r_{xx}) of the BMCT (Bennett, 1969)

Group	Group/Grade	Form	*N*	Mean	*SD*	*SEM*[1]	r_{xx}[2]
Applicants for process training jobs at an oil refinery		S and T	100	44.2	10.5	3.2	.91
Applicants for skilled trades jobs at an automobile company		S and T	435	46.0	9.5	3.4	.87
Applicants for a union apprentice training program in the construction trades		S and T					
	Group 1		271	42.7	8.7	3.5	.84
	Group 2		204	43.6	9.0	3.5	.85
Academic high school students		S and T					
	Grade 11		128	36.9	8.6	3.8	.81
	Grade 12		109	37.4	11.3	3.0	.93
Technical high school students		S and T					
	Grade 11		254	40.6	9.6	3.3	.88
	Grade 12		85	42.2	8.7	3.8	.81

[1] Computed by the formula $SEM = SD\,(1-r_{xx})$.
[2] Odd-even coefficients corrected by the Spearman-Brown formula.

Current Studies of Internal Consistency Reliability

Cronbach's alpha and the standard error of measurement (*SEM*) were calculated for the current norm groups (see Table 7.2). Reliability estimates for these samples were similar to those found in previous studies and range from .84 to .92. Consistent with previous research, these values indicate that BMCT scores possess adequate reliability. In addition, internal consistency reliability estimates were calculated separately for the two forms in an overall sample of examinees who took the BMCT from 2003 to 2005. *The internal consistency reliability coefficients were .90 for both Form S* (n = 3,725) *and Form T* (n = 754).

Table 7.2 Means, Standard Deviations (*SD*), Standard Error of Measurement (*SEM*) and Internal Consistency Reliability Coefficients (r_{alpha})

Group	*N*	Mean	*SD*	*SEM*	r_{alpha}
Occupation					
Automotive Mechanic	95	46.3	11.2	3.2	.92
Engineer	105	52.7	8.7	3.0	.88
Installation/Maintenance/Repair	112	49.1	8.4	3.1	.86
Industrial/Technical Sales Representative	133	47.4	7.9	3.2	.84
Skilled Tradesperson (e.g., electrician, welder, carpenter)	153	47.4	8.9	3.2	.87
Transportation Trades/Equipment Operator (e.g., truck driver, heavy equipment operator)	126	42.8	9.4	3.4	.87
Combined Occupation					
Mechanical Trades (e.g., mechanic, installation/maintenance/repair, electrician, welder, carpenter)	388	47.5	9.4	3.3	.88
Automotive and Aircraft Mechanic	123	46.2	10.7	3.2	.91
Industry					
Manufacturing/ Production	580	43.8	10.5	3.3	.90
Energy/Utilities	135	48.1	9.2	3.3	.87

Evidence of Test-Retest Reliability

Test-retest reliability was evaluated for Form S in a sample of job incumbents representing various occupations and industries (N = 74). The test-retest intervals ranged from 1 to 29 days, with a mean interval of 10.6 days. As the data in Table 7.3 indicate, scores on Form S demonstrate acceptable test-retest reliability (r_{12} = .93). The differences in mean scores between the first testing and the second testing on Form S are statistically small (d' = .15). The 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.

Table 7.3 Test-Retest Reliability of the BMCT

	First Testing		Second Testing			
BMCT	**Mean**	***SD***	**Mean**	***SD***	r_{12}	**Difference (d′)**
Form S	50.4	10.5	52.0	10.6	.93	.15

N = 74

Effects of Training and Practice

Scores on the BMCT may be influenced by environmental factors, but not to the extent that important difficulties in interpretation are introduced. Research indicates that, while an individual's scores on the BMCT can generally be improved through training and experience, it is unlikely that dramatic improvement will result (Bennett, 1940). In a study involving applicants for technical defense courses, the average score on the BMCT for individuals reporting previous training in physics was 41.7 ($N = 220$, $SD = 8.6$), while the average score for individuals reporting no previous physics training was 39.7 ($N = 95$, $SD = 8.9$). Similarly, candidates for positions as fire fighters or police officers who reported previous physics training had an average BMCT score of 39.6 ($N = 488$, $SD = 9.6$), while candidates reporting no previous physics training had an average score of 34.0 ($N = 983$, $SD = 9.8$). Point-biserial correlations calculated using data from this sample revealed that physics training was moderately related to years of education ($r = .26$). Further analyses of these data indicated that the effect on BMCT test scores as a result of prior training in physics was to raise the mean by four raw score points, or less than one-half the standard deviation (Bennett, 1969).

Evidence of Validity 8

The validity of a test is the single most fundamental and important aspect of test development and evaluation. Validity refers to the degree to which specific evidence supports the interpretation of test scores for their intended purpose (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). Although test developers are responsible for providing initial evidence of validity, the test user must evaluate whether the evidence supports the use of the test for his or her intended purpose.

Evidence of Validity Based on Test Content

Evidence of content validity is based on the degree to which test items adequately represent and relate to the trait being measured. Test content also includes the wording and formatting of items, and the procedures for administering and scoring the test. Evaluation of content-related evidence is usually a rational, judgmental process (Cascio & Aguinis, 2005). 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). For the BMCT, evidence based on test content should be established by demonstrating that the jobs for which the test will be used require the mechanical aptitude measured with the BMCT. Content-related evidence of the BMCT in training and instructional settings may be demonstrated by the extent to which the BMCT measures a sample of the specified objectives of such instructional programs.

Evidence of Validity Based on Test-Criterion Relationships

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.

Evidence of criterion-related validity refers to the statistical relationship between scores on the test and one or more criteria (e.g., performance ratings, grades in a training course, productivity measures). By collecting test scores and criterion scores, one can determine how much confidence may be placed in the use of test scores to predict job success. Typically, correlations between criterion measures and test scores 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 with a uniform criterion measure. The criterion itself should be reliable and job-relevant, and should provide a wide range of scores. 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 to .35 are considered "likely to be useful," .11 to .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 reasoning is based on the principle that the practical value of a test depends not only on validity but on other factors as well, 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 reviewed. 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.

Studies have generally shown strong support for the validity of the BMCT, based on evidence of test-criterion relationships. Kolz, McFarland, and Silverman (1998) reported that, in a study of 176 manufacturing employees, BMCT scores correlated .40 with supervisory ratings of mechanical understanding, and .27 with supervisory ratings of problem solving ability.

Muchinsky (1993) evaluated the relationships between the BMCT, a general mental ability test, and an aptitude classification test focused on mechanics, and supervisory ratings of overall performance for 192 manufacturing employees. Of the three tests, he found the BMCT to be the best single predictor of job performance ($r = .38$, $p < .01$). He also found that the incremental gain in predictability from the other tests was not significant.

Callender and Osbourn (1981) reported that the estimated true validity coefficients for the BMCT were .31 (*SD* = .17) for job performance criteria and .52 (*SD* = .07) for training performance criteria. These correlations were based on a sample of 38 job performance and 11 training performance validity studies. A wide variety of jobs were represented in the Callender and Osburn study, including maintenance crafts, process operators, and laboratory technicians.

Schmidt, Hunter, and Caplan (1981) found an estimated true score correlation of .33 between the BMCT and overall job/training performance for both operator and maintenance positions within the petroleum industry. Their findings were based on data for 1,800 operators and 706 maintenance personnel from 13 different organizations.

In studies conducted for the revision of this manual in 2006, the relationship between BMCT scores and on-the-job performance of incumbents in various occupations, organizations, and industries was examined. Job performance was defined as supervisory ratings on behaviors determined through research to be important to jobs requiring mechanical reasoning ability (e.g., "demonstrates mechanical knowledge and expertise"). Supervisory ratings on a single-item measure of job potential were also obtained. The study found that BMCT scores correlated with job performance for all groups. Specifically, uncorrected correlations with supervisory ratings of mechanical understanding/problem solving ranged from .32 (manufacturing employees) to .50 (transportation trades/equipment operators). Furthermore, in four of the six groups, BMCT scores correlated significantly with supervisory ratings of potential (uncorrected r = .29 to .45).

Table 8.1 presents criterion-related validity evidence for BMCT Forms S and T. The table includes statistical information for the studies described previously in this chapter, as well as additional studies. Only studies that reported validity coefficients are shown.

The first column of the table gives a description of each sample. The column entitled *N* shows the number of cases in the sample. The form of the BMCT administered is then noted, followed by the mean score and standard deviation. The criterion measures include training course grades and supervisor ratings, among others. Means and standard deviations are shown for the criterion measures. The validity coefficient for the sample appears in the last column.

