



TABLE OF CONTENTS

INTRODUCTION	1
Directions for Completing Chapter Quizzes	2
CHAPTER 1 - OVERVIEW	3
Basics	3
History.....	5
Aptitude Testing.....	6
Standardized Achievement Tests	7
Personality Tests	8
Competency Tests.....	9
TESTS AND TESTING OVERVIEW QUIZ	12
CHAPTER 2 - VALIDITY AND RELIABILITY	13
Validity	13
Reliability	18
VALIDITY AND RELIABILITY QUIZ	21
CHAPTER 3 - EXAM PLANNING	22
Exam Plan Decisions.....	22
Review	23
Test Types	24
The Test Development Process	27
Ten Suggestions for Item Writing:	29
EXAM PLANNING QUIZ	32
CHAPTER 4 - TEST ADMINISTRATION.....	33
Test Loan.....	33
Process Summary	36
Administering the Test	36
Examinees With Disabilities.....	38
TEST ADMINISTRATION QUIZ	41
CHAPTER 5 - TEST SCORING.....	42
Introduction	42
Concept of Standard Scores.....	42

Statistical Terms	44
More statistical terms	46
TEST SCORING QUIZ	48
APPENDIX A	A
TEST TYPES.....	A 1-6
APPENDIX B	B
READABILITY MEASURES	1
APPENDIX C	C
SECURITY AGREEMENT	2-4

TESTS AND TESTING

INTRODUCTION

This training manual is written in a format and style designed to help you learn the material more easily and retain what you have learned longer than other manuals you may have used in the past. You will need to answer questions about what you read, as you read. The questions will help you to be sure that you understand the important ideas in the text before you go on. Answering the questions is simple. We provide the correct answers so you can check your responses.

On each page you will find information you will need to know to be effective in selection work. Numbered blanks scattered through the text mark the places where you need to fill in the missing word to make the sentence correct. The best word for some blanks will be obvious, for other blanks, you may need to read the rest of the sentence or even the next sentence to get the meaning and know the word to fill in the blank. The correct answers are in the shaded areas at the bottom of the page.

The best way to use this manual is to cover the answers as soon as you turn to a new page. Write the answers in the blanks as you read. When you reach the bottom of the page, remove the cover and check your answers with the correct answers in the shaded area. You may change any wrong answers you filled in, so that you can review the material knowing that the answer in the blank makes the sentence correct.

At the end of each chapter you will find a review quiz with questions that can be answered in four to six sentences each. Use the space provided to type your answer to each question.

The SPCP Administrator or designated representative will score the quizzes. The scoring is based on evidence that students can:

- express thoughts clearly in writing;
- demonstrate knowledge of the principles and concepts in the chapter;
- support opinions with facts.

After the review of all the quizzes you will be notified of the results.

Directions for Completing Chapter Quizzes

At the end of each chapter you will find a quiz which can be answered in four to six sentences. Write enough to demonstrate your knowledge on the subject.

Use your computer to complete the quizzes, recording your name, department/institution, date and answers. When you have completed all quizzes for the manual e-mail them to Jennifer.Clayman@state.co.us. A "report card" will be issued when all quizzes for all manuals have been completed.

Ms. Clayman will notify you of the results of your quizzes. Do not be surprised if you are requested to rewrite an answer and be more specific or elaborate. *Putting ideas and concepts in written form which can be communicated to readers* is an important competency for human resource practitioners.

CHAPTER 1 - OVERVIEW

Basics

The purpose of this manual is to help students understand the theories and principles of occupational testing for employment decisions. Although every part of the material may not apply to every Human Resources position, understanding the principles will help you perform better in most Human Resources assignments. Much of what you will learn pertains to HR practices within the state of Colorado, but certain basic principles are relevant throughout the field of industrial organizational psychology of which employment testing is a part.

When you finish this manual and your on-job training you should be able select the most appropriate type of test, develop a set of test items to accomplish your objective, and use the test(s) properly to fill a vacancy.

Testing is a way of predicting which candidates will be:

- successful in the position for which they are applying,
- which will be barely acceptable, and
- which candidates are likely to be unsuccessful.

If the employer could give all candidates on-the-job tryouts, it would be relatively easy to select the best workers in the position, and hire them permanently.

From the applicants' point of view, one problem with an (1) _____ is the feeling of failure and rejection if they are not (2) _____. A better way to make employment decisions is to measure the candidates' knowledge, skills, abilities and other competencies in areas that are necessary to be (3) _____ on the job. Measuring KSAs and competencies in advance of employment decisions is called (4) _____.

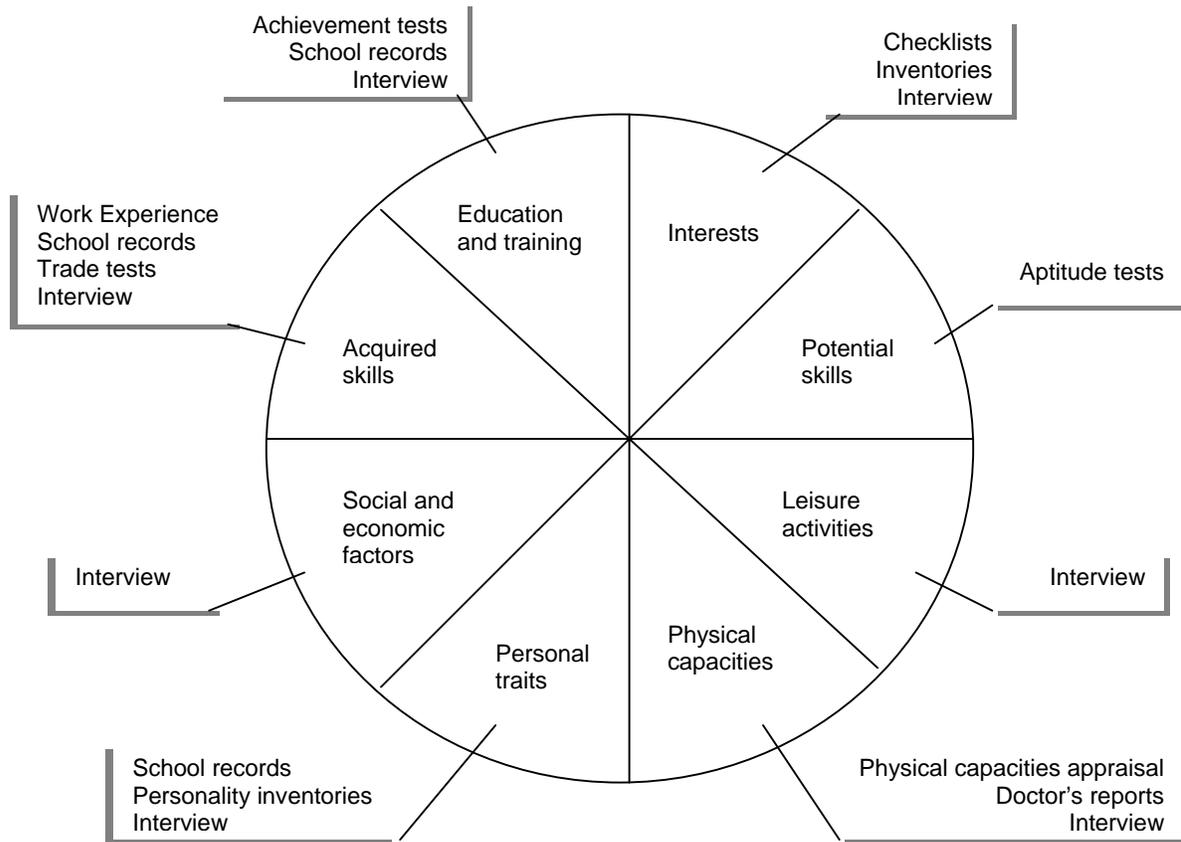
Some of the many types of employment tests are:

- training and experience
- performance
- personality
- written objective.

Each has its own advantages and disadvantages. This manual will discuss these test types, when it is appropriate to use each test type, and the advantages and disadvantages of using that test type.

1. on-the-job tryout 2. hired permanently 3. successful 4. testing

Appraisal of an Individual Figure 1



In Figure 1 the attributes being measured are inside the circle and the methods of measuring the attributes are on the outside with a line pointing to the appropriate section in the circle.

In a manual about tests and testing, one would expect the main topics of interest in this illustration to be Aptitude tests and Achievement tests. Actually, **all** of the topics shown, including interviews, personality inventories, and evaluating school records, are forms of (5)_____. The “whole person” approach to making selection decisions retains the traditional cognitive (6)_____ tests and incorporates personality or social skill measurements in addition. Social skills are represented in the (7)_____ trait section of the graph. Interviews, school records, and personality inventories measure personal traits. Which of these three is a test? (8)_____ of these are tests.

5. tests 6. abilities 7. personal 8. All

History

Why study the history of testing? It is important because hidden between the names and dates there is essential information which is reflected in current testing issues and practice.

Testing is a comparatively young science. Sir Francis Galton, an English biologist, is credited with launching the testing movement. His interest in human heredity led Galton to realize the need for measuring the characteristics of related and unrelated persons. In 1884, Galton started (9)_____ physical traits and simple sensorimotor functions of volunteers to establish the exact degree of resemblance between parents and offspring, brothers and sisters, and so on. Galton's records became the first large, systematic body of data on individual differences in simple psychological processes.

James Cattell, an American psychologist, began exploring differences in reaction time. Contact with Galton reinforced Cattell's interest in measuring individual (10)_____. Cattell thought a measure of intellectual functions could be obtained through tests of sensory discrimination and (11)_____ time. In 1890, Cattell was the first to use the term "mental test."

Alfred Binet criticized most available tests as being too largely sensory and as concentrating unduly on simple, specialized abilities. In 1905, the Binet & Simon scale was devised to measure judgment, comprehension, and reasoning, which Binet considered to be three essential components of intelligence or "mental level." In 1916, the Stanford- (12)_____ scale was first used to establish intelligence quotient (IQ) or the ratio between mental age and chronological age. Recent revisions of the Stanford-Binet are still widely used for clinical and counseling purposes.

When the United States entered World War I in 1917, there was a need to classify the million and a half military recruits for administrative decisions such as: assignment to different types of service, admission to officer-training camps, and acceptance or discharge from the service. Arthur Otis contributed the first group intelligence (13) _____. A major contribution of Otis's test was the introduction of multiple-choice and other "objective" item types. This test was revised into reading and non-reading versions, the Army Alpha and the Army Beta. Both were suitable for administration to large (14) _____.

After World War I, the Army Alpha and Army Beta were released for civilian use. Army Alpha and Army Beta went through revisions and are still in use. They also served as models for most (15)_____ tests that followed. By redesigning a version of the Stanford-Binet intelligence scale to allow administration to groups, psychologists were able to test almost two million recruits from 1917 until the end of World War I. The Army testing project and the body of data it generated legitimized the use of standardized, group-administered tests as tools for making selection and placement decisions.