Table 8.1 Studies Showing Evidence of Criterion-Related Validity

		BMCT			Criterion			
Group	***N***	**Form**	**Mean**	***SD***	**Description**	**Mean**	***SD***	***r***
Manufacturing employees at a large Midwestern manufacturing company (Kolz, McFarland, & Silverman, 1998)	176	S/T	—	—	**Supervisor's Ratings:**			
					Mechanical Understanding	5.4	2.1	.40**
					Problem Solving	5.3	1.9	.27**
Manufacturing employees (e.g., machine operators, maintenance mechanics, quality control inspectors) at two plants (Muchinsky, 1993)					**Supervisor's Ratings:**			
	96	S	45.8	8.4	Overall Performance (Plant 1)	139.8	28.2	.40**
	97	S	48.4	8.6	Overall Performance (Plant 2)	141.2	30.1	.33**
	—	—	—	—	Overall Performance (Plants Combined)	—	—	.38**
Job incumbents across multiple jobs (e.g., maintenance crafts, process operators, laboratory technicians) and industries (Callender & Osburn, 1981)	—	—	—	—	Training Performance	—	—	.52**
					Job Performance	—	—	.31**
Operator positions within the petroleum industry (Schmidt, Hunter, & Caplan, 1981)	1,800	—	—	—	Job/Training Performance	—	—	.33**
Maintenance positions within the petroleum industry (Schmidt, Hunter, & Caplan, 1981)	706	—	—	—	Job/Training Performance	—	—	.33**
Machine operators at a large paper products manufacturer (Pearson Education, Inc. 2006)	32–41	S/T	43.0	8.0	**Supervisor's Ratings:**			
					Mechanical Understanding/ Problem Solving	93.1	22.6	.45**
					Potential	2.3	1.1	.29
Job incumbents across multiple jobs (e.g., electricians, truck drivers, equipment operators, mechanics, mining laborers) from the western site of a large mining company (Pearson Education, Inc., 2006)	58–68	S/T	45.7	9.5	**Supervisor's Ratings:**			
					Mechanical Understanding/ Problem Solving	4.3	0.9	.41**
					Potential	1.9	0.8	.32*

(continued)

Table 8.1 Studies Showing Evidence of Criterion-Related Validity *(continued)*

		BMCT			Criterion			
Group	***N***	**Form**	**Mean**	***SD***	**Description**	**Mean**	***SD***	***r***
Job incumbents (primarily engineers) from a large chemical engineering company (Pearson Education, Inc., 2006)	40–42	S/T	53.0	7.7	**Supervisor's Ratings:**			
					Mechanical Understanding/ Problem Solving	42.7	7.8	.32*
					Potential	2.4	0.9	.40**
Transportation trades/Equipment operators (Pearson Education Inc., 2006)	52–58	S/T	43.6	8.8	**Supervisor's Ratings:**			
					Mechanical Understanding/ Problem Solving	4.4	1.1	.50**
					Potential	2.3	1.1	.45*
Manufacturing employees (e.g., equipment operators) (Pearson Education, Inc., 2006)	42–48	S/T	40.3	13.3	**Supervisor's Ratings:**			
					Mechanical Understanding/ Problem Solving	4.8	1.0	.32*
					Potential	2.4	1.1	.29*
Energy/Utilities employees (e.g., maintenance supervisors) (Pearson Education, Inc. 2006)	38–40	S/T	43.8	9.9	**Supervisor's Ratings:**			
					Mechanical Understanding/ Problem Solving	4.8	1.0	.34*
					Potential	2.6	1.1	.01
College students in a work sample laboratory study (Mount, Muchinsky, & Hanser, 1977)	20	S/T	38.8	10.7	Work sample	30.9	7.7	.58**
					Criterion measure	62.8	13.4	.55*
	20	S/T	36.1	11.5	Work sample	27.3	7.4	.51*
					Criterion measure	56.6	11.5	.62**
Operators at a Southern chemical plant (Bennett, 1994)	87	S/T	45.1	7.0	**Supervisor's Ratings:**			
					Reaction to emergency	8.0	1.5	.36**
					Safety rules	7.4	1.0	.21
					Mechanical ability	7.1	1.4	.32**
					Job knowledge	7.7	1.7	.39**
Operators at a Southern chemical plant (Bennett, 1994)	136	S/T	48.0	7.5	Job knowledge test	64.0	12.8	.63**
Trainees at an Eastern refinery and chemical plant (Bennett, 1994)	54	S/T	51.8	6.9	Training course grade	—	—	.48**

(continued)

Table 8.1 Studies Showing Evidence of Criterion-Related Validity *(continued)*

		BMCT			Criterion			
Group	***N***	**Form**	**Mean**	***SD***	**Description**	**Mean**	***SD***	***r***
Mechanics at a Northeastern utility company (Bennett, 1994)	35	S/T	47.0	10.4	Weighted criterion score (work samples & trade information test)	50.0	10.0	.64**
Apprentices at a major domestic steel producer (Bennett, 1994)	30	S/T	50.0	6.8	Average course grade	88.2	5.3	.54**
Knitting machine mechanics at a Southern textile plant (Bennett, 1994)	32	S/T	41.2	8.8	Supervisor's ratings	31.2	5.3	.37*
Experienced coal miners at a Midwestern underground coal mining company (Bennett, 1994)	83	S/T	48.5	9.0	Combined rating & ranking index	93.6	16.6	.23*
Inexperienced coal miners at a Midwestern underground coal mining company (Bennett, 1994)	178	S/T	50.9	8.1	Combined rating & ranking index	86.9	15.5	.23**
Technician trainees at a Northeastern utility (Bennett, 1994)	83	S/T	39.6	9.7	Speed in solving training modules	228.1	73.8	.52**
					Number of modules completed	4.4	0.8	.40**
Technical assistants at a Northeastern electrical utility (Bennett, 1994)	29	S/T	51.8	7.1	Weighted criterion score derived from performance on specific tasks & trade information test	50.0	10.0	.49**
Equipment operators at a Northeastern utility (Bennett, 1994)	31	S/T	51.6	7.0	Weighted criterion score derived from performance on specific tasks & trade information test	50.0	10.0	.39*
Customer service trainees at a Northeastern electrical utility (Bennett, 1994)	26	S/T	50.5	6.9	Special training exam	18.7	5.3	.56**
	35	S/T	51.1	7.5	Training exams	13.6	—	.41*

* $p < .05$.

** $p < .01$. The studies by Callender and Osburn (1981) and Schmidt, Hunter, and Caplan (1981) were meta-analyses. With the exception of the two meta-analyses, all correlation coefficients reported are uncorrected. Pearson (2006) studies reported for manufacturing, energy, and transportation trades employees include some participant overlap with Pearson (2006) organization-specific studies.

Criterion-related validity information for BMCT forms developed prior to S and T (i.e., Forms AA and BB) can be found in the 1994 BMCT Manual. Criterion-related validity information for the Differential Aptitude Tests' Mechanical Reasoning subtest (DAT MR), a statistically equivalent form of the BMCT, may be found in the *Fifth Edition Manual for the Differential Aptitude Tests, Forms S and T* (The Psychological Corporation, 1974), the *Differential Aptitude Tests for Personnel and Career Assessment Technical Manual* (Bennett, Seashore, & Wesman, 1991), and the *Differential Aptitude Tests Fifth Edition Technical Manual* (The Psychological Corporation, 1992).

Published validity coefficients such as those reported in Table 8.1 apply to the specific samples listed. Test users should not automatically assume that these data constitute sole and sufficient justification for use of the BMCT. Inferring validity for one group from data reported for another group is not appropriate unless the organizations and jobs being compared are demonstrably similar.

Careful examination of Table 8.1 can help test users make an informed judgment about the appropriateness of the BMCT for their organization. However, the data presented here are not intended to serve as a substitute for locally obtained data. Locally conducted validity studies, together with locally derived norms, provide a sound basis for determining the most appropriate use of the BMCT. Whenever technically feasible, test users should study the validity of the BMCT, or any selection test, at their own location.

Validity coefficients such as those reported in Table 8.1 apply to the specific samples listed. Sometimes it is not possible for a test user to conduct a local validation study. There may be too few incumbents in a particular job, an unbiased and reliable measure of job performance may not be available, or there may not be a sufficient range in the ratings of job performance to justify the computation of validity coefficients. In such circumstances, evidence of a test's validity reported elsewhere may be relevant, provided that the data refer to comparable jobs.

Evidence of Convergent and Discriminant Validity

Evidence of convergent validity is demonstrated when scores on a test relate to scores on other tests or variables that purport to measure similar traits or constructs. Evidence of relations with other variables can involve experimental (or quasi-experimental) as well as correlational evidence (AERA et al., 1999). Evidence of discriminant validity is demonstrated when scores on a test do not relate closely to scores on tests or variables that measure different traits or constructs.

Evidence of convergent validity for the BMCT has been demonstrated in studies that examined its relationship with other mechanical ability tests. In 2006, Pearson Education conducted a study of the relationship between the BMCT and the Mechanical Reasoning subtest of the *Differential Aptitude Tests for Personnel and Career Assessment* (DAT MR; Bennett, Seashore, & Wesman, 1991). The DAT MR originally was designed as a statistically parallel version of the BMCT. The study consisted of 107 individuals employed in various roles and industries. As expected, BMCT scores correlated strongly ($r = .85$, $p < .01$) with DAT MR scores.

Other studies have focused on correlations between BMCT scores and scores on cognitive ability tests without a specific mechanical component (e.g., verbal and quantitative abilities tests). These studies have typically reported a pattern of correlations that supports both the convergent and discriminant validity of the BMCT. That is, they generally have found that BMCT scores correlate higher with abilities that are more conceptually related (e.g., spatial ability) than unrelated (e.g., spelling ability). The largest of these studies focused on correlations of DAT MR (the statistically parallel version of the BMCT), with other abilities measured within the *Differential Aptitude Test* battery. This series of eight studies included sample sizes of 2,790 to 11,120 students, and so provides relatively stable estimates of the relationships among these abilities. Results indicated that Mechanical Reasoning scores correlated highly with Spatial Reasoning (r = .56 to .62) and Abstract Reasoning (r = .57 to .60) scores; moderately with Verbal Reasoning (r = .48 to .53), Numerical Reasoning (r = .40 to .50), and Language Usage (r = .31 to .43) scores; and low with Spelling (r = .15 to .25) and Perceptual Speed and Accuracy (r = .02 to .12) scores.