Several categories and sub-categories of tests have evolved from these historical beginnings. A few major categories are described in the following pages.

<i>9. measuring</i>	<i>10. differences</i>	<i>11. reaction</i>	<i>12. Binet</i>
<i>13. test</i>	<i>14. groups</i>	<i>15. intelligence</i>	

Aptitude Testing

In the 1920s the application of group intelligence tests far outran their technical improvement. Group intelligence tests were still crude instruments and this fact was often forgotten in the rush of gathering scores and drawing practical conclusions from the results. The claims made for the Army Alpha and the interpretation of test results based on heredity did not go entirely unchallenged. In *The New Republic* (1922) Walter Lippmann wrote, "I admit it. I hate the impudence of a claim that in fifty minutes you can judge and classify a human being's predestined fitness for life. I hate the sense of superiority which it creates, and the sense of inferiority which it imposes."

Although intelligence tests were originally designed to sample a wide variety of functions in order to estimate an individual's general intellectual level, it soon became apparent that such tests were quite limited in their coverage. Not all important (16)_____ were represented. In fact, most intelligence tests were primarily measures of verbal ability and to a lesser extent, the ability to handle numerical and other abstract and symbolic relations. Gradually psychologists came to recognize that the label (17)_____ was a misnomer because only certain aspects of intelligence were measured by such tests.

Even prior to World War I, psychologists had begun to recognize the need for tests of special aptitudes to supplement the global intelligence tests. A process called "factor analysis" indicated the presence of a number of relatively independent (18)_____ or traits. One of the outcomes of factor analysis was the development of *multiple aptitude batteries*.

⇒ The term **aptitude test** has been traditionally employed to refer to a test which measures clearly defined segments of ability. **Aptitude** is a measure of ability to learn in areas which apply to performing tasks. **Aptitude** (19)_____ can then be said to measure **potential** to (20)_____ in a specified occupation.

⇒ In comparison, the term **intelligence test** means a broad test yielding a single global score such as IQ.

In place of a total score such as IQ, aptitude testing results in separate scores for such traits as mechanical aptitude, verbal comprehension, numerical aptitude, spatial visualization, reasoning, and perceptual speed. (21)_____ batteries provide a suitable instrument for making employment selection decisions. Job analysis indicates which traits are required to perform a specified occupation. An aptitude test measures the candidates' potential abilities in the identified traits. The identified (22)_____ as a group are called profiles. A profile is the combination of identified traits that a person should have to succeed on the specified job. Aptitude is a measure of potential, and some workers do not perform up to their potential. It is important to remember that the person with the highest profile score is more **likely** to succeed, but it is not automatically assured.

16. functions 17. intelligence test 18. factors 19. tests 20. learn or succeed 21. multiple aptitude 22. traits

Nearly all Multiple Aptitude Batteries have appeared since 1945. The work of the military psychologists during World War II should also be noted. Much of the test research conducted in the armed services was based on (23)_____ analysis and was directed toward the construction of (24)_____. The Armed Services Vocational Aptitude Battery (ASVAB) and the Department of Labor's General Aptitude Test Battery (GATB) are two of today's most respected multiple aptitude batteries. Both are built on the foundations of the World War II (25)_____ psychologists.

Standardized Achievement Tests

Shortly after 1900, the first standardized tests for measuring the outcomes of school instruction began to appear. By 1930, it was widely recognized that essay tests were not only more time-consuming for examiners and examinees, but also yielded less reliable results than the "new" objective tests.

Standardized (26)_____ utilized measurement principles developed in psychological laboratories. Examples include scales for rating the quality of handwriting and written compositions, as well as tests in spelling, arithmetic computation and arithmetic reasoning.

The publication of the Stanford Achievement Test in 1923 provided comparable measures of performance in different (27)_____ subjects. An individual's scores were compared to the scores of a single normative group. Achievement tests are used not only for education purposes. They are also used in the selection of job applicants.

The purpose of norms is to show an individual's relative standing in some appropriate reference group. For example, (28)_____ for the Department of Labor's General Aptitude Test Battery (GATB) are based on what the Department of Labor calls the General Working Population Sample. An individual's scores are compared to the norms of the most working people. A sample of 4,000 workers was chosen to be representative of the work force. The General (29)_____ Population Sample is the reference group in which each GATB aptitude is standardized to have a mean (average) score of 100 for workers in general.

State, regional, and national testing programs established standardized tests. Measurement of scholastic (30)_____ by students continues to be a prime factor in evaluating school systems, teachers, and student progress. The reliance on standardized achievement scores to evaluate education leads to teachers "teaching the test," and does not account for many factors outside of standardized tests.

Recently, the technical aspects of achievement tests increasingly resemble those of intelligence and aptitude tests. Efforts increased to prepare achievement tests that would measure the attainment of broad educational goals. There is less emphasis on recall of factual minutiae as the content of achievement tests more closely resembles that of intelligence tests. Today the difference between these two types of tests is chiefly one of degree of specificity of the content, and the extent to which the test assumes prior instruction.

23. factor	24. multiple aptitude batteries	25. military	26. achievement tests
27. school	28. norms	29. Working	30. achievement

Personality Tests

Personality tests are concerned with the nonintellectual aspects of behavior. Personality tests most often refer to measures of such characteristics as emotional adjustment, interpersonal relations, motivation, interests, and attitudes.

“**Can** he/she do it?” can be answered using some of the many tests which have been developed to assess basic (31)_____. If the answer is, “no,” there is no reason to continue. If the answer is, “yes,” the decision maker must answer the second question, “**Will** he/she do it?” Vocational interest questionnaires establish an order of preference for (32)_____ fields. Interests that have persisted over several years form an important component of a person’s attitude. (33)_____. and motivation are most often measured by *self-report questionnaires*.

Another approach measuring personality is the *performance or situational test*. In performance or situational tests, the subject is given a task to perform . Most of these tests simulate every day situations. The real purpose of performing the (34)_____ is often disguised. These tests are designed to measure behavior such as:

- cheating
- lying
- stealing
- cooperativeness
- persistence

With a large body of data, it is possible to find *norms* for the group and develop objective, quantitative (35)_____. A situational test developed by Harshorne in 1930 was standardized on school children. Other tests have attempted to measure personality with situational tests, but the interpretation was subjective and required a highly trained interpreter.

Projective techniques are worth mentioning even though they are seldom used as employment selection instruments. This category includes:

- problem solving with few guidelines
- free association
- sentence completion
- drawing
- arranging toys to create a scene
- interpreting pictures or inkblots

All available types of personality tests present serious difficulties. (36)_____ tests have not progressed as far as aptitude (37)_____. The special difficulties encountered in the measurement of personality account for the slow advances in this field of testing.

31. aptitudes or abilities	32. occupational or vocational	33. Interests	
34. task	35. scores	36. Personality	37. tests

Competency Tests

Competency based tests have been used extensively in education. Educational competency makes up a large part of the body of competency literature. Most often a set of standards in various school subjects defines the level of competency a student should have learned before going on to the next grade level. For example, certain school systems grant high school diplomas only to students who pass a competency test at high school graduate level.

For selection purposes, we are more concerned with employment competencies. Employment competencies have been defined in many ways in selection literature. Competencies are:

- The set of knowledge, skills, and attributes that differentiate and define exemplary performance in a given work process
- Knowledge, skills, and abilities demonstrated by organization members that are critical to the effective and efficient function of the organization
- Knowledge, skills, and abilities a person needs to be successful in doing a particular job
- An underlying characteristic of an individual which is causally related to effective or superior (one standard deviation above the mean) performance on the job

At this time, the following definition (also found in the *Job Analysis Manual*, page 40) will serve as a working definition for selection purposes:

⇒ *A competency is an observable behavior that contributes to success on the job.*

In a total competency-based selection process neither seniority nor minimum requirements matter. The focus is on maximum requirements and the applicant who comes the closest to meeting them.

The Federal Office of Personnel Management (OPM) says of competency testing, "Competency-based hiring can facilitate the streamlining of the testing process while producing better candidates. It will shift the value of the job (minimum qualifications) to the individual (competencies) and defines what successful workers are expected to be able to do. The crosswalk conducted to compare Federal jobs to the State of Colorado jobs will aggregate job titles down to fewer competency groupings. These fewer competency groupings will then become the foundation in which tests can be developed or procured. The United States Employment Service (USES) has an array of assessment instruments available for procurement."

Additionally, OPM says, “Traditionally, civil service tests have primarily been cognitive (38)_____ tests. Occupational analyses have long demonstrated, however, that, while ability to reason is indeed important for almost all jobs, interpersonal and social skills are equally important. This *whole person* approach to (39)_____ has led to new measures to supplement the traditional cognitive ability tests. In addition to measures of cognitive abilities, measures of social skills that are crucial for good customer (40)_____, measures of suitability to identify applicants who are likely to exhibit counterproductive behavior on the job, and diagnostic measures to assess training needs or assure appropriate placement are available.”

The “TIPS on the Hiring Process” interactive Web-based training program by consultants Bill Prince and Patrick Milliken lists the following seven competencies to be rated by managers evaluating employment candidates. (This system uses a five point rating scale with values of 1-5.)

Competencies:

- Effective Planning and Organization
- Initiative
- Ability to Learn/Technical Aptitude
- Judgment/Decisiveness
- Systematic Problem Solving
- Teamwork
- Interpersonal Communication

This list is fairly typical of (41) _____ identified as important to a variety of competency-based schemes. Other competencies often mentioned are:

- Honesty
- Integrity
- Conscientiousness
- Conflict Management
- External Awareness
- Innovative Thinking
- Self-direction

The reader should recognize that competencies are linked to Knowledge, Skills, and Abilities through job analysis. Competencies describe behavior in personality terms. Personality measures are less objective than specific job knowledge measures. Research shows that personality measures added to General Mental Ability (GMA) increase the predictive power of tests. In addition, research also shows that adding (42)_____ measures to GMA decreases adverse (43)_____. (Deniz S. Ones, Viswesvaran, and Schmidt; *Journal of Applied Psychology* 1993).

38. ability	39. selection or testing	40. service
41. competencies	42. personality	43. impact

Suitability assessment instruments identify the extent to which applicants have characteristics that make them likely to exhibit counterproductive behaviors once they are on the job. Although the definition of (44) _____ behaviors varies from job to job, it typically includes:

- excessive use of force
- drug and alcohol abuse
- theft
- misuse of employer's property

Even a few unsuitable employees can disrupt work systems and have a widespread effect on morale. In addition to increasing (45) _____ power and decreasing (46) _____, suitability tests are less costly than the psychological interview, polygraph or other procedures commonly used to detect unsuitable behavior tendencies.