Table 8.2 presents correlations between the BMCT and other tests. Additional studies are reported in the previous version of the BMCT manual (Bennett, 1994). A description of the study participants appears in the first column on the left. The second column lists the total number of participants (*N*), followed by the BMCT form for which data were collected, the mean and standard deviation (*SD*) of BMCT scores, and the comparison test name. The mean and standard deviation of scores on the comparison test are reported next, followed by the correlation coefficient (*r*) indicating the relationship between scores on the BMCT and the comparison test. In general, higher correlation coefficients indicate a higher degree of overlap between the construct measured by the BMCT and the construct measured by the comparison test. Likewise, lower correlation coefficients generally indicate less overlap between the construct measured by the BMCT and the construct measured by the comparison test.

Table 8.2 BMCT Convergent Validity Evidence

		BMCT			Other Test			
Group	***N***	**Form**	**Mean**	***SD***	**Description**	**Mean**	***SD***	***r***
Job incumbents across multiple jobs and industries	107	S/T	44.4	10.5	Mechanical Reasoning subtest of the Differential Aptitude Tests for Personnel and Career Assessment	31.9	7.7	.85**
Adults aged 16 to 54 (Kellogg & Morton, 1999)	28	S	45.8	11.1	Beta III	109.8	13.6	.46*
College students at a Southwestern university (Lowman & Williams, 1987)	149	S/T	41.8	6.1	Revised Minnesota Paper Form Board Test	43.5	9.1	.49**
					Watson-Glaser Critical Thinking Appraisal	56.1	8.9	.39**
					Raven's Standard Progressive Matrices	50.8	5.0	.50**
					Wide-Range Achievement Test, Arithmetic	27.1	4.9	.37**
Manufacturing employees at two plants (e.g., machine operators, maintenance mechanics, quality control inspectors) (Muchinsky, 1993)	192	S	45.8–48.4	8.4–8.6	Mechanics subtest of the Flanagan Aptitude Classification Test	7.4–8.2	2.5–3.1	.64**
					Thurstone Language	28.3–31.6	8.7–9.5	.33**
					Thurstone Quantitative	21.6–2.2	6.4–7.2	.37**
College students in a work sample laboratory study (Mount, et al. 1977)	20	S/T	38.8	10.7	Wonderlic Personnel Test	30.9	7.6	.62**
	20	S/T	36.1	11.5	Wonderlic Personnel Test	31.8	5.7	.46*

(continued)

Table 8.2 BMCT Convergent Validity Evidence *(continued)*

		BMCT			Other Test			
Group	***N***	**Form**	**Mean**	***SD***	**Description**	**Mean**	***SD***	***r***
Parents of mathematically gifted students (Benbow, Stanley, Kirk, & Zonderman, 1983)	43 (males)	S/T	40.0	11.0	Concept Mastery Test	118.0		.48**
					California Test of Mental Maturity, Language	24.1		.53**
					Kirk's Synonym-Antonym Test	35.0		.30
					Kirk's General Information Test	25.0		.67**
					Kirk's Test of Semantic Comprehension	15.0		.50**
					Kirk's Cubes Test	15.0		.53**
					Kirk's Rotation-Inversion Test	12.0		.69**
	45 (females)	S/T	26.0	11.0	Concept Mastery Test	120.0		.46**
					California Test of Mental Maturity, Language	23.5		.66**
					Kirk's Synonym-Antonym Test	35.0		.40**
					Kirk's General Information Test	23.0		.67**
					Kirk's Test of Semantic Comprehension	14.0		.51**
					Kirk's Cubes Test	11.0		.49**
					Kirk's Rotation-Inversion Test	12.0		.64**

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

Using the BMCT as an Employment Assessment Tool

9

BMCT Forms S and T were developed for adult assessment applications. In employee selection, the BMCT may be used to predict success in certain industrial, mechanical, and engineering jobs. It is also useful in monitoring the effectiveness of mechanical comprehension instruction and training programs, and in researching the relationship between mechanical comprehension and other abilities or traits.

Employment Selection

Many organizations use testing as a component of their employment selection process. Employment selection test programs typically use cognitive ability tests, aptitude tests, personality tests, basic skills tests, and work values tests to screen out unqualified candidates, to categorize prospective employees according to their probability of success on the job, or to rank order a group of candidates according to merit.

The BMCT has been used in the assessment of applicants for a wide variety of jobs, including electrical and mechanical positions in metals plants and aviation companies, skilled trades in automobile companies, computer operators, mechanics and equipment operators at utility companies, and operators in manufacturing. The BMCT may also be used effectively with other tests to evaluate candidates for jobs that require aptitudes other than mechanical comprehension. For example, if the job requirements involve the use of technical manuals or textbooks, a reasonably demanding mental ability test, such as the *Miller Analogies Test for Professional Selection* (Pearson Education 2005), may also be useful. When job requirements involve critical and analytical thinking skills (e.g., for management-level positions), an instrument such as the *Watson–Glaser Critical Thinking Appraisal* (Watson & Glaser, 2005) may provide valuable supplementary information. Where manipulative skill is important, measures such as the *Bennett Hand–Tool Dexterity Test* (Bennett, 1981) or the *Crawford Small Parts Dexterity Test* (Crawford & Crawford, 1981) may be useful.

Do not assume that the type of mechanical comprehension required in a particular job is identical to that measured by the BMCT. Job analysis and validation of the BMCT for selection purposes should follow accepted human resource research procedures and conform to existing guidelines concerning fair employment practices. Local validation is particularly important when a selection test may have adverse impact; this is likely to occur with cognitive ability tests such as the BMCT (refer to the section in chapter 6, "Differences in Mechanical Comprehension Test Scores Based on Sex"). While it is not unlawful to use a test with adverse impact (Equal Employment Opportunity Commission, 1978), the

testing organization must be prepared to demonstrate that the instrument is job-related and consistent with business necessity. If use of the test is challenged, the organization must provide historical evidence of the test's validity in similar situations, evidence of content validity, or evidence of locally obtained criterion-related validity.

Making a Hiring Decision Using the BMCT

Human resource professionals can look at the percentile that corresponds to the candidate's raw score in several ways. Candidates' scores may be rank ordered by percentile so that those with the highest scores are considered further. Alternatively, a cut score (e.g., the 50th percentile) may be established so that candidates who score below the cut score are not considered further. In general, the higher the cut score is set, the higher the probability that a given candidate who scores above that point will be successful. However, the need to select high scoring candidates typically needs to be balanced with situational factors, such as the need to keep jobs filled and the supply of talent in the local labor market. Ultimately, it is the responsibility of the hiring authority to determine how it uses the BMCT scores. If the hiring authority establishes a cut score, that person should consider examinees' scores 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 be based on multiple job-relevant measures rather than relying on any single measure (e.g., using only BMCT scores to make hiring decisions).

When interpreting BMCT scores, it is useful to know the specific behaviors that an applicant with a high BMCT score may be expected to exhibit. In validity studies conducted by Pearson for the revision of this manual (see chapter 8), the following behaviors, as rated by supervisors, were consistently found to be positively related to BMCT scores across different occupations that require mechanical aptitude:

- diagnosing complex problems, involving machinery or equipment;
- installing equipment, machines, wiring, or programs to meet specifications;
- imagining how something will look after it is has been moved around or when its parts are moved or rearranged;
- developing new ideas to address work-related problems;
- understanding the implications of new information;
- discerning subtle interrelationships among seemingly disparate pieces of information;
- demonstrating knowledge of physical principles and laws;
- demonstrating knowledge of machines and tools; and/or
- demonstrating mechanical knowledge and expertise.

Thus, applicants who score higher on the BMCT tend to display a higher level of competence in these behaviors. Conversely, applicants who score low on the BMCT may not exhibit these behaviors, or they may find it challenging to effectively demonstrate these behaviors.

Human resource professionals who use the BMCT should document and examine the relationship between examinees' scores and their subsequent performance on the job. Using locally obtained information provides the best foundation for interpreting scores and most effectively differentiating examinees who are likely to be successful from those who are not. Pearson Education does **not** establish or recommend a cut score for the BMCT due to the multiple situational factors that must be considered in doing so.

Fairness in Selection Testing

Fair employment regulations and their interpretation are continuously subject to changes in the legal, social, and political environments. Therefore, a user of the BMCT should consult with qualified legal advisors and human resources professionals as appropriate.

Legal Considerations

Governmental and professional regulations 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 (1999) *Testing and Assessment: An Employer's Guide to Good Practices*.

Group Differences/Adverse Impact

Local validation is particularly important when a selection test may have adverse impact. According to the *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, 1978), adverse impact is indicated when the selection rate for one group is less than 80% (or 4 out of 5) of another. Adverse impact is likely to occur with aptitude tests such as the BMCT. Though it is within the law to use a test with adverse impact (Equal Employment Opportunity Commission, 1978), the testing organization must be prepared to demonstrate that the selection test is job-related and consistent with business necessity. A local validation study, in which scores on the BMCT are correlated with indicators of on-the-job performance, can provide evidence to support the use of the test in a particular job context. In addition, employers should conduct an evaluation that demonstrates that the BMCT (or any employment assessment tool) is equally predictive for protected subgroups, as outlined by the Equal Employment Opportunity Commission, to assist in the demonstration of fairness of the test.

Monitoring the Selection System

An organization's abilities to evaluate selection strategies and to implement fair employment practices depend on its awareness of the demographic characteristics of applicants and incumbents. Monitoring these characteristics and accumulating test score data are clearly necessary for establishing legal defensibility of a selection system, including those systems that incorporate the BMCT. The most effective use of the BMCT is with a local norms database that is regularly updated and monitored.

Employee Development

Some organizations provide career counseling services to employees as part of a formal career development program or as outplacement counseling as a result of employment termination. These services assist the employee in identifying and exploring career alternatives that are well-matched with his or her interests and aptitudes. In addition, career development counseling assists the organization in identifying candidates for training or other suitable roles within the organization.