44. *unsuitability or counter productive* 45. *predictive* 46. *adverse impact*

CHAPTER 2 - VALIDITY AND RELIABILITY

Validity

Do test takers receiving high scores on the test have a higher probability of doing better on the job than test takers who receive lower scores? Ability to predict is the central question in employment testing.

An employment test that (1) _____ employment performance accurately describes a valid use of a test. A test is not simply valid or invalid. There are varying degrees of (2) _____. Validity depends on how the test is used. The user must determine if the test is appropriate for measuring the candidates' attributes that are necessary for successful work performance.

Validity is the degree to which a test measures what it purports to do. In the case of college entrance exams, the test purports to predict which students will be successful in college classes. Employment selection tests (3) _____ which workers are more likely to be successful on the job. The more accurate the predictions, the higher the degree of (4) _____. Although some tests have a higher degree of (5) _____ than others, no test is 100% (6) _____. That is why we say, "more likely to succeed."

Further discussion of comparisons between predictors and performance requires an introduction to the **statistical concept of correlation**.

A correlation is a comparison of degree of relatedness, usually between two variables (bivariate). For employment selection purposes, most often the two variables are test (7) _____ and work (8) _____.

As the value of one variable increases the value of the second variable increases. This is called a positive correlation. If the value of one variable increases as the second value decreases, the (9) _____ is negative.

As age increases, the incidence of measles decreases. This is an example of a (10) _____ correlation. As age increases, interest in social security reform usually increases. This is an example of a positive correlation. When comparing test scores with work performance, we hope to find a (11) _____ correlation to prove that use of the test is (12) _____.

Correlation is calculated in a mathematical formula. The result is a number called a correlation coefficient.

- | | | | | | |
|-------------|----------------|----------------|--------------|--------------|---------------|
| 1. predicts | 2. validity | 3. predict | 4. validity | 5. validity | 6. predictive |
| 7. scores | 8. performance | 9. correlation | 10. negative | 11. positive | 12. valid |

When the change in one variable is **exactly** proportional to the change in the other variable, the (13) _____ is 1.0. Generally, as one variable (test scores) changes, the other variable (work performance) changes more or less, but not exactly proportional. In conducting criterion-related validity studies with a large sample, a correlation (14) _____ of between .2 and .5 is considered acceptable. Good predictive test scores may forecast success in training with a correlation coefficient of .3 to .75. The closer to 1.0, the stronger the relationship. A perfect negative relationship results in a -1.0 correlation coefficient.

Is a correlation coefficient of .25 meaningful?

If we examine two people and one of them has blue eyes, can we say 50% of the population has blue eyes? If we examine 10,000 people and 5,000 of them have blue eyes, we can have much more confidence in the 50% figure. To know if a correlation coefficient of .25 is meaningful, we must know the number of events (sample size).

Significance expresses the chances that a finding (correlation coefficient) is a coincidence. As in the eye-color example above, the larger the number of people in the research sample, the more significant the findings. As the sample size increases, the value of the correlation coefficient needed to be (15) _____ decreases. A correlation coefficient of .25 may not be significant for a sample of ten, but a correlation coefficient of .25 would be significant for a sample of ten thousand.

Significance is expressed in decimals. A significance level of .05 means there are five chances out of 100 that a finding is a coincidence. We are 95% sure the research proves what it says it proves. A significance level of .05 is usually acceptable in test validation. Often the significance level in test validity research is .01 or even .001 (one chance out of (16) _____ that the finding was by chance or coincidence).

Review: statistical terms

VALIDITY

The test is being used to measure what it was designed to measure.



CORRELATE

Comparison of two (or more) variables.



CORRELATION COEFFICIENT

Degree of relatedness expressed as a decimal



SIGNIFICANCE

Confidence that the results of research did not occur by coincidence

13. correlation 14. coefficient 15. significant 16. 1,000

There are three accepted methods for investigating test validity:

- Content validity
- Criterion-related validity
- Construct validity

Content Validity is the degree to which the evidence supports the assertion that the sample of test items is representative of the knowledge, skills, abilities, and competencies necessary to perform the job and which characterize outstanding as opposed to average job performance. In other words, a test may be considered (17)_____ if the specific knowledge, skills, abilities, and competencies it measures accurately represent the behaviors necessary to perform the job in a successful way.

Job analysis is at the heart of content validity. The job analysis portion of the content validity evidence is thorough if:

1. The sampling of subject matter experts used to generate tasks or elements was representative and of sufficient in size.
2. The knowledge, skills, abilities, and competencies flow logically from the tasks or (18)_____ of the job.
3. The importance of tasks and duties, in evaluating knowledge, skills, abilities, and competencies, was measured.
4. The ratings or measurements of (19)_____ were done by a representative sample of (20)_____ matter experts.

There are three sub-categories of **Content Validity**.

1. **Face Validity** refers to what the test **appears** to measure. The test may or may not actually measure what it (21)_____ to measure. Face validity is most important in creating acceptance on the part of examinees and appointing authorities. The greatest hazard that face validity presents is that a test that is actually inappropriate may be chosen on the basis of its title and apparent content. Many tests with (22)_____ are assumed to measure one factor but are actually measuring another factor such as reading comprehension or oral communication.
2. **Logical or Sampling Validity** is so commonly used that it is often simply called content validity. Like face (23)_____, sampling validity depends on a judgment about the contents of the test. However, it is far more than a cursory look at the items by the user. The job content is broken down logically into content sub-areas, sometimes loosely called factors or attributes. SMEs rate tasks within (24)_____ for criticality of the tasks in good job performance. The test contents are then logically related to the (25)_____ content.

17. content valid	18. duties	19. tasks and duties	20. subject	
21. appears	22. face validity	23. validity	24. factors	25. job

3. **Factorial Validity** can be thought of as a sampling validity strategy (described above) in, which the rational judgment methods of producing job factors and test factors are replaced by a statistical method. The statistical method is factor (26)_____.
- In the job analysis, task elements, critical incidents, etc., are interrelated and factor analyzed. The SME ratings of criticality depend on SME opinions about the extent the task contributes toward being a good worker.

The rest of the content validity approach depends on whether:

- the test contains enough items in each factor to measure its KSAs and competencies reliably,
- the factors are logically measuring the relevant KSAs and competencies,
- the (27)_____ items appear capable of discriminating between acceptable and outstanding work performance.

Test items should principally cover the areas shown by the (28) _____ to be most important or critical to good job performance. It is this characteristic that provides indirect evidence that higher scorers will generally do (29)_____ on the job.

Criterion-Related Validity is also referred to as empirical validity and sometimes as statistical validity. Regardless of name, this type of validity is a correlation of a predictor (test scores) and a direct measure of the behavior (job performance, school grades, etc.) the test was designed to predict.

The direct measure of behavior is called the criterion (plural is criteria). The (30)_____ is the standard against which the accuracy of the (31)_____ is measured. In validating employment selection tests the criterion used is usually supervisors' evaluations of work performance. The criterion (supervisors' evaluations) shows us which workers are actually good workers and the correlation (comparison) shows us how accurately the predictor (test scores) forecasts which would be good workers.

In Criterion-Related Validity, job analysis is important but it is not the only source of validity as it is in (32)_____ validity.

There are two ways to collect data for criterion-related validity studies:

- concurrent
- predictive (longitudinal)

Concurrent validity research means that the criterion scores are collected at approximately the same time as the predictor/test scores. A concurrent (33)_____ research design usually calls for testing incumbents in a specific occupational group, and collecting work performance evaluations from their (34)_____. The test scores (one variable) are then correlated with the supervisory ratings (the second variable). A significant positive correlation means the use of the test was valid. A (35)_____ correlation says, "if we had tested these workers before we hired them, the test would have predicted well which ones would turn out to be good workers because most of our good workers score high on the test."

26. analysis	27. test	28. job analysis	29. better	30. criterion	31. predictor
32. content	33. validity	34. supervisors	35. positive		

A Predictive or Longitudinal Research design takes place over a period of time. In a Predictive design the research sample consists of applicants who are tested before they are hired. Ideally, all of the workers are hired regardless of their test scores. Realistically, only some of the applicants are hired. In either case, the hiring is decided by a method other than the specific test scores. After a period of time required for new hires to complete training and become fully competent workers, usually about six months, supervisors' ratings or production records are collected for the research sample members.

The supervisors' ratings or production records serve as the (36)_____ in the correlation against test scores. Predictive studies give researchers a fairly accurate simulation of how successfully a test will separate good workers from poor workers in an applicant pool. From this simulation, researchers can show at what rate the test's predictions are true over time. Just how accurately the test predicts is shown in the correlation (37)_____ and the level of (38)_____.

In both sub-types of criterion-related validity research, the supervisory ratings **must** be reliable and accurate for the results to be meaningful. It is a good practice to have the minimum number of raters possible rate the entire sample of workers. Given one or two workers to evaluate, most supervisors rate them as average or slightly above average. The larger the number of workers to evaluate, the more likely the supervisor will spread the ratings from the highest rating category to the lowest category. The more the ratings on the criterion are spread, the greater the opportunity for the predictor to identify which sample members are (will be) good workers.

A weakness of concurrent criterion-related studies is that workers with only marginal ability will have quit or will have been discharged from the job. The incumbents tested will not represent the full range of (39) _____ to do the job. This is called **range restriction**. Range restriction means the supervisors' ratings will probably range from excellent to average (below average workers will no longer be on the job) and most of the (40)_____ scores will be clustered between medium to high. **Range** refers to the lowest score subtracted from the highest score. Test scores and supervisors' ratings both have ranges.

A weakness of predictive criterion-related studies is the time involved. The waiting period between testing applicants and the evaluations at the end of training to see if the predictions prove true can be six months to a year. Meanwhile many employment decisions must be made without knowing the results of the research. Also, there is only criterion information (supervisors' ratings) for the applicants who were hired and stay through the training period, therefore there is (41)_____.

Construct Validity is the third type of validity after content and criterion-related. A construct is a personality trait. For example, we can not directly observe IQ. We develop tasks we think measure intelligence in an indirect way. Construct validation is the process of gathering data to support our contention that a test of intelligence (or whatever trait) is actually a reflection of the attribute it is designed to reflect.

36. <i>criteria</i>	37. <i>coefficient</i>	38. <i>significance</i>
39. <i>abilities</i>	40. <i>test</i>	41. <i>range restriction</i>

Sometimes data are collected in a criterion-related design. Another means of collecting data in construct validity is called **cross validation**. In cross validation, the researcher identifies an attribute that is needed to be successful on a certain job. The researcher then develops a test to measure that (42)_____. The members of the research sample also take another test from a separate and independent source that measures the same attribute. If use of test XYZ is valid and the same research subjects who score high on XYZ also score high on test ABC, there is a positive correlation and we can assume that use of test ABC is also valid.

The key to construct validity is to accurately identify and describe the construct needed to succeed on a specified job. In the past, construct validity did not play a large part in employment testing. With the increased interest in competency testing there is a greater emphasis on personality traits such as integrity, conscientiousness, and (43)_____.

Reliability

In a chapter about Validity and Reliability, reliability receives less attention. Don't misunderstand, reliability is an important characteristic of a good test.

Reliability refers to consistency of scores. Consistency of scores obtained by the same person when reexamined with the same test on different occasions tell us how (44) _____ the test is. Reliability also refers to consistency of scores obtained with different sets of equivalent items, or other variable examining conditions.

In its broadest sense, test reliability (consistency) indicates the extent to which individual differences in test scores are attributable to "true" differences in the characteristics under consideration and the extent to which they are attributable to chance errors. If we think of each person's test score as being made up of a "true score" component and a measurement error component, reliability can be defined as the fraction of the observed test score that is "true score".

If the observed score is 50% (45)_____ and 50% is error, its reliability is 0.5. In many tests reliability reaches .75 or even higher. If a test is measured without error of measurement, its reliability is (46)_____. Even with optimum testing conditions, no test is perfectly reliable. A statement of its reliability should accompany every test.

A test may be reliable but not valid because it does not measure correctly, but does it consistently. A test can **not** be valid if it is not reliable because being reliable is a part of validity. As confusing as this may seem at first, think of a ruler that is 11 inches long, but is incorrectly marked as 12 inches long. Everything you measure with this ruler will be measured consistently (47)_____. On the other hand, if a ruler or test used is valid and does what it says it is going to do, then it follows it must also be reliable.

For the mathematically inclined, validity (correlation coefficient) can be no larger than the square root of the reliability. Try this on your calculator. A test with a reliability of .4 cannot have validity greater than about (48)_____.

<p>42. attribute or construct 43. fill in a competency you think is often important for success 44. reliable 45. true score 46. 1.0 47. wrong 48. 0.63</p>

Reliability Types

There is no direct way to be certain what fraction of the observed score is “true score”. Several ways have been devised to determine this indicator of (49)_____.

The test-retest approach has the most intuitive appeal. (Remember face validity?) Essentially, the test constructor gives the test to the same sample of persons on two separate occasions and calculates the correlation of test scores on occasion one with the test scores on occasion two. The correlation is an estimate of the (50)_____ of “true score” contained in the observed score.

There are two major reasons for **not** using the test-retest approach. First, for some tests the trait measured is not stable over time. Traits like anxiety and depression can change from time to time and the observed reliability will appear weak. Second, on achievement tests, practice can increase “true scores.” Thus for tests where the “true scores” can (51)_____, test-retest reliability will be underestimated.

The other approach to determine reliability is the internal consistency method. Internal (52)_____ means the test items are being compared with each other.

There are three sub-types of internal consistency methods:

- parallel-form
- split-half
- KR-20

Parallel-form reliability is the correlation of two tests that are constructed of different but very similar items designed to measure exactly the same trait in exactly the same way. Like the (53)_____ method, the correlation between parallel forms of a test is a theoretical estimate of the fraction of “true score” in the observed score.

The two parallel forms are given on separate occasions. As in the test-retest approach, the timing of the second test is key. If the time between tests is too short, the person may remember how he/she answered the first time. If the time between tests is too long, the personal trait being measured may change. The researcher should also be aware that no two tests are exactly parallel.

Split-half reliability is similar to parallel-form reliability except only one administration is required. The test is divided into (54)_____ halves, usually odd numbered and even numbered items. By correlating one half of the test with another, we are obtaining a measure of the extent the two halves of the test measure the (55)_____ trait.

KR-20, the Kuder-Richardson reliability measure, is another commonly reported reliability measure. If all possible ways to split a test in half were used to calculate all the possible split-half reliability coefficients the average split-half reliability across all these splits would be formally equivalent to the (56)_____ formula. KR-20 is a better measure of split-half reliability. Like the others, this measure can be shown to be an estimate of the “true score” fraction of the (57)_____ score.

<p>49. reliability 50. fraction 51. change 52. consistency 53. test-retest 54. two, 55. same 56. KR-20 57. observed</p>

Variables Influencing Reliability

- Test length
- Speededness
- Range of scores

If a test has two items, the examinee might guess one question correctly, or on a good day he/she might even guess both questions correctly. If there are 200 questions, the chances of guessing half of them correctly are remote and the chances of guessing all 200 correctly are almost nil. The larger the sample of items, the nearer the approximation of the subjects' "true scores" to their observed scores.

A speeded test is one in which items are often of low difficulty and the scores candidates receive are, therefore, more indicative of how (58)_____ they work than how many items they are able to answer substantively. A pure (59)_____ test would be one for which every item reached is answered correctly. Some clerical tests are close to pure speed tests. For speeded tests the evaluator can expect to see high reliabilities for split-half or KR-20 measures. On a pure speeded test the score on odd numbered items and even numbered items will be almost the same, given that speed is more important than difficulty in the design of the test.

Range of scores refers to the same problems discussed in "range restriction" in the validity section of this chapter. If incumbents are tested, the range between the (60)_____ score and the (61)_____ score will be decreased by the workers who "don't work out," and are no longer on the job. One would presume that workers who were not successful on a specific job would have lower test scores than the incumbents. The incumbents' observed scores will usually be restricted to high scores and the fraction resulting from comparing observed with "true scores" will underestimate reliability.

Responsibility for Validity and Reliability

In this section you have learned that reliability means consistency and the degree to which the test results of a candidate remain the same from one administration to another or one half of a test to the other half is referred to as "true score".

Another way of defining reliability appears in the 1999 edition of Standards for Educational & Psychological Testing (Joint Technical Standards) (Sec 2) *Reliability refers to the degree to which a test is free from error.*

Section 2 also states: Typically, test developers and publishers have the primary responsibility for obtaining and reporting evidence concerning reliability and errors of measurement.

Section 1 states: *Validity always refers to the degree to which evidence supports the inferences that are made from the scores. It is the inferences regarding specific uses of a test that are validated, not the test itself.*

Test developers and publishers cannot know in advance how the user will use the test. The (62)_____ is responsible for showing that the inferences (interpretations/conclusions) based on test scores are valid.

58. fast 59. speeded or speed 60. minimum or lowest 61. maximum or highest 62. user

Name: _____

Department or Institution: _____

Date: _____

VALIDITY AND RELIABILITY QUIZ

1. Explain the advantage Content Validity has over Criterion-related Validity.
2. Explain the advantage Criterion-related Validity has over Content Validity.
3. Which is more important, Validity or Reliability? Please explain.
4. You have developed a clerical-detail test in which the candidates compare a word in the left column with a word in the right column and decide if the two words are exactly the same or different. Five minutes are allowed to make 500 decisions. What method would you use to estimate the Reliability of this test?
Why?
5. Consider this statement, "This is a valid test." Why is the statement incorrect?

CHAPTER 3 - EXAM PLANNING

Exam Plan Decisions

Proper test development always starts with job analysis. Once the Knowledge, Skills, and Abilities (KSAs) and traits that are necessary to succeed on the (1)_____ have been identified and linked to competencies, the test developer is able to make basic test development decisions.

- Which competencies are most important for test development?
- Which test type measures those competencies most accurately?
- Is more than one test or test type needed to measure all of the most important competencies?
- What resources are needed for each test type and what resources are available?

After considering the basics, think about these more specific details before settling on an exam plan.

- What is the level of the position: entry level, full operational, supervisory level?
- How extensive was recruitment?
- Are applicants likely to be widely scattered geographically?
- How quickly must the vacancy be filled?
- How complex is the job? How much on-the-job training is provided?
- Is adverse impact a problem with selection devices used to fill past vacancies in this class?

After preliminary planning, the next decisions formulate an exam plan. The appropriateness of an exam plan is dependent upon several factors that vary from vacancy to vacancy.

- **Factors to be Examined:** The main criterion for using any selection device should be its ability to accurately measure the significant (2)_____ identified by the (3)_____. For example, a position or class involves significant oral communication or interpersonal skills. An exam plan for this position should include an oral exam because other test types cannot assess these factors as well.
- **Level and Nature of Position:** As a general rule, it is not appropriate to use only a Type C (checklist) T&E for higher level (professional/technical/management) positions because other devices (written exams, oral interviews) can usually measure the critical (4)_____ more accurately. On the other hand, service and maintenance classes do not usually require written or oral (5)_____ skills. It is therefore best to avoid exam types such as Narrative T&Es and written exercises that require these skills.
- **Location of Applicant Pool:** When the applicants are widely scattered across the state, it would be wise to consider an unassembled test for the first step. **Unassembled** test means that it is not necessary to test all of the candidates at the same place and often not at the same time.

<p>1. job 2. competencies 3. job analysis 4 KSAs or competencies 5. communication</p>
--

Factors to consider when designing an exam plan (continued from the previous page).

- Number of Applicants: Although oral and performance exams may be preferred devices for some classes, such methods may be impractical or not cost effective for large numbers of applicants. One may use screening devices such as multiple choice exams and T&E Type C to process (6)_____ numbers of applicants. In many cases, it may be reasonable to use costly and time-consuming devices for testing only the most highly qualified applicants. On the other hand, if the size of the applicant pool is (7)_____, ten or less, there is little need for a screening device.
- Adverse Impact: If an exam is known to have adverse impact, the testing specialist should take steps to reduce those effects or search for an alternative selection device that has a lesser degree of (8)_____ impact. Enlisting protected class members as Subject Matter Experts in test development and on oral boards is an accepted practice.
- Frequency of Vacancies: The extent of use of an eligible list influences the amount of resources one should expend on developing a test for that class. For example, a general clerical class may justify the extra (9) _____ to develop a written multiple choice test, but a glass blower class probably would not.
- Impact of Position: Extra attention and effort should be devoted to exam construction for positions where the consequences of substandard performance are great.
- Availability of Banked Exams: A final consideration in the selection of an exam plan is the availability of (10) _____ exam devices. A specialist's preparation time can be reduced substantially by using previously - developed exam materials. If banked materials are used, they need to be reviewed for quality and job relatedness and the **competencies tested should be compared to those identified in the JOB ANALYSIS.**

Review

From the material in this chapter so far, it is clear there are many questions the specialist needs to ask him/herself in developing an exam plan. For each item below, can you explain the significance and implications of the test specialist's decisions?

- * Which competencies should be examined?
- * What is the level and nature of the position?
- * Are the applicants nearby or widely scattered across the state?
- * How many candidates are expected to apply?
- * Has there been adverse impact? Is the class under utilized?
- * How frequently are there vacancies in the class?