The BMCT may be used by employee development counselors to assist adults with career exploration. In this application, the counselor seeks to identify career options with mechanical comprehension requirements that are well-suited to the counselee's measured abilities. The BMCT may also be used to evaluate an employee's need for training in preparation for advancement within the organization.

Training and Instruction

The ability to comprehend mechanical principles is an important objective of many instructional programs in both business and educational settings. The BMCT may be used to measure the extent to which examinees have mastered certain aspects of mechanical comprehension or need training. The availability of comparable forms (Form S and Form T) makes it possible to conduct pre- and post-testing to gauge the efficacy of instructional programs, as well as measure development of these aptitudes over an extended period of time.

Appendix A

Normative Data

Table A.1 Industry Characteristics, Means, and Standard Deviations by Occupation

Occupation Characteristics	Industry Characteristics	Mean	*SD*
Automotive Mechanic *N* = 94	Automotive mechanic positions in various industries, including Transportation/Warehousing, and Government/Public Service/Defense.	46.2	11.2
Engineer *N* = 105	Engineer positions in various industries, including Energy/Utilities, Government/Public Service/Defense, and Manufacturing/Production.	52.7	8.7
Installation/Maintenance/Repair *N* = 112	Installation, maintenance, and repair positions in various industries, including Construction, Energy/Utilities, Government/Public Service/Defense, Manufacturing/Production, and Transportation/Warehousing.	49.1	8.4
Industrial/Technical Sales Representative *N* = 133	Sales positions in various industrial and technical industries, including Information Technology/High-Tech/Telecommunications, Manufacturing/Production, and Retail/Wholesale.	47.4	7.9
Skilled Tradesperson (e.g., electrician, welder, carpenter) *N* = 153	Skilled trades positions in various industries, including Construction, Energy/Utilities, Manufacturing/Production, Aerospace/Aviation, Government/Public Service/Defense, and Information Technology/High-Tech/Telecommunications.	47.4	8.9
Transportation Trades/Equipment Operator (e.g., truck driver, heavy equipment operator) *N* = 126	Transportation Trades/Equipment Operator positions in various industries, including Transportation/Warehousing, Manufacturing/Production, and Energy/Utilities.	42.8	9.4

Table A.2 Industry Characteristics, Means, and Standard Deviations by Combined Occupation

Combined Occupation Characteristics	Industry Characteristics	Mean	*SD*
Mechanical Trades (e.g., mechanic, installation/maintenance/repair, electrician, welder, carpenter) *N* = 387	Mechanical trade occupations in various industries, including Aerospace/Aviation, Construction, Energy/ Utilities, Government/Public Service/Defense, Information Technology/High-Tech/Telecommunications, Manufacturing/ Production, and Transportation/Warehousing.	47.5	9.4
Automotive and Aircraft Mechanic *N* = 122	Automotive and aircraft mechanic positions in various industries, including Aerospace/Aviation, Government/ Public Service/Defense, and Transportation/Warehousing.	46.1	10.7

Table A.3 Occupation and Position Type/Level Characteristics, Means, and Standard Deviations by Industry

Industry Characteristics	Occupation and Position Type/Level Characteristics		Mean	*SD*
Manufacturing/Production	Various occupations in manufacturing and production industries.		43.8	10.5
N = 580	**Occupation Characteristics** (of those who reported occupation)			
	0.3%	Accountant, Auditor, Bookkeeper		
	1.3%	Administrative Assistant, Secretary, Clerk, Office Support		
	0.7%	Aircraft Mechanic, Service Technician		
	0.3%	Automotive Mechanic		
	1.0%	Customer Service Representative		
	10.4%	Engineer		
	0.3%	Food Services Occupations		
	0.3%	Housekeeper, Janitor, Building Cleaner		
	0.3%	Human Resource Occupations		
	0.7%	Information Technology Occupations		
	7.7%	Installation, Maintenance, Repair		
	9.1%	Sales Representative		
	19.5%	Skilled Tradesperson		
	10.1%	Transportation Trades/Equipment Operator		
	37.9%	Other		
	Position Type/Level Characteristics			
	19.1%	Blue Collar		
	56.7%	Hourly/Entry Level		
	2.9%	Supervisor		
	5.7%	Manager		
	0.3%	Director		
	0.5%	Executive (e.g., CEO, CFO, VP)		
	8.3%	Professional/Individual Contributor		
	0.5%	Self-Employed/Business Owner		
	5.9%	Not Specified		

(continued)

Table A.3 Occupation and Position Type/Level Characteristics, Means, and Standard Deviations by Industry *(continued)*

Industry Characteristics	Occupation and Position Type/Level Characteristics		Mean	*SD*
Energy/Utilities	Various occupations in energy and utilities industries.		48.1	9.2
N = 135	**Occupation Characteristics** (of those who reported occupation)			
	0.8%	Automotive Mechanic		
	0.8%	Customer Service Representative		
	5.5%	Engineer		
	0.8%	Housekeeper, Janitor, Building Cleaner		
	1.6%	Human Resource Occupations		
	13.3%	Installation, Maintenance, Repair		
	4.7%	Sales Representative		
	20.3%	Skilled Tradesperson		
	11.7%	Transportation Trades/Equipment Operator		
	40.6%	Other		
	Position Type/Level Characteristics			
	36.3%	Blue Collar		
	32.6%	Hourly/Entry Level		
	8.2%	Supervisor		
	6.7%	Manager		
	0.7%	Director		
	0.0%	Executive (e.g., CEO, CFO, VP)		
	8.9%	Professional/Individual Contributor		
	1.5%	Self-Employed/Business Owner		
	5.2%	Not Specified		

Table A.4 BMCT Norms by Occupation

Raw Score	Automotive Mechanic	Engineer	Installation, Maintenance, Repair	Industrial/ Technical Sales Representative	Skilled Tradesperson	Transportation Trades/ Equipment Operator	Mechanical Trades	Automotive and Aircraft Mechanic	Raw Score	Scaled Score
68	≥99	≥99	≥99	≥99	≥99	≥99	≥99	≥99	68	>90
67	≥99	≥99	≥99	≥99	≥99	≥99	≥99	≥99	67	90
66	≥99	97	≥99	≥99	≥99	≥99	≥99	≥99	66	82
65	≥99	94	≥99	≥99	≥99	≥99	≥99	≥99	65	78
64	≥99	92	≥99	≥99	≥99	≥99	≥99	≥99	64	75
63	≥99	89	≥99	≥99	97	≥99	98	≥99	63	72
62	98	85	97	98	96	98	97	98	62	70
61	95	81	95	96	94	98	95	96	61	68
60	93	77	92	94	92	97	93	95	60	66
59	92	73	89	91	92	96	91	93	59	65
58	90	70	84	89	90	96	88	91	58	63
57	86	65	80	87	85	95	84	85	57	62
56	81	60	77	86	81	92	80	81	56	61
55	77	56	73	83	79	90	76	77	55	60
54	72	53	69	80	76	87	73	72	54	58
53	68	50	65	76	72	85	69	69	53	57
52	64	44	58	71	67	82	64	66	52	56
51	61	39	51	67	63	79	60	63	51	55
50	57	35	46	65	59	76	55	60	50	54
49	53	31	43	61	55	74	52	56	49	53
48	49	28	40	55	51	71	48	53	48	52
47	45	25	37	48	46	68	44	48	47	51
46	40	23	34	43	41	65	40	43	46	50
45	38	19	32	38	37	60	36	40	45	49
44	35	15	29	31	34	56	33	36	44	49
43	31	14	27	25	32	54	30	32	43	48

(continued)

Table A.4 BMCT Norms by Occupation *(continued)*

Raw Score	Automotive Mechanic	Engineer	Installation, Maintenance, Repair	Industrial/ Technical Sales Representative	Skilled Tradesperson	Transportation Trades/ Equipment Operator	Mechanical Trades	Automotive and Aircraft Mechanic	Raw Score	Scaled Score
42	30	11	24	22	28	49	27	30	42	47
41	27	9	20	21	25	40	24	27	41	46
40	25	7	17	19	20	35	21	25	40	45
39	24	7	14	15	16	32	18	23	39	44
38	21	6	13	11	14	29	15	20	38	44
37	19	5	10	9	11	26	13	18	37	43
36	18	3	8	8	10	23	11	17	36	42
35	17	3	7	7	9	20	10	16	35	41
34	16	2	5	6	8	18	9	15	34	41
33	15	2	3	5	7	15	8	14	33	40
32	13	2	2	5	6	13	7	12	32	39
31	12	2	≤1	4	5	10	5	11	31	38
30	10	2	≤1	2	3	8	5	10	30	38
29	10	2	≤1	≤1	3	7	4	9	29	37
28	10	≤1	≤1	≤1	2	6	4	9	28	36
27	9	≤1	≤1	≤1	2	4	3	8	27	35
26	7	≤1	≤1	≤1	2	4	3	7	26	35
25	6	≤1	≤1	≤1	≤1	3	2	5	25	34
24	5	≤1	≤1	≤1	≤1	2	2	4	24	33
23	5	≤1	≤1	≤1	≤1	2	≤1	4	23	32
22	4	≤1	≤1	≤1	≤1	2	≤1	3	22	31
21	4	≤1	≤1	≤1	≤1	2	≤1	3	21	30
20	3	≤1	≤1	≤1	≤1	2	≤1	2	20	29
19	2	≤1	≤1	≤1	≤1	2	≤1	2	19	28
18	2	≤1	≤1	≤1	≤1	2	≤1	≤1	18	27
17	≤1	≤1	≤1	≤1	≤1	2	≤1	≤1	17	26

(continued)