6. large 7. small 8. adverse 9. resources 10. banked

- * What are the consequences of substandard performance?
- * Are there banked tests that measure the competencies identified in the job analysis?

Test Types

Early in exam planning, test specialists need to select the type of test(s) they will use. Almost any method of collecting data used to make employment decisions is a test. Review of MQs, interviews, urine analysis, polygraph, are all (11) _____. The test types below are highlights of the most common types. You will find a detailed comparison chart in Appendix A.

Assessment Centers are groups of performance exercises in which observers evaluate the performance of candidates. The performance exercises often include in-basket exercises, panel discussions, and negotiating. Many of the activities involve interaction with other candidates. For this reason, there is often one observer for each candidate, rating his/her assigned candidate as they interact. Usually it takes 4-8 hours for candidates to complete the assessment.

Obviously Assessment Centers are very (12)_____ intensive. The level of validity makes the high cost worthwhile when filling a vacancy in which consequences of error by the incumbent are high in terms of money or peoples lives. Assessment Centers generally are used for filling management positions.

Application Review is the least valid and most subjective method of testing. In this test type, the Minimum Qualifications become the test. Interpreting data from boxes on a form requires highly professional, technical judgment. Most application reviewers claim to have this skill even if H.R. workers in general do not.

Application Review has the advantage of quickly screening a large pool of applicants down to a desired size group for more costly steps in the exam plan. Little or no outside resources are required other than the (13)_____ time.

Canvassing Letter is a method of screening large applicant pools down to the truly qualified candidates who are available and interested in the position. Candidates are asked to respond to a letter if they are interested in a specified position. If the position has unique characteristics, the letter may direct the candidate to complete a supplemental application to determine the applicant's qualifications in greater detail.

The advantage of a Canvassing letter is that applicants who have found alternative employment or for other reasons are no longer (14)_____ in the position are eliminated before the testing specialist applies more costly selection devices.

The greatest disadvantage is the time required for sending the letter, a reasonable length of time for the applicant to complete a (15)_____ application, respond, and the time for the return mail. The testing specialist can save time by making canvassing phone calls, but the rate of contact, even with the widespread use of phone answering machines, is not good.

11. tests	12. labor	13. reviewer's	14. interested	15. supplemental
-----------	-----------	----------------	----------------	------------------

Oral Exams are the best way to determine oral (16)_____ skills. Subject Matter Experts (usually a minimum of three but any number of SMEs can make up a panel) examine the candidates one at a time. In addition to testing time, the human resources specialist briefs the panel on rating scales of responses before the examination starts, and the SMEs complete rating forms after each candidate is tested.

The SMEs rate the candidates' oral responses according to prescribed standards. This method is usually not suitable if there are more than 10 or 12 candidates in the pool, because of the time required of the (17)_____. The rating scales for answers make oral exams more objective than they would be without guidelines. It is difficult to determine if the SMEs are following the scales as they rate the candidates.

In many ways Oral Exams are typical of other exams, and oral exams are the test type most commonly developed by agency test developers. A sample exam development process based on an oral exam but suitable for other test types appears in Appendix A.

Panel Assessment Devices (PADs) are combinations of test types. Most often a PAD results if the job analysis shows both written and oral communication as critical competencies. In this case, the PAD would include a written exercise rated by the panel **and** oral questions also rated by the (18)_____. In some cases candidates are asked to explain their written answers orally. Reading and rating the written exercises is time consuming. Therefore, PADs are most appropriate when the (19)_____ is small.

Performance Exams are used to test for manual skills, skilled trades, and interpersonal role playing. Typing tests are by far the most common performance exam. Performance exams provide a sample of the actual work, making it easy to show content validity. One disadvantage is that few jobs are easily simulated.

Physical Agility tests may be desirable when physical fitness is a factor in job performance. Jobs requiring (20)_____ fitness include fire fighters, police, and other security-related jobs. Hiring Authorities like physical agility tests for security and protection related jobs.

There are many reasons **not** to include Physical Agility in the exam plan. Many incumbents are not able to pass the physical agility test required of recruits. If incumbents cannot pass the physical test, but are doing well on the job, the argument for content validity is weak. There is liability if someone is hurt while taking the test. There are gender and disability issues. A test may be administered in exactly the same manner in repeated administrations in the same location, but the test and test scores lose (21)_____ if the test is given in multiple locations. When scores from multiple testing sites are combined in state wide applicant pools, non-standard scores are not fair for some candidates.

In short, Physical Agility tests are open to challenge for many reasons.

Training and Experience (T&E) tests evaluate a person's work and education history. Often there are no correct or incorrect answers. The information to be evaluated comes from the applicant. T&Es are easy to administer and do **not** require assembled applicants. There are several types of T&Es.

16. communication	17. SMEs	18. oral board or panel
19. applicant pool	20. physical	21. standardization

T&E Type C (checklist) tests are used most often for skilled, labor, and craft positions. If job analysis shows that written comprehension is **not** a critical competency, a T&E checklist can be a good choice of test types. Candidates are requested to rate their backgrounds in certain tasks. The ratings typically range from “have not done this” to “have supervised and trained others to do this.”

It is easy to demonstrate the (22)_____ validity of a T&E checklist.

To discourage inflated rating responses by the applicants, the checklist includes null items called inflation items that are fictitious. Candidates who rate themselves high in non-existent tasks are lying and therefore receive a penalty against their score for each (23)_____ item for which they claim experience.

In addition to the inflated ratings, another disadvantage of the Type C T&E is that the checklist evaluates amount of experience, but not the quality of experience. T&E Type C's are likely to test for skills easily learned on the job. If most people can learn to do a task, there is no point in (24)_____ for it.

T&E Type A (application) is similar to an Application Review. The difference is that the testing specialist evaluates and **rates** information in the applications. If results of the (25)_____ will be included with other scores in the exam plan, the T&E Type A is more appropriate than the Application review, although the validity of these ratings is typically very low. A testing specialist can review Type A T&Es quickly and applicants are not required to (26)_____ for testing. This method works best where there is a (27)_____ consequence of (28)_____.

This test type can be very subjective. Often application entries that appear to be deficient can be explained in a way that qualifies the applicants if the (29)_____ know why they were rejected and have a chance to explain. Applicants must be informed that applications will be (30)_____.

T&E Type N (narrative) instructions ask the candidates to answer training and experience questions in a narrative form. Job analysis should indicate that the job requires (31)_____ beyond the level of making check marks next to the appropriate level of experience as the applicant does when completing a T&E (32)_____.

The time it takes for the applicants to write and for the SMEs to review (33)_____ answers makes the Type N impractical for (34)_____ applicant pools.

Thought and Strategy Paper is another test type in which a panel of SMEs rate written responses. The Thought and Strategy Paper requires candidates to solve problems, formulate plans, and make decisions. The questions in Thought and Strategy Papers give the candidates a problematic situation and ask how they would solve the (35)_____. This approach to testing is most appropriate for professional/technical and (36)_____ positions because they are more likely to involve (37)_____ making on the job.

22. content 23. inflation 24. testing or measuring 25. review 26. assemble 27. low
28. error 29. applicants 30. reviewed or evaluated or read 31. written communication
32. type c 33. narrative 34. large 35. problem 36. management 37. decision

Written Exercises can be used when no written test is available. Written exercises may measure writing ability more than the ability they claim to measure. Written exercises are scored manually and rated by a panel of raters. It may be difficult to establish scoring criteria, but scoring models must be used to provide a basis for rater consistency.

Written Multiple Choice Tests (Written Objective) are the most objective of all tests. There is not any subjective (38)_____ by a (39)_____ of SMEs. Written Objective tests can be scored by electronic scanner. Validity and Reliability are generally very high. Written objective tests are used as screeners when there is a large number of applicants in the pool and there is an appropriate written objective test available. One (40)_____ of written objective tests is the extensive time required to develop each test.

OTHER LESS USED TEST TYPES

Background check
Biodata
Driving record check
Drug Test
Polygraph
Psychological Evaluation
Written Essay

You are filling a vacancy for a Deputy Director of State Division of Public Relations. The duties include giving presentations at clubs and schools, and editing a monthly newsletter. There are not any Deputy Director tests in the test bank. The position must be filled three days from yesterday.

What type of test(s) would you use?

Please see the complete chart of test types in Appendix A.

The Test Development Process

The following explains developing written objective (multiple choice) tests, but most of the information also applies to other types of tests. If you understand developing written objective tests, you will understand how to develop other tests. Written objective test development is **not** decentralized, and most agency human resources staff members will not develop written objective tests. However, knowing the process will help in understanding and developing many other types of tests.

Once the competencies to be measured are identified through a comprehensive job analysis and it is decided to use a written test to measure them, the test constructor should consider and possibly test those behaviors the applicants should be able to do. The critical job tasks linked to competencies will suggest test items.

Example: The competency is "Ability to file printed material quickly and accurately." The applicant must be able to alphabetize, place large numbers such as social security numbers in order, and categorize types of correspondence into existing categories. All of these (41)_____ suggest different types of items. To cover the types of tasks that characterize the competencies, more than one test or test section may be necessary.

38. rating 39. panel 40. disadvantage 41. skills or behaviors

At this point the opinions of several subject matter experts (SMEs) are invaluable. Only individuals skilled and experienced in doing the job will understand what competencies are necessary. Ordinarily SMEs should not write the test. They are experts in the type of job being filled, but they are not testing experts. SMEs may be asked to write (42)_____ initially, but these will generally need to be reworked. The test developer is ultimately responsible for the finished test.

Ideally, the items should be written so that persons possessing the higher levels of the competency (and who will do (43)_____ on the job) will answer the item in one way and those possessing lower levels of the competency will answer it in another way. That is, applicants with higher levels of the competency are more likely to answer the item as keyed.

There are many test types from which to choose. In this section we will describe written (44)_____ choice tests (also referred to as written objective tests) because in many ways they are typical of tests in general.

The standard multiple-choice test item consists of a stem or body, which states the question, and a set of options or choices. One of the options is the keyed or (45)_____ answer and the other options are called distracters or foils. The test takers' task is to select the correct or best alternative from all the (46)_____.

Multiple-choice items have one correct answer and (47)_____ the foils are plausible but (48)_____. Some multiple-choice tests specify that more than one answer might be correct, but the examinee is to select the **best** answer because there are degrees of correctness (*not recommended*). Another way in which multiple-choice items may vary is that some ask a complete question with the options being possible answers. Other questions consist of incomplete statements requiring one of the options to make a (49)_____ and true sentence.

Multiple-choice tests have several **advantages**: (please see the Test Types section above and Appendix A for further information)

- versatility in measuring objectives in the full range from simple to (50)_____
- test taker writing is minimized
- the tester can sample a substantial amount of material in a relatively short time
- scoring is objective with no interpretation required
- (51)_____ can be done electronically
- asking for the **best** option requires the test taker to discriminate among options that vary in degree of correctness
- this format lends itself to item analysis to determine which questions were successful in discriminating between candidates having a specified knowledge and those who did not

Written Objective/Multiple-choice Tests also have **disadvantages**:

- multiple choice tests are (52)_____ consuming to write and produce
- test takers sometimes argue that more than one answer can be true under some circumstances (and they may be right)
- correctly answering a question usually involves recall and recognition but does not require evaluation or synthesis of information

42. questions or test items	43. better	44. multiple	45. correct	46. options	
47. all	48. incorrect	49. complete	50. complex	51. scoring	52. time

Ten Suggestions for Item Writing:

1. The stem should introduce what is expected of the test taker in clear and grammatically correct language, with a vocabulary at an (53)_____ level.
2. Avoid specific determiners. **Specific Determiners** are clues found in the item, which could lead the test taker to the correct answer for that question or other items in the same test. For example:

The type of standardized test used to measure academic achievement is called an:

- a. achievement test because formal classroom learning is measured
- b. case study
- c. special aptitude test
- d. test of intelligence

In the example, option “a.” is much longer than the other options. “Achievement” in the option matches “achievement” in the stem. The stem ends in “an” (Please look at blank number 53. on this page. Did you see the clue?) Any of the three clues could lead the test taker to the correct answer.

3. The stem should measure a single important objective at the intended level of complexity. The stem should contain only relevant information without being unnecessarily long. Watch for stems in one part of the test that provide information needed to answer a question(s) in another part of the test.
4. Stems and options should be positive whenever possible. If a negative is used, it should be emphasized in **bold**. Do **not** use a negative in both the stem and the options.
5. Items should have only one defensible correct or best option. Avoid items that request an opinion. All distracters should be plausible.
6. Avoid overlapping alternatives. For example:
 - a. boy
 - b. girl
 - c. lad
 - d. son
7. Use “None of the Above” as an option only if there is supposed to be an absolutely right answer, Do **not** use “None of the Above” if the instructions are to choose the **best** answer.
8. Avoid “All of the Above” because once the test taker finds more than one correct option, it becomes clear that “All of the Above” is the proper choice. Also remember that if the instruction is to select the **best** choice, there can be only one best choice.
9. Vary the position of the correct option, using some sort of randomized method to arrive at approximately the same number of A, B, C, and D, correct answers.
10. Check and recheck the stem and options to be sure there is no evident bias on the basis of race, gender, age, religion or national origin.

53. *appropriate*

It is critical to use language appropriate for the level of the position or class for which you are testing, (*Please see item 1. on the list of ten item writing suggestions.*) There are software programs which count the words per paragraph or syllables per sentence. The result is expressed as density and reading level is derived from density. Density calculating software is not entirely accurate when applied to multiple choice tests because the options are incomplete sentences.

There are a number of readability measure references listed in Appendix B which will provide you more detail about the different readability measures.

The test constructor must remember that questions must be clear and easy to understand. If the language used in the (54) _____ is complex, the item becomes a reading comprehension question and may not measure the competency (other than reading comprehension) that it was designed to measure.

The number of items needed in a test depends on various factors. The number of applicants expected in a typical pool determines how many items are needed to spread the highest scores from the lowest scores and avoid ties. Generally, the more items the greater the test (55)_____. Most important, the number of items must give examinees sufficient opportunities to demonstrate their abilities in the competencies being measured.

After you have an adequate number of items plus 10% more, and you have organized them and arranged them in some logical order, let someone else read the draft. More than (56)_____ outside opinion is recommended. It is often amazing how readers can misinterpret written questions that are perfectly clear to the writer. Be prepared to throw out your favorite questions, but remember, that is the reason you wrote (57)_____ % extra items. Revise the draft to a nearly final form.

The next step in the test development process is the **Pilot Test**. The more people who take the (58) _____ test, the more significant the findings. Test developers call it Testing the Test. Sometimes a sample (group of research subjects) can be sufficient with an “N” (total number) of 10-20 pilot test examinees. For a test that is used over a broad range of classes at various levels, a proportionately higher (59)_____ is required in the sample. Sample members should be in the same occupation and at or above the classification level the test is expected to measure. The sample should consist of subjects who will not have a reason to take the actual test for employment reasons.

Record the test starting time, and the time each subject finishes. Use this information to set time constraints for the actual test.

⇒ **Power Tests** measure knowledge of a subject or dimension. Speed is not a factor. Set time limits so at least 90% of the examinees complete the test before time expires.

⇒ **Speeded Tests** measure how quickly examinees can complete a task. The items should **not** be difficult. Ideally, there should be enough items to prevent any examinees from finishing in the time allowed. If examinees finish before the time limit, we do not know how many more items they could have done.

54. stem 55. reliability 56. one 57. ten 58. pilot 59. N

Give the pilot study research subjects an evaluation form to complete after the test.

- Was the test too long? Too short?
- Did you understand what the questions were asking? Which questions need clarification?
- Were there any areas you thought had too much (or too little) emphasis?
- Comments?

Each test requires different feedback, but the examples above are typical of the types of information that pilot study (60)_____ members can provide. Be prepared to accept constructive criticism and follow frank advice. Again, it will probably be your favorite test items that will need to be dropped. *If it can be misunderstood, it will be!*

Item Analysis is the breakdown of test results on a question by question basis. Many useful data can be gleaned using (61)_____.

- Was the question too easy or too hard? Compare the number of correct answers to the number of incorrect answers.
- Who got the question right? Ideally the highest scoring examinees (best workers) should give the correct answer and lowest scoring examinees should give an (62)_____ answer. If more of the low scoring examinees get a question right than the number of high scoring examinees answering correctly, the question is suspect.
- Were all of the (63)_____ attractive or were there some that were not selected by anybody? If all or none of the examinees choose a foil, it does not contribute toward separating the potentially good workers from the potentially average or substandard workers.

Item analysis software produces data describing the degree to which an item separates the high scoring examinees from the low (64)_____ examinees (discrimination in a good sense.)

Other issues that can be tested with Item Analysis include:

- Validity (if job performance measures are available)
- Reliability (split-half or test retest)
- Adverse Impact

It is not enough to conduct item analysis for the Pilot Test only. Users should examine their applicant pools and the test results periodically to determine the effectiveness and validity of the (65)_____ they are making based on the test scores.

IS THE TEST VALID FOR THE WAY I AM USING IT?

Once the test developer has adjusted and fine-tuned the items, and prepared the administrators' instructions (*please see the section on test administration*) the test developer is ready to publish the test and use it.

The test development process **does not apply only** to written objective tests. The same underlying principles apply to other tests such as oral boards, written exercises, interviews, and (66)_____.

60. sample 61. item analysis 62. incorrect 63. distracters or foils 64. scoring, 65. inferences or interpretations 66. see "Test Types" and choose any appropriate test

Name: _____

Department or Institution: _____

Date: _____

EXAM PLANNING QUIZ

1. Do you agree or disagree with the statement, "It is not appropriate to use only a T&E Type C for higher level positions?" What are your reasons for agreeing or disagreeing?
2. Item analysis shows that 37 examinees selected option "A," 1 selected option "B," 0 selected "C" and 12 selected option "D". The keyed answer is "A". Why is this or why is this **not** a good test question? (b) Given the same number of examinees selecting each option, but changing the keyed answer to "D," why is this or why is this **not** a good test question?
3. In what situation(s) should you use a power test? In what situations should you use a speeded test?
4. Which test type is actually a combination of two or more test types? Describe how and why it might be used.
5. Describe situations in which you would use an assembled test.

CHAPTER 4 - TEST ADMINISTRATION

Test Loan

After test users have decided on a test type and the competencies they will measure, they have several options:

- use existing tests from the Human Resource Services test vault
- buy commercial tests from test publishers, after review and approval by the Department of Personnel
- select a combination of existing modules to measure the desired competencies
- construct a new test when no existing tests meet the need

Existing tests are housed in the Department of Personnel, Selection System Services (SSS) vault. The test inventory includes all of the active (1) _____ tests (coded WOQAxxx), one copy of each archived test, and copies of oral exams and other exam types used in the past.

Borrow written objective tests by contacting the test loan administrator by phone (303) 866-4229, fax (303) 866-2458, walk-in (not preferred) or E-mail, (preferred). The only people authorized to order test materials are Selection PCP certified signatories and alternates on the test-security agreement.

Test loans can be either temporary or permanent depending on the agreement negotiated between (2) _____ and the user. You may be able to arrange for a permanent loan of a limited number of test booklets for tests you administer frequently. SSS does on-site inventories of all (3) _____ loan test material annually.

Assuming you have done the (4) _____ for the vacancy you wish to fill and have decided which is the appropriate combination of tests and modules, reserve test booklets in advance (before scheduling a test). This will increase the likelihood of getting the test material you need on the date you need it.

Minimum Advance Notice

Call 5 working days in advance for material to be picked up by agency personnel.

Call 10 working days in advance for material to be mailed. If you have not received mailed material, do not wait until the last day to inquire about the shipment.

Call 15 working days in advance for material that must be printed.

Call Selection Systems Services before you schedule an exam.
Avoid a crisis!

1. *written objective* 2. *SSS* 3. *permanent* 4. *job analysis*

Test Use and Security Agreement (Appendix C) describes the terms and conditions for using tests borrowed from Consulting Services. The following **highlights** are important enough to appear in both the text and Appendix C. The trainee should read both.

Terms and Conditions:

The principal signer of this agreement for the recipient accepts personally and on behalf of the named organization the continuing responsibility for carrying out its (5)_____. The principal signer further agrees that all necessary administrative steps will be taken to assure that staff members, special consultants or others who may have access to the (6)_____ supplied will be informed of this agreement and required to comply with it.

Test materials obtained from the supplier will be used only for the official purposes of the recipient in testing candidates for employment and promotion, test research, and development. Under no circumstances will the supplied material be made available to prospective job seekers or other unauthorized persons for purposes of study, copying or publication.

The recipient may develop training (7)_____ which might assist potential examinees in test wiseness and subject matter improvement so long as such materials do not contain any of the actual (8)_____ materials provided by the supplier.

All supplied testing material in the possession of the (9)_____ will be handled and stored in a manner that will prevent unauthorized persons from having access to it and which will ensure that tests are not defaced or damaged. **Booklets that are marked will be erased completely before returning them to Selection System Services.**

Worn, defaced or damaged booklets must be returned to Selection System Services for proper disposal.

Colorado State agencies may not delete, add or modify the test items.

No reproduction of test materials is (10)_____.

The supplier can make no assessment of the adequacy of the recipient's job analysis. It is the responsibility of the (11)_____ to establish the (12)_____ of the supplier's material for the recipient's use.