Table A.4 BMCT Norms by Occupation *(continued)*

Raw Score	Automotive Mechanic	Engineer	Installation, Maintenance, Repair	Industrial/ Technical Sales Representative	Skilled Tradesperson	Transportation Trades/ Equipment Operator	Mechanical Trades	Automotive and Aircraft Mechanic	Raw Score	Scaled Score
16	≤1	≤1	≤1	≤1	≤1	2	≤1	≤1	16	25
15	≤1	≤1	≤1	≤1	≤1	≤1	≤1	≤1	15	24
14	≤1	≤1	≤1	≤1	≤1	≤1	≤1	≤1	14	23
13	≤1	≤1	≤1	≤1	≤1	≤1	≤1	≤1	13	22
12	≤1	≤1	≤1	≤1	≤1	≤1	≤1	≤1	12	21
11	≤1	≤1	≤1	≤1	≤1	≤1	≤1	≤1	11	20
10	≤1	≤1	≤1	≤1	≤1	≤1	≤1	≤1	10	18
9	≤1	≤1	≤1	≤1	≤1	≤1	≤1	≤1	9	17
8	≤1	≤1	≤1	≤1	≤1	≤1	≤1	≤1	8	15
7	≤1	≤1	≤1	≤1	≤1	≤1	≤1	≤1	7	14
6	≤1	≤1	≤1	≤1	≤1	≤1	≤1	≤1	6	12
≤5	≤1	≤1	≤1	≤1	≤1	≤1	≤1	≤1	≤5	<10
Raw Score Mean	46.2	52.7	49.1	47.4	47.4	42.8	47.5	46.1	**Raw Score Mean**	
Raw Score *SD*	11.2	8.7	8.4	7.9	8.9	9.4	9.4	10.7	**Raw Score *SD***	
N	94	105	112	133	153	126	387	122	***N***	

Table A.5 BMCT Scores by Industry

Raw Score	Energy/ Utilities	Manufacturing/ Production	Raw Score	Scaled Score
68	≥99	≥99	68	>90
67	≥99	≥99	67	90
66	≥99	≥99	66	82
65	≥99	≥99	65	78
64	98	98	64	75
63	97	98	63	72
62	96	97	62	70
61	95	95	61	68
60	93	94	60	66
59	90	92	59	65
58	85	91	58	63
57	81	89	57	62
56	78	87	56	61
55	74	85	55	60
54	71	82	54	58
53	67	80	53	57
52	62	77	52	56
51	57	74	51	55
50	53	72	50	54
49	51	68	49	53
48	49	64	48	52
47	44	60	47	51
46	39	57	46	50
45	35	53	45	49
44	31	50	44	49
43	29	47	43	48
42	25	44	42	47
41	21	39	41	46
40	19	35	40	45
39	16	32	39	44
38	13	29	38	44
37	12	26	37	43
36	11	23	36	42
35	10	20	35	41
34	9	17	34	41
33	7	16	33	40
32	6	14	32	39
31	4	12	31	38
30	4	10	30	38
29	3	7	29	37

(continued)

Table A.5 BMCT Scores by Industry *(continued)*

Raw Score	Energy/ Utilities	Manufacturing/ Production	Raw Score	Scaled Score
28	3	6	28	36
27	3	4	27	35
26	2	4	26	35
25	2	4	25	34
24	≤1	3	24	33
23	≤1	3	23	32
22	≤1	2	22	31
21	≤1	2	21	30
20	≤1	≤1	20	29
19	≤1	≤1	19	28
18	≤1	≤1	18	27
17	≤1	≤1	17	26
16	≤1	≤1	16	25
15	≤1	≤1	15	24
14	≤1	≤1	14	23
13	≤1	≤1	13	22
12	≤1	≤1	12	21
11	≤1	≤1	11	20
10	≤1	≤1	10	18
9	≤1	≤1	9	17
8	≤1	≤1	8	15
7	≤1	≤1	7	14
6	≤1	≤1	6	12
≤5	≤1	≤1	≤5	<10
Raw Score Mean	48.1	43.8	**Raw Score Mean**	
Raw Score *SD*	9.2	10.5	**Raw Score *SD***	
N	135	580	***N***	

References

Aiken, L. R. (1979). *Psychological testing and assessment* (3rd ed.). Boston: Allyn & Bacon.

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, 42 U.S.C.A. § 12101 *et seq*. (West 1993).

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Bechtoldt, H. E. (1972). Review of the Bennett Mechanical Comprehension Test. In J. V. Mitchell, Jr. (Ed.), *The seventh mental measurements yearbook* (pp. 1484–1485). Lincoln: University of Nebraska Press.

Benbow, C. E., Stanley, J. C., Kirk, M. K., & Zonderman, A. B. (1983). Structure of intelligence in intellectually precocious children and in their parents. *Intelligence, 7,* 129–152.

Bennett, G. K. (1940). *Manual of the Test of Mechanical Comprehension, Form AA*. San Antonio: The Psychological Corporation.

Bennett, G. K. (1969). *Manual for the Bennett Mechanical Comprehension Test, Forms S and T.* San Antonio: The Psychological Corporation.

Bennett, G. K. (1981). *Bennett Hand–Tool Dexterity Test manual*. San Antonio: The Psychological Corporation.

Bennett, G. K. (1994). *Manual for the Bennett Mechanical Comprehension Test, Forms S and T.* (2nd ed.). San Antonio: The Psychological Corporation.

Bennett, G. K., & Cruikshank, R. M. (1942). Sex differences in the understanding of mechanical problems. *Journal of Applied Psychology, 26,* 121–127.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1947). *A manual for the Differential Aptitude Tests*. New York: The Psychological Corporation.

Bennett, G., Seashore, H., & Wesman, A. (1991). *Technical manual for the Differential Aptitude Tests for Personnel and Career Assessment.* San Antonio, TX: The Psychological Corporation.

Callender, J. C., & Osburn, H. G. (1981). Testing the constancy of validity with computer-generated sampling distributions of the multiplicative model variance estimate: Results for petroleum industry validation research. *Journal of Applied Psychology, 66,* 274–281.

Cascio, W. F. (1982). *Applied psychometrics in personnel management* (2nd ed.). Reston, VA: Reston Publishing.

Cascio, W. F., & Aguinis, H. (2005). *Applied psychology in human resource management* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Ciechalski, J. C. (2005). Review of the Bennett Mechanical Comprehension Test. In R. A. Spies & B. S. Plake (Eds.), *The sixteenth mental measurements yearbook*. Retrieved from http://buros.unl.edu/buros/jsp/reviews.jsp?item=16182568.

Cohen, B. H. (1996). *Explaining psychological statistics*. Pacific Grove, CA: Brooks & Cole.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Crawford, J. E., & Crawford, D. M. (1981). *Crawford Small Parts Dexterity Test manual*. San Antonio, TX: The Psychological Corporation.

Cronbach, L. (1970). *Essentials of psychological testing*. New York: Harper & Row.

Dagenais, F. (1992). Bennett Mechanical Comprehension Test: Normative data for a sample of Saudi Arabian technical trainees. *Perceptual and Motor Skills, 74,* 107–113.

de Wolf, V. (1981). High school mathematics preparation and sex differences in quantitative abilities. *Psychology of Women Quarterly, 5,* 555–567.

Equal Employment Opportunity Commission. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*(166), 38295–38309.

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist, 43*, 95–103.

Fortson, H. D. (1991). An investigation of gender differences in mechanical aptitude. Unpublished doctoral dissertation, California School of Professional Psychology.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Erlbaum.

Halpern, D. (1986). *Sex differences in cognitive abilities*. Hillsdale, NJ: Erlbaum.

Hambleton, R. K. (1972). Review of the Bennett Mechanical Comprehension Test. In J. V. Mitchell, Jr. (Ed.), *The seventh mental measurements yearbook* (pp. 720–721). Lincoln: University of Nebraska Press.

Pearson Education, Inc. (2005). *Miller Analogies Test for Professional Selection manual*. San Antonio: Author.

Harris, A. J., & Jacobson, M. D. (1982). *Basic reading vocabularies*. New York: Macmillan.

Hegarty, M., Just, M. A., & Morrison, I. R. (1988). Mental models of mechanical systems: Individual differences in qualitative and quantitative reasoning. *Cognitive Psychology*, *20,* 191–236.

Kellogg, C. E., & Morton, N. W. (1999). *Beta III manual*. San Antonio: The Psychological Corporation.

Kolz, A. R., McFarland, L. A., & Silverman, S. B. (1998). Cognitive ability and job experience as predictors of work performance. *The Journal of Psychology, 132*(5), 539–548.

Lowman, R. L., & Williams, R. E. (1987). Validity of self-ratings of abilities and competencies. *Journal of Vocational Behavior, 31,* 1–13.

Lunneborg, P. W., & Lunneborg, C. E. (1985). Nontraditional and traditional female college graduates: What separates them from the men? *Journal of College Student Personnel, 26,* 33–36.

McCall, J. N. (1973). Birth-order differences in special ability: Fact or artifact? *Psychological Reports, 33,* 947–952.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449–458.

Mount, M. K., Muchinsky, P. M., & Hanser, L. M. (1977). The predictive validity of a work sample: A laboratory study. *Personnel Psychology, 30,* 636–645.

Muchinsky, P. M. (1993). Validation of intelligence and mechanical aptitude tests in selecting employees for manufacturing jobs. *Journal of Business and Psychology, 7*(4), 373–382.

The Psychological Corporation. (1974). *Fifth edition manual for the Differential Aptitude Tests, Forms S & T.* San Antonio: Author.

The Psychological Corporation. (1992). *Differential Aptitude Tests Fifth Edition technical manual*. San Antonio: Author.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Roberts, O. H. (1972). [Review of the Bennett Mechanical Comprehension Test]. In J. V. Mitchell. Jr. (Ed.), *The seventh mental measurements yearbook* (pp. 1485–1486). Lincoln: University of Nebraska Press.