The signer of this agreement for the recipient, the person who is officially responsible for requesting test material from the supplier, will be the one to whom such material is sent, and will be regarded by the supplier as having responsibility for carrying out the terms of this (13)_____ except that the agency involved has the option of designating named alternates to the principal signer. If this option is used, the names and titles of the (14)_____ will be regarded by the supplier as sharing responsibility with the principal signer for carrying out the (15)_____ of this agreement.

5. terms 6. test materials 7. material 8. test 9. recipient 10. permitted or allowed 11. recipient or user 12. validity 13. agreement 14. alternates 15. terms

At least the principal signer, if representing a State agency, must have successfully completed the Personnel Certification Program in Selection. Materials may be ordered or supplied only to current signers of this agreement.

Commercial tests do not require item development. Instruction manuals are generally provided and often some of the test development is already done. They are easier to use, and may be less expensive to use than local development of a similar test would be.

Commercial tests also have disadvantages. The user must establish that the chosen test is the right one for the job. Tests may not be appropriate even if the job title and (16)_____ title match. Commercial tests may be prone to security problems. Often test publishers do not provide an (17)_____ key. They charge a fee for scoring tests, and there is a cost associated with each use.

Does the test development research data support the test's reliability and its **validity for the purpose intended by the user**? If the test has adverse impact, is a test of similar validity but less (18)_____ available? The user should be able to answer these questions with information provided by the publisher. The test development data should include validity for the job in question, its reliability, and its adverse impact on protected classes, its normative characteristics, readability, and the adequacy of its administration manual.

In choosing a commercial test, identify job competencies with a (19)_____, then consider the cost and whether there is a recurring cost for booklets, answer sheets, or scoring. Contact other state agencies to determine if the use and the cost can be shared for similar jobs. If there is not an existing test available in the SSS test vault, it may be practical to buy a commercial test. Check with SSS first to review the data and consider the possibilities for using an alternate test or developing a new one. If buying a commercial test appears to be the best option, SSS has catalogs of tests from test publishers that will help in your selection. Remember, too, the SSS test development specialist must approve the use of any commercial test by a state agency.

When the sample test or tests arrive, review the items and their logical relationship to the competencies you wish to measure. Next, consider reliability. The test may have little or no predictive validity because, although the job tasks are related to the test items, the tasks may represent features of the job that can easily be (20)_____ by anyone. Or a test with apparent linkage between tasks and test items can be less (21)_____ because it is too (22)_____ for the level of the job.

When reviewing a publisher's test manual and the reported test development data, keep in mind that reliabilities of .4 have been reported for some personality tests and KR20 reliabilities in the .90s are not unreasonable to expect in unspeeded ability tests. (23)_____ much lower than the .70s are not desirable. If the manual also reports validity, consider this, validity (24)_____ coefficients should be in the range of .2 to .5 when supervisors' ratings are used for criteria. Correlation coefficients above .30 are considered good, but more than .40 are suspect. The validities may be reported as corrected for **attenuation**, a statistical correction that corrects for test scores and criterion scores not being perfectly reliable.

16. test 17. answer 18. adverse impact 19. job analysis 20. learned or done 21. reliable 22. hard or difficult 23. reliabilities 24. correlation

To be **transportable**, a comparison of the research job analysis and the job analysis of the job the user is testing for, shows the analyses require substantially the same (25)_____. This means that they should show approximately the same tasks, importance for those (26)_____, and criticality for those tasks. The rationale in defense of the transportability of the commercial test should be reduced to written form.

When you have determined that a written objective test is the most appropriate test (27)_____, you have three options:

- Use an existing test (or combination of tests or modules).
- Buy or rent (a charge for each test use) a commercial test, after approval by SSS.
- Request the Test Development Unit to develop a new test.

Process Summary

For all types of tests:

1. Develop an exam plan.
2. Decide which tests to use to measure the appropriate (28)_____.
3. Examine potential tests to determine whether or not the behaviors that characterize the competencies are, in fact, those sampled by the (29)_____ in the test.
4. If more than one test is appropriate, select the test with the highest validity. If the validities are equal, choose the one with the highest reliability. If (30)_____ impact is an issue, balance both the validity and reliability against the relative adverse impact. If all else is equal consider the readability of the tests and choose the one that is most in line with required reading levels on the job.
5. Submit a report to SSS, describing a desired test and test development data, for approval if you have decided to use a commercial test.
6. Decide the weight of the test along with the other selection devices listed in your exam plan.
7. Administer and score the test. Analyze for adverse impact.
8. If a new test of any type was developed, submit appropriate exam records to the SSS exam bank (*please!*), as required by your agency delegation agreement.

Administering the Test

You will find many useful tips for day to day test administration in the Administrative Support and Related (ASR) Test Administrators' Handbook. A few steps in the process apply to the ASR Basic Test only, but most of the general instructions such as assembling test packets and filling out answer sheet information, are valuable for most group testing sessions.

25. <i>competencies or KSAs</i>	26. <i>tasks</i>	27. <i>type</i>
28. <i>competencies or KSAs</i>	29. <i>questions</i>	30. <i>adverse</i>

The practical information found in the ASR Administrators' Handbook is supported by material in the following sections. Again, the underlying philosophy of Selection PCP is that understanding the principles and theories makes "step by step, how-to" training more meaningful. When something occurs which is not covered in the step by step process, the PCP trained individual will have a resource of knowledge from which to draw.

The objective of test administration is to increase the probability that the applicants hired will be successful on the job. To achieve this goal, the test administration should be **standardized** and give individual examinees an equal opportunity to demonstrate their full abilities.

⇒ **Standardized = Administered Exactly the Same, Regardless of Time or Location of the Administration**

The administrator should review the test instructions. From these, a list of materials needed for the (31)_____ should be compiled: pencils, scratch paper, answer sheets, disqualification sheets, tape recorder, tape, stopwatch, etc. all should be listed for a check-off list. Even the most obvious things such as the test booklets should be included on the (32)_____. If ordering test booklets from the Selection System Services (SSS) test vault, you must reserve them (33)_____ the test is (34)_____.

On the day before the tests, check-off the items as you assemble them. Be sure you have extras and backups on hand in case an item is faulty, such as a page missing from a test (35)_____, or a stopwatch that works until you are ready to start the test. The administrator should attempt to administer the test under substantially the same conditions as it was piloted or its use validated.

Reviewing the instructions ahead of time also serves to locate any unusual testing conditions or steps found in the particular test that might differ from other tests.

The **physical setting** of the room in which the test takes place should be large enough to accommodate the group to be tested. If there is enough room and flexibility, the seating should be far enough apart to (36)_____ cheating, and close enough together so the monitor is able to observe most of the examinees without looking away from others. There should be ample desk or table space for the (37)_____. Personal articles must **not** be allowed on the desk or table during testing. The test room should be well lighted, at a comfortable (38)_____, quiet and free of interruptions.

The (39)_____ setting does not require a great deal of special attention, but it cannot be left to chance. Tests have been challenged in court for such seemingly trivial things as an electric fan in the room causing noise problems. Applicants needing accommodations are required to make the request at least three days in advance. Nevertheless, it is best to use wheelchair accessible rooms for testing whenever possible, in case the candidate did not make the request, but there is a need. Wheelchair-accessible restrooms should be available in the building, but do not necessarily need to be on the same floor as the (40)_____ room.

31. exam or test 32. list 33. before 34. scheduled 35. booklet 36. discourage
37. test material 38. temperature 39. physical 40. testing

Administrator Instructions usually are included in commercial tests. Instructions for administering state developed tests are available in the *ASR Competency Test, Test Administrators' Handbook*. There are special instructions for some specific tests, so check for administrators' instructions when checking out test booklets. Usually the general instructions in (41) _____ will cover most test situations.

Administrators, monitors, and proctors should not try to administer tests extemporaneously. Even experienced administrators will forget to include some (42) _____; if they attempt to administer a test from memory. If all candidates in all locations and all testing session do not receive the same instructions *verbatim*, the test administration will not be (43) _____; some examinees will not have an equal opportunity to compete.

Conduct the testing session in a friendly, calm and competent manner. You are neither entertainer nor buddy for the applicants. Examinees will be put at ease by treating them in a relaxed confident approach. Answer questions about notification of (44) _____ or positions available to the best of your ability. Do not be afraid to say, "I don't know," when that is the correct answer. Do **not** answer questions about test content either before, during, or (45) _____ a test. Most questions concerning test strategy should be referred to the formal instructions, either written or spoken. Answering strategy questions such as, is there is a penalty for guessing, etc. will unstandardize the test in sessions where the question does not come up. A professional manner does not imply an unsmiling, unfeeling, robotic approach. Nothing is more professional than knowing what you are doing.

Deal with complaints on a case-by-case basis. If the complaint is about (46) _____, before the test, these conditions should be changed if physically possible. Complaints occurring during the test are rare. Unless the complaint is about an easily changed physical condition, monitors should advise the candidate to speak with them after the test or outside of the (47) _____ room. Most complaints come after the test. Listen, make notes, but do not comment on the merits of the complaint. Tell the complainant that you will look into the problem, then do it. Report the results to the (48) _____.

Examinees With Disabilities

Reasonable accommodations must be negotiated on a case by case basis. An examinee's disability may or may not affect the administration of a test. Some disabilities do not affect an (49) _____ exam performance at all. Some require minimal physical adjustments. Examinees who use a wheelchair but have full upper body coordination may need only adequate access to the testing room and a table at which they can take the test. More severely disabled persons may require specialized equipment or access, such as "talking calculators" and personal (50) _____ with special software.

As a general rule, where a disabled individual can be reasonably accommodated, but such accommodations might be disruptive to other examinees, schedule the individual for a separate test session. This will not only be less disruptive to the other examinees, it will also be less difficult for the person with a (51) _____.

41. Appendix D 42. instructions 43. standardized 44. results 45. after
46. physical conditions 47. testing 48. complainant or applicant
49. examinee's 50. computers 51. disability

Some disabilities will require that the examinees have another person present to communicate the test materials to them in a form they can understand. Deaf or severely hearing-impaired examinees will require an interpreter. (52)_____ most often will use manual or sign language, but some are skilled at mouthing words in such a way as to reduce ambiguity for those deaf individuals who can lip read, but do not sign. Seeing-impaired examinees will generally require a reader to read instructions and questions and to record the examinee's (53)_____ on the answer sheet. Some learning-disabled individuals, particularly those who are dyslexic, may also require a (54)_____. In general, interpreters and readers should be selected and their services paid for by the testing agency.

Deaf or blind candidates may wish to bring a friend or family member as an interpreter or reader. This should be avoided, as it is conceivable that person might inadvertently help the candidate with answers. Monitoring will prevent this type of collusion in the case of readers. In the case of interpreters for the deaf, only certified interpreters who are trained in ethical standards should be allowed.