Sapitula, L. L., & Shartzer, M. C. (2001). Predicting the job performance of maintenance workers using a job knowledge test and a mechanical aptitude test. *Applied HRM Research, 6*(1), 71–74.

Schmidt, E. L., Hunter, J. E., & Caplan, J. R. (1981). Validity generalization results for two job groups in the petroleum industry, *Journal of Applied Psychology, 66,* 261–273.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Spangler, M. (2005). [Review of the Bennett Mechanical Comprehension Test]. In R. A. Spies & B. S. Plake (Eds.), *The sixteenth mental measurements yearbook*. Retrieved from http://buros.unl.edu/buros/jsp/reviews.jsp?item=16182568.

Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science? A critical review. *American Psychologist, 60*(9), 950–958.

Taylor, S. E., Frackenpohl, H., White, C. E., Nieroroda, B. W, Browning, C. L., & Birsner, E. E. (1989). *EDL core vocabularies in reading, mathematics, science, and social studies*. Columbia, SC: EDL.

U.S. Department of Labor. (1999). *Testing and assessment: An employer's guide to good practices*. Washington, DC: Author.

Watson, G., & Glaser, E. M. (2005). *Watson-Glaser critical thinking appraisal short form manual*. San Antonio: Pearson Education, Inc.

Wing, H. (1992). [Review of the Bennett Mechanical Comprehension Test]. In J. V. Mitchell, Jr. (Ed.), *The eleventh mental measurements yearbook* (pp. 106–107). Lincoln: University of Nebraska Press.

Yang, Z., Wang, S., Trent, J., & Rose, M. (2006, April). *The effect of calibration methods on accuracy of item parameter estimation with the Rasch Model*. Presented at the Chinese American Educational Research and Development Association (CAERDA) International Conference, San Francisco. http://www.caerda.org/conference.htm.

Research Bibliography

Anderson, R. G. (1946). A comparative study of test scores and supervisors' efficiency ratings of machinists [Abstract]. *American Psychologist, 1,* 243.

Anderson, R. G. (1947). Test scores and efficiency ratings of machinists. *Journal of Applied Psychology, 31,* 377–388.

Ash, P. (1960). Validity information exchange, No. 13-0.6: D.O.T.; Code 5-83.127, Typewriter Serviceman. *Personnel Psychology, 13,* 455.

Avolio, B. J., & Waldman, D.A. (1987). Personnel aptitude test scores as a function of age, education, and job type. *Experimental Aging Research, 13,* 109–113.

Barnette, W. L., Jr. (1949). *Occupational aptitude patterns of counseled veterans*. Unpublished doctoral dissertation, New York University.

Barnette, W. L., Jr. (1949). Occupational aptitude patterns of selected groups of counseled veterans. *Psychological Monographs: General and Applied, 65* (5, Whole No. 322). Washington, DC: American Psychological Association.

Barnette, W. L., Jr. (1950). Occupational aptitude pattern research. *Occupations, 29,* 5–12.

Barrett, R. S. (1958). The process of predicting job performance. *Personnel Psychology, 11,* 39–57.

Belmont, L. (1977). Birth order, intellectual competence, and psychiatric status. *Journal of Individual Psychology, 33,* 97–104.

Belmont, L. (1978). Birth order, intellectual competence, and psychiatric status. *Annual Progress in Child Psychiatry and Child Development,* 51–58.

Belmont, L., Wittes, J., & Stein, Z. (1977). Relation of birth order, family size and social class to psychological functions. *Perceptual and Motor Skills, 45,* 1107–1116.

Bennett, G. K., & Wesman, A. G. (1947). Industrial test norms for a southern plant population. *Journal of Applied Psychology, 31,* 241–246.

Bordieri, J. E. (1988). Reward contingency, perceived competence, and attribution of intrinsic motivation: An observer simulation. *Psychological Reports, 63,* 755–762.

Borg, W. R. (1950a). Does a perceptual factor exist in artistic ability? *Journal of Educational Research,* 44, 47–53.

Borg, W. R. (1950b). Some factors relating to art school success. *Journal of Educational Research,* 43, 376–384.

Bradley, A. D. (1958). Estimating success in technical and skilled trade courses using a multivariate statistical analysis (Doctoral dissertation, University of Minnesota). *Dissertation Abstracts, 21,* 313.

Bradshaw, O. L. (1968). The relationship of selected measures of aptitude, interest, and personality to academic achievement in engineering and engineering technology (Doctoral dissertation, Oklahoma State University). *Dissertation Abstracts International, 30,* 979A.

Bruce, M. M. (1952). The importance of certain personality characteristics, skills, and abilities in effectiveness as a factory foreman (Doctoral dissertation, New York University). *Dissertation Abstracts, 13,* 116.

Bruce, M. M. (1953). The prediction of effectiveness as a factory foreman. *Psychological Monographs,* 67, 1–17.

Bruce, M. M. (1954a). Validity information exchange, No. 7-076; D.O.T. Code 5-91.101, Foreman II. *Personnel Psychology, 7,* 418–419.

Bruce, M. M. (1954b). Validity information exchange, No. 7-079: D.O.T. Code 7-83.058, Electrical Appliance Serviceman. *Personnel Psychology, 7,* 425–426.

Bruce, M. M. (1956a). Normative data information exchange, No. 17. *Personnel Psychology, 9,* 393.

Bruce, M. M. (1956b). Normative data information exchange, Nos. 18, 37. *Personnel Psychology, 9,* 394, 552–553.

Carter, G. C. (1952). Measurement of supervisory ability. *Journal of Applied Psychology, 36,* 393–395.

Carter, L., & Nixon, M. (1949). Ability, perceptual, personality, and interest factors associated with different criteria of leadership. *Journal of Psychology, 27,* 377–388.

Case, H. W. (1952). The relationship of certain tests to grades achieved in an industrial class in aircraft design. *Educational and Psychological Measurement, 12,* 90–95.

Cass, J. G., & Tiedman, D. V. (1960). Vocational development and the election of a high school curriculum. *Personnel and Guidance Journal, 38,* 538–545.

Chandler, R. E. (1956). Validation of apprentice screening tests in an oil refinery (Doctoral dissertation, Purdue University). *Dissertation Abstracts, 27,* 325B.

Clegg, H. D., & Decker, R. L. (1962). The evaluation of a psychological test battery as a selective device for foremen in the mining industry. *Proceedings of the West Virginia Academy of Sciences, 34,* 178–182.

Coleman, W. (1953). An economical test battery for predicting freshman engineering course grades. *Journal of Applied Psychology, 37,* 465–467.

Cottingham, H. F. (1947). The predictive value of certain paper and pencil mechanical aptitude tests in relation to woodworking achievement of junior high school boys (Doctoral dissertation, Indiana University). *Studies in Education, 1945–1949,* 5–10.

Cottingham, H. F. (1948). Paper-and-pencil tests given to students in woodworking. *Occupations, 27,* 95–99.

Crane, W. J. (1962). Screening devices for occupational therapy majors. *American Journal of Occupational Therapy, 16,* 131–132.

Cuomo, S. (1955). Validity information exchange, No. 8-17: D.O.T. Code 5-92.601, Foreman II. *Personnel Psychology, 8,* 268.

Cuomo, S., & Meyer, H. H. (1955a). Validity information exchange, No. 8-16: D.O.T. Code 5-92.601, Foreman II. *Personnel Psychology, 8,* 267.

Cuomo, S., & Meyer, H. H. (1955b). Validity information exchange, No. 8-19: D.O.T. Code 6-78.632, Floor Assembler. *Personnel Psychology, 8,* 270.

Decker, R. L. (1958). A study of the value of the Owens-Bennett Mechanical Comprehension Test (Form CC) as a measure of the qualities contributing to successful performance as a supervisor of technical operations in an industrial organization. *Journal of Applied Psychology, 42,* 50–53.

DeNisi, A. S., & Shaw, J. B. (1977). Investigation of the uses of self-reports of abilities. *Journal of Applied Psychology, 62,* 641–644.

Dicken, C. F., & Black, J. D. (1965). Predictive validity of psychometric evaluations of supervisors. *Journal of Applied Psychology, 49,* 34–47.

Drew, A. S. (1964). The relationship of general reading ability and other factors to school and job performance of machine apprentices. *Journal of Industrial Teaching Education, 2,* 47–60.

DuBois, P. H. (1950). The selection of patrolmen. *Journal of Applied Psychology, 34,* 90–95.

DuBois, P. H., & Watson, R. I. (1954). Validity information exchange, No. 7-075: D.O.T. Code 2-66.23, Policeman. *Personnel Psychology, 7,* 414–417.

Dunham, R. E. (1954). Factors related to recidivism in adults. *Journal of Social Psychology, 39,* 77–91.

Durrett, H. L. (1961). Validity information exchange, No. 14-03: D.O.T. Code 5-21.010, Continuous Miner Operator (Bituminous Coal Industry). *Personnel Psychology, 14,* 453–455.

Ferson, R. F. (1951). *The probabilities of success in trade training as estimated by standardized tests*. Unpublished doctoral dissertation. University of Pittsburgh.

Fiske, D. W. (1947). Validation of naval aviation cadet selection tests against training criteria. *Journal of Applied Psychology*, *31,* 601–614.

Fitzpatrick, E. D., & McCarty, J. J. (1955). Validity information exchange, No. 8-35: D.O.T. Code 9-00.91, Assembler VII (Electrical Equipment). *Personnel Psychology*, *8,* 501–504.

Forehand, G. A., & McQuitty, L. L. (1959). Configurations of factor standings as predictors of educational achievement. *Educational and Psychological Measurement*, *19,* 31–43.

Forster, C. R. (1955). The relationship between test achievement and success in training of a selected group of tuberculosis patients (Doctoral dissertation, New York University). *Dissertation Abstracts*, 15, 1201.

Gilbert, H. G. (1952). The use of tests and other objective data in the selection of camp counselors [Abstract]. *American Psychologist*, *7,* 369.

Glennon, J. R., Smith, W. J., & Albright, L. E. (1958). Normative data information exchange, Nos. 11–35, 11–36. *Personnel Psychology*, *11,* 601–602.

Gordon, T. (1949). The airline pilot's job. *Journal of Applied Psychology*, *33,* 122–131.

Gough, H. G., Lazzari, R., Fioravanti, M., & Stracca, M. (1978). An adjective check list scale to predict military leadership. *Journal of Cross-Cultural Psychology*, *9,* 381–400.

Greene, R. R. (1946). Ability to perceive and react differentially to configurational changes as related to the piloting of light aircraft (Doctoral dissertation, Ohio State University). *Abstracts of Dissertations, 1946–1947*, 65–72.

Grimes, J. W., & Scalise, J. J. (1986). *An analysis of variables affecting high school students' vocational choices*. Louisiana State Department of Education, Division of Vocational Education, Baton Rouge, LA.

Grohsmeyer, E. A. (1954). Validation of a personnel test for a paper mill (Doctoral dissertation, Purdue University). *Dissertation Abstracts*, *14,* 1796.

Halliday, R. W., & Fletcher, E. M. (1950). The relationship of Owens-Bennett scores to first-year achievement in an engineering college [Abstract]. *American Psychologist*, 5, 353.

Halliday, R. W., Fletcher, E. M., & Cohen, R. M. (1951). Validity of the Owens-Bennett Mechanical Comprehension Test. *Journal of Applied Psychology, 35,* 321–324.

Halstead, H. (1950). Abilities of male mental hospital patients. *Journal of Mental Science, 96,* 726–733.

Hamilton, J. W., & Dickinson, T. L. (1987). Comparison of several procedures for generating J-coefficients. *Journal of Applied Psychology, 72,* 49–54.

Hanes, B. (1952). A factor analysis of the MMPI, aptitude test data, and personal information using a population of criminals (Doctoral dissertation, Ohio State University). *Dissertation Abstracts, 18,* 1483.

Harrison, R., Hunt, W., & Jackson, T. A. (1955). Profile of the mechanical engineer: I, ability *Personnel Psychology, 8,* 219–234.

Hinman, S. L. (1967). *A predictive validity study of creative managerial performance.* Greensboro, NC: Creativity Research Institute of the Richardson Foundation, Inc.

Hodgson, R. W. (1964). Personality appraisal of technical and professional applicants. *Personnel Psychology, 17,* 167–187.

Holland, J. L., & Nafziger, D. H. (1975). A note on the validity of the Self-Directed Search. *Measurement and Evaluation in Guidance, 7,* 259–262.

Hueber, J. (1954). Validity information exchange, No. 7-089: D.O.T. Code 5-83.641, Maintenance Mechanic II. *Personnel Psychology, 7,* 565–566.

Jacobsen, E. E. (1943). An evaluation of certain tests in predicting mechanic learner achievement. *Educational and Psychological Measurement, 3,* 259–267.

Jensen, M., & Rotter, J. B. (1947). The value of thirteen psychological tests in officer candidate screening. *Journal of Applied Psychology, 31,* 312–322.

Johnson, R. H. (1950). Reading ease of commonly used tests. *Journal of Applied Psychology, 34,* 319–324.

Johnson, R. H. (1955). Factors related to the success of disabled veterans of World War II in the rehabilitation training program approved for mechanics and repairmen, motor vehicle (Doctoral dissertation, University of Minnesota). *Dissertation Abstracts*, *15,* 2460.

Juergensen, E. M. (1958). The relationship between success in teaching vocational agriculture and ability to make sound judgments as measured by selected instruments (Doctoral dissertation, University of Pennsylvania). *Dissertation Abstracts, 19,* 96.

Jurgensen, C. E. (1948). Norms for the Test of Mechanical Comprehension. *Journal of Applied Psychology, 32,* 618–621.

Kazmier, L. J. (1959). Normative data information exchange, No. 12–23. *Personnel Psychology, 12, 505.*

Kirkpatrick, J. J. (1956). Validation of a test battery for the selection and placement of engineers. *Personnel Psychology, 9,* 211–227.

Krathwohl, D. R., Ewing, T. N., Gilbert, W. M., & Cronbach, L. (1952). Prediction of success in architecture courses [Abstract]. *American Psychologist, 7,* 288–289.

Lane, G. G. (1947a). *Prediction of success in learning to fly light aircraft.* Unpublished doctoral dissertation, Ohio State University.

Lane, G. G. (1947b). Studies in pilot selection: I, The prediction of success in learning to fly light aircraft. *Psychological Monographs, 61,* 1–17.

Laney, A. R. (1951). Validity of employment tests for gas-appliance service personnel. *Personnel Psychology, 4,* 199–208.

Lee, M. C. (1952). Relationship of masculinity-femininity to tests of mechanical and clerical abilities. *Journal of Applied Psychology, 36,* 377–380.

Lingwood, J. (1952). Test performance of ATS recruits from certain civilian occupations. *Occupational Psychology, 26,* 35–46.

Lipsman, C. K. (1967). The relation of socio-economic level and occupational choice to needs and vocational behavior (Doctoral dissertation, Catholic University of America). *Dissertation Abstracts, 28,* 2073A.

Littleton, I. T. (1952). Prediction in auto trade courses. *Journal of Applied Psychology, 36,* 15–19.

MacKinney, A. C., & Wolins, L. (1960). Validity information exchange, No. 13-01, Foreman II, Home Appliance Manufacturing. *Personnel Psychology, 13,* 443–447.

McCarty, J. J. (1954). Validity information exchange, No. 7-077: D.O.T. Code 5-92.621, Foreman II. *Personnel Psychology, 7,* 420–421.

McCarty, J. J. (1957). Normative data information exchange, No. 10–31. *Personnel Psychology*, *10,* 365.

McCarty, J. J., & Fitzpatrick, E. D. (1955). Validity information exchange, No. 9-26: D.O.T. Code 5-92.621, Foreman II. *Personnel Psychology*, *9,* 253.

McCarty, J. J., Westberg, W. C., & Fitzpatrick, E. D. (1954). Validity information exchange, No. 7-091: D.O.T. Code 5-92.621, Foreman II. *Personnel Psychology*, 7, 568–569.

McDaniel, J. W., & Reynolds, W. (1944). A study of the use of mechanical aptitude tests in the selection of trainees for mechanical occupations. *Educational and Psychological Measurement, 4,* 191–197.

McElheny, W. T. (1948). A study of two techniques of measuring "mechanical comprehension." *Journal of Applied Psychology, 32,* 611–617.

McGehee, W., & Moffie, D. J. (1942). Psychological tests in the selection of enrollees in engineering, science, management, defense training courses. *Journal of Applied Psychology*, *26,* 584–586.

McMurry, R. N., & Johnson, D. L. (1945). Development of instruments for selecting and placing factory employees. *Adv*[sic] *Management*, *10,* 113–120.

Meadow, L. (1964). Assessment of students for schools of practical nursing. *Nursing Resources*, *13,* 222–229.

Miller, G. E. (1951). Some components of mechanical composition. *Proceedings of the Iowa Academy of Sciences, 58,* 385–389.

Moffie, D. J., & Goodner, S. (1967). *A predictive validity study of creative and effective managerial performance.* Greensboro, NC: Creativity Research Institute of the Richardson Foundation, Inc.

Mollenkopf, W. G. (1957). An easier "male" mechanical test for use with women. *Journal of Applied Psychology, 41,* 340–343.

Moore, B. V. (1941). Analysis of results of tests administered to men in engineering defense training courses. *Journal of Applied Psychology, 25,* 619–635.

Moore, C. L., McNaughton, J. E., & Osburn, H. G. (1969). Ethnic differences within an industrial selection battery. *Personnel Psychology, 22,* 473–482.

Nair, R. K. (1950). *Predictive value of standardized tests and inventories in industrial arts teacher education.* Unpublished doctoral dissertation, University of Missouri.

Onarheim, J. (1947). Scientific selection of sales engineers. *Personnel, 24,* 24–34.

Otterness, W. B., Patterson, C. H., Johnson, R. H., & Peterson, L. R. (1956). Trade school norms for some commonly used tests. *Journal of Applied Psychology, 40,* 57–60.

Owens, W. A. (1950). A difficult new test of mechanical comprehension. *Journal of Applied Psychology, 34,* 77–81.

Owens, W. A. (1959). A comment on the recent study of the Mechanical Comprehension Test (CC) by R. L. Decker. *Journal of Applied Psychology, 43,* 31.

Patterson, C. H. (1955). Test and background factors related to drop-outs in an industrial institute (Doctoral dissertation, University of Minnesota). *Dissertation Abstracts, 15,* 1024.

Patterson, C. H. (1956). The prediction of attrition in trade school courses. *Journal of Applied Psychology, 40,* 154–158.

Penfield, R. V. (1966). The psychological characteristics of effective first-line managers (Doctoral dissertation, Cornell University). *Dissertation Abstracts*, *27*, 1610B.

Poe, W. A., & Berg, I. A. (1952). Psychological test performance of steel industry production supervisors. *Journal of Applied Psychology, 36,* 234–237.

Richardson, Bellows, Henry & Co., Inc. (1963). *Normative information: Manager and executive testing* (pp. 45). New York: Author.

Riland, L. H., & Upshall, C. C. (1958). Normative data information exchange, No. 11–10. *Personnel Psychology*, *11,* 275.

Rinsland, H. D. (1949). The prediction of veterans' success from test scores at the University of Oklahoma. *In The sixth yearbook of the National Council on Measurements Used in Education, 1948–1949* (Part 1, pp. 59–72). Fairmont, WV: The Council, Fairmont State College.

Ronan, W. W. (1964). Evaluation of skilled trades performance predictors. *Educational and Psychological Measurement, 24,* 601–608.

Sartain, A. Q. (1945). The use of certain standardized tests in the selection of inspectors in an aircraft factory. *Journal of Consulting Psychology*, *9,* 234–247.

Sartain, A. Q. (1946). Relation between scores on certain standard tests and supervisory success in an aircraft factory. *Journal of Applied Psychology, 30,* 328–332.

Saunders, W. J. (1957). Normative data information exchange, No. 10–32. *Personnel Psychology, 10,* 366.

Schmitz, R. M., & Holmes, J. L. (1954). Relationship of certain measured abilities to freshman engineering achievement. In W. L. Layton (Ed.), *Selection and counseling of students in engineering*. Minneapolis: University of Minnesota Press.

Schorgmayer, H., & Swanson, R. A. (1975). *The effect of alternative training methods on the trouble-shooting performances of maintenance technicians*. Bowling Green State University and Johns-Manville Corporation.

Sell, J. M., & Torres-Henry, R. (1979). Testing practices in university and college counseling centers in the United States. *Professional Psychology*, *December,* 774–779.

Shukia, N. N. (1958). The relation of intelligence and ability to scholastic achievement of pupils in the S.S.C. class. *Journal of Educational and Vocational Guidance,* 5, 38–44.

Shultz, I. T., & Bentley, B. (1945). Testing for leadership in industry. *Transactions of the Kansas Academy of Sciences, 48,* 160–164.

Smith, O. B. (1955). *Predicting grade success of high school students in radio and drafting.* Unpublished master's thesis, Alabama Polytechnic Institute, Auburn.

Sorenson, W. W. (1966). Test of mechanical principles as a suppressor variable for the prediction of effectiveness on a mechanical repair job. *Journal of Applied Psychology, 50,* 348–352.

Super, D. E. (1949). *Appraising vocational fitness by means of psychological tests* (pp. 246–260). New York: Harper & Brothers.

Super, D. E., & Critter, J. O. (1962). *Appraising vocational fitness by means of psychological tests, revised edition* (pp. 242–256). New York: Harper & Brothers.

Swanson, R. A., & Sawzin, S. A. (1975). *Industrial training research project*. Bowling Green State University and Johns-Manville Corporation.

Taylor, D. W. (1963). Variables related to creativity and productivity among men in two research laboratories. In C. W. Taylor & E. Barren, *Scientific creativity: Its recognition and development* (pp. 228–250). New York: John Wiley & Sons.

Thumin, E. J. (1979). Performance on tests of power and speed as related to age among male job applicants. *International Journal of Aging and Human Development, 9,* 255–261.

Topetzes, N. J. (1957). A program for the selection of trainees in physical medicine. *Journal of Experimental Education, 25,* 263–311.

Torres, L. (1963). A study of the relationship between selected variables and the achievement of industrial arts students at Long Beach State College (Doctoral dissertation, Colorado State College). *Dissertation Abstracts, 25,* 316.

Travers, R. M. W., & Wallace, W. L. (1950). Inconsistency in the predictive value of a battery of tests. *Journal of Applied Psychology, 34,* 237–239.

Traxler, A. E. (1943). Correlations between mechanical aptitude scores and "mechanical comprehension" scores. *Occupations, 22,* 42–43.

Validity information exchange, No. 7-064: D.O.T. Code 5-92.411, Foreman I. *Personnel Psychology, 7, 300.*

Validity information exchange, No. 7-065: D.O.T. Code 5-92-411, Foreman II. *Personnel Psychology, 7,* 301.

Vernon, R. E. (1949). The structure of practical abilities. *Occupational Psychology, 23,* 81–96.

Walker, E. C. (1956). Normative data information exchange, No. 10. *Personnel Psychology, 9,* 276.

Welder, A. (1951). Some aspects of an industrial mental hygiene program. *Journal of Applied Psychology, 35,* 383–385.

Welsch, L. A. (1967). The supervisor's employee appraisal heuristic: The contribution of selected measures of employee aptitude (Doctoral dissertation, University of Pittsburgh). *Dissertation Abstracts, 28,* 4321A.

Whitlock, J. B., & Crannel, C.W. (1949). An analysis of certain factors in serious accidents in a large steel plant. *Journal of Applied Psychology, 33,* 494–498.

Wolff, W. M., & North, A. J. (1951). Selection of municipal firemen. *Journal of Applied Psychology, 35,* 25–29.

Yeslin, A. R., Vernon, L. N., & Kerr, W. A. (1958). The significance of time spent in answering personality inventories. *Journal of Applied Psychology, 42,* 264–266.

Glossary of Measurement Terms

This glossary of terms is intended as an aid in the interpretation of statistical information presented in the Bennett Mechanical Comprehension Test Manual, as well as other manuals published by Pearson Education, Inc. The terms defined are fairly common and basic. In the definitions, certain technicalities have been sacrificed for the sake of succinctness and clarity.

achievement test—A test that measures the extent to which a person has "achieved" something, acquired certain information, or mastered certain skills—usually as a result of planned instruction or training.

alternate-form reliability—The closeness of correspondence, or correlation, between results on alternate forms of a test; thus, a measure of the extent to which the two forms are consistent or reliable in measuring whatever they do measure. The time interval between the two testing events must be relatively short so that the examinees are unchanged in the ability being measured. See Reliability, Reliability Coefficient.

aptitude—A combination of abilities and other characteristics, whether innate or acquired, that are indicative of an individual's ability to learn or to develop proficiency in some particular area if appropriate education or training is provided. Aptitude tests include those of general academic ability (commonly called mental ability or intelligence tests); those of special abilities, such as verbal, numerical, mechanical, or musical; tests assessing "readiness" for learning; and prognostic tests, which measure both ability and previous learning and are used to predict future performance—usually in a field requiring specific skills, such as speaking a foreign language, taking shorthand, or nursing.

arithmetic mean—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.

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 term "average" is used without designation as to type, the most likely assumption is that it is the arithmetic mean. See Central Tendency, Arithmetic Mean, Median, Mode.

battery—A group of several tests standardized on the same population so that results on the several tests are comparable. Sometimes applied to any group of tests administered together, even though not standardized on the same subjects.

ceiling—The upper limit of ability that can be measured by a test. When an individual earns a score that is at or near the highest possible score, it is said that the "ceiling" of the test is too low for the individual, who should be given a higher level test.

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

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 several 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 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 ratings, etc., with which a test is designed to measure, to predict, or to correlate. See Validity.

cut-off score (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 examinees' 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.

equipercentile method—A method of equating tests or forms of a test in which the percentile equivalent of a raw score on one test or form is matched to the percentile equivalent on a second test or form; the corresponding raw score is considered "equivalent."

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 indice, 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.

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)*—*See* Arithmetic Mean, 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 scores are below the median and half above it, except when the median itself is one of the obtained scores. *See* Central Tendency.

mode—The score or value that occurs most frequently in a distribution.

multitrait-multimethod matrix—An experimental design to examine both convergent and discriminant validity, involving a matrix showing the correlations between the scores obtained (1) for the same trait by different methods, (2) for different traits by the same method, and (3) for different traits by different methods. Construct valid measures show higher same trait-different methods correlations than the correlations obtained for different traits-different methods and different traits-same method correlations.

N—The symbol commonly used to represent the number of cases or subjects in a group.

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 standardization sample(s) for the test. Because 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 occupation, age, grade, gender, or ethnicity.

part-whole correlation—A correlation between one variable and another variable representing a subset of the information contained in the first; in test construction, the correlation between a score based on a set of items and another score based on a subset of the same items.

percentile (***P***)—A point (score) in a distribution at or below which fall the percent of cases indicated by the percentile. A score coinciding with the 35th percentile (*P*35) is interpreted as equaling or surpassing 35% of the persons in the group, such that 65% of the performances exceed this score. "Percentile" has nothing to do with 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 chief drawback of using a raw score-to-percentile conversion is the resulting inequality of units, especially at the extremes of the distribution of scores. That is, as the distribution of scores approaches a normal shape, observations cluster closely at the center and scatter widely at the extremes. 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 the same or similar scores, causing a one- or two-point raw score difference to result in a 10- or 15-unit percentile difference. Such clustering is most likely to occur on short tests, where only a limited number of scores are possible. 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—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 testings 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, 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 areas 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, because 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—An IRT technique 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, uninterpreted 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 three measure somewhat different aspects of reliability, but all are properly spoken of as reliability coefficients. See Alternate-Form Reliability, Split-Half Reliability Coefficient, Test-Retest Reliability Coefficient.

representative sample—A 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.

split-half reliability coefficient—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. This type of reliability coefficient is inappropriate for tests in which speed is an important component.

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 (68.3%) 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 because 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 deviations, 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 *SE*s 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 2.58 *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,

Standard Score = (Score – Mean) / 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 (standard test)—A test designed to provide a systematic sample of individual performance, administered according to prescribed directions, scored according to 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.

statistical 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.

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," because the validity of a test is always specific to the purposes for which the test is used. The term validity, then, has different connotations for various types of tests and, thus, a different kind of validity evidence is appropriate for each.

(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) 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.

(3) construct validity—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. Tests of personality, verbal ability, mechanical aptitude, critical thinking, etc., are validated in terms of their construct and the relation of their scores to pertinent external data.

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

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