For blind examinees, complicated instructions may need to be repeated several times. Math questions and charts will take (55)_____ time. Some of the more commonly used tests in the SSS test vault have been translated to Braille. Many blind individuals do **not** read Braille, so do not assume that (56)_____ is a suitable accommodation for all blind applicants.

There are many compensating devices that may be familiar to the applicant, but not to the person scheduling the test. The person (57)_____ the tests should explain the general nature of the specific test and let applicants suggest how they can accomplish similar tasks on the job and what accommodations they need for testing. Test accommodations should be based on (58)_____ accommodations.

Give applicants ample opportunity to explain how they can perform the essential functions of the job and the corresponding selection exam.

Often deaf or blind applicants do not take much longer to complete tests than non-disabled applicants for the same test. In scheduling, it is wise to plan for twice the time, depending on the nature of the disability, tests, and accommodations. Certain types of tests can take even longer than (59)_____ the time. For longer tests with several parts, splitting the exam into two testing sessions will reduce fatigue.

According to the federal Office of Personnel Management guidelines, when administering **power tests** (speed is not a factor, 90 percent of examinees finish in the time allowed) **do not limit the time allowed test takers with disabilities.**

<p>52. interpreters 53. answers 54. reader 55. extra or additional 56. Braille 57. scheduling 58. job 59. twice</p>

Speeded tests (short, closely timed exercises that nobody quite finishes) reflect the need for speed on the job. Applicants should be tested accordingly. Deaf examinees may require more time for instructions, but should have the **same time limits** as all other examinees on **speeded tests**. Because (60)_____ required by blind test takers are inherently slow, blind applicants can not compete when speed is a factor, and they should not be tested on speeded test, where speed is a job requirement. If it is necessary to administer a speeded module in order to obtain a complete set of scores for an overall test, consider **doubling the time limit**. While (61)_____ the time on one or two modules is not an optimum solution, it is preferable to invalidating an entire test that may include many power tests which are within blind applicants' capabilities.

Learning (62)_____ and motor disabilities must be evaluated on an (63) _____ basis. Disabilities that are not apparent by observation should be documented. Documentation can be obtained from doctors, rehabilitation counselors, etc. Such (64) _____ should indicate the degree of disability and the percent of additional time required to fairly compensate for the disability. Again, there are many factors determining extra time, but one is the issue of (65) _____ tests compared to (66)_____ tests.

If in doubt, give the advantage to the disabled applicant if the resulting test score will truly represent the applicant's ability to perform the job.

60. accommodations 61. doubling 62. disabilities 63. individual
64. documentation 65. power 66. speeded, or the other way around

Name: _____

Department or Institution: _____

Date: _____

TEST ADMINISTRATION QUIZ

1. Describe three things a Test Administrator can do to promote standardization.
2. When administering a test, why is it important to read test administration instructions from the book even though you have already memorized them?
3. The job description says, "Walks across the street to pick up the mail from the department annex and distributes the mail to three locations, one on each floor in the main building." You discover that one applicant who is permanently using a wheelchair applies for the job. Please explain what you would do.
4. At the break between Book 1 and Book 2, a candidate says the room is too cold. Please explain what you would do.

CHAPTER 5 - TEST SCORING

Introduction

Scoring is extremely important because it makes test results meaningful and helps in the interpretation of measurements. In the absence of additional interpretive data, a raw score on any selection test is meaningless.

Colorado Department of Personnel, Human Resource Division endorses two methods of test scoring: 1) Z-scoring; and 2) Percentage. Both methods are available in the current automated applicant data system. Selection System Services recommends the use of the stand or Z-scoring method unless the pool size is very small. The Z-scoring method is designed to (1) _____ exam validity. Regardless of the method selected the raw score is reported in converted score form on a scale of 70 to 100.

Concept of Standard Scores

The concept of standard scores is an accepted (2) _____ to provide some measure of comparability between two tests that may be very different in ranges, means and standard deviations. Transforming scores to Z scores puts both test scores into a metric that has a mean of zero and a standard deviation of one. This allows for comparisons to be made between relative scores on the two tests. The standard score does not change the (3) _____ of the original scale intervals, so that the relative distances between the variant values remains unchanged.

A primary goal of statistical method is the organization and summarization of (4) _____ data to facilitate understanding of the data. A typical first step in organizing data is the tabulation of scores into **frequency distributions**, which group the same or similar scores together into intervals. A simple example of this might be:

Score	Frequency
1	5
2	10
3	20
4	10
5	5

	50

A group of scores may also be described in terms of (5) _____, which provide a single, typical, score that characterizes the performance of the total group. A common measure of this sort is the average, or **mean**, usually labeled M. This is calculated by adding up all raw scores and dividing the total by the number of cases, labeled N. Other measures of central tendency are the **mode** (the score that has the greatest number of occurrences) and the **median** (the score that is in the exact center of the distribution of scores). For the distribution above, the mean is 3, the median is 3 and the mode is also.

1) maximize (2) statistical transformation (3) proportionality
(4) quantitative (5) measures of central tendency

Measures of variability also describe (6) _____ of a group of test scores. The (7) _____ is the distance between the lowest and the highest scores. A better description of the (8) _____ in a distribution is the **standard deviation**, SD, which is a more stable measure of the differences in variability between two sets of scores. When a distribution of scores approximates a **normal** distribution, (9) _____, there is an exact relationship between the SD and the proportion of total cases (N). It is this relationship that is important to the concept of standard scores. The illustration that follows displays the relationship between the proportions of the total sample and the score points corresponding to standard deviations of + or - 1, 2 or 3.

NOTE: THE SYMBOL σ MEANS STANDARD DEVIATION (SD), WHICH IS THE SAME AS "Z".

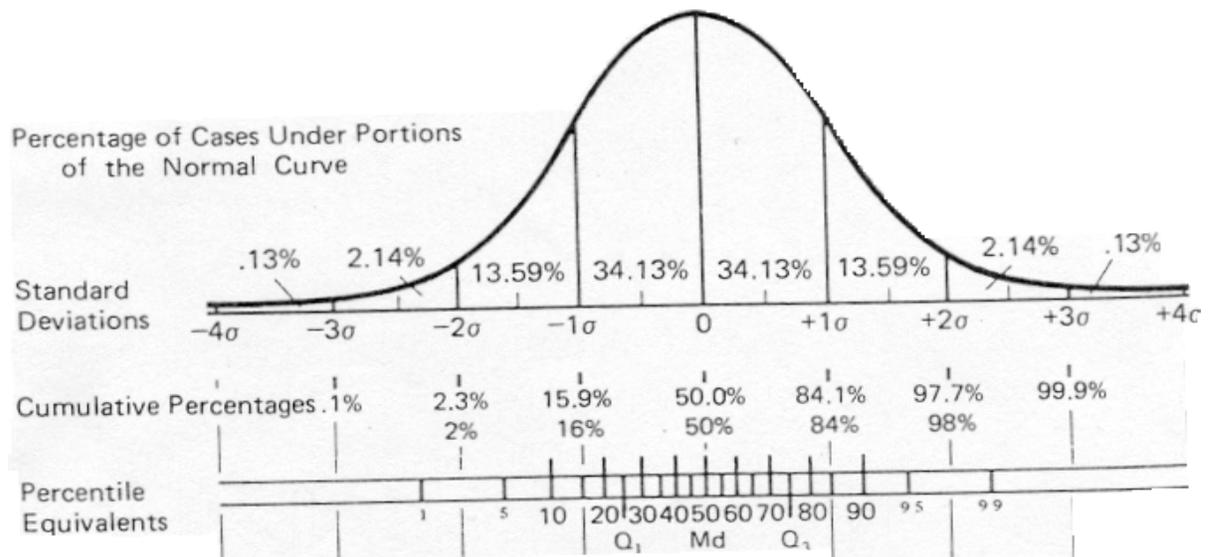


Figure 1

Notice that a Z score of +1 is the same as +1 and that 84.13% of all applicants will have scores equal to or below this value. Only 15.87% of applicants will score above +1Z (1 SD).

(6) characteristics (7) range (8) variability (9) the bell-shaped curve

Scores on selection tests are most commonly interpreted by reference to **norms**, which represent the test performance of the standardization sample (test results of a representative group of people). Individuals' raw scores are then compared to the distribution of the (10)_____ to see where they fall in the (11)_____. Reviewing the distribution would answer the following questions about the individual's score:

- ▶ Does an individual's raw score coincide with the average performance of the standardization group?
- ▶ Is the score slightly above average?
- ▶ What does "slightly above average" mean?

To determine more precisely an individual's exact position compared to the (12)_____ convert the (13)_____ score into a relative measure called a **derived score**. Derived scores indicate individuals' relative standing compared to the norms of the standardization group. Describing the process to derive scores (convert raw scores) requires the use of some simple statistical terms.

Statistical Terms

Frequency Distribution: Tally and group the scores by a convenient interval such as the number of correct answers or correct answers expressed in a range. An example of an interval is a range of 5, like 40-44 or 86-90. Indicate how many subjects scored at each interval, this is called the frequency at each interval. All of the (14)_____ added together equal the total number of people in the sample. Total number of people is usually called **N**.

Central Tendency: The typical or average score of group performance is called (15)_____.

- The most familiar way of describing the representative score is called the **mean**. Mean is a statistical term for average. (16)_____ = sum of the scores, divided by N.
- A second way to express central tendency is called **mode**. Mode is the most frequently occurring score in the distribution tally.
- The third way to express central (17)_____ is **median**. Exactly half of the people scored above the median, and half of the people scored (18)_____ the median. The mean score and the median score are generally close to being the same, but not always.

10. standardization sample 11. distribution 12. standardization sample 13. raw 14. frequencies 15. central tendency 16. mean or average 17. tendency 18. below,
--

Maximum: The highest score.

Minimum: The lowest score.

Range: Subtract the (19) _____ from the (20) _____ score to find the range.

Exercise:

A group of employment candidates took a 50 question written objective test. Their results have been grouped and tallied below.

Test Score	Number of people obtaining this score
45	1
43	1
41	7
40	10
39	8
38	12
37	5
34	2

What is the N of this group? (21) _____

What is the range of scores? (22) _____

What is the mean score? (23) _____

What is the median score? (24) _____

What score is the mode? (25) _____

What is the list of scores and number obtaining each score called? (26) _____

What is the interval used in this frequency distribution? (27) _____

19. minimum 20. maximum 21. 46 22. 11 23. 39
24. 39 approximately 25. 38 26. frequency distribution 27. 1

More statistical terms

Standard Deviation: A method of measuring variability based on the difference between each individual's score and the mean of the group. Are the scores grouped around the mean or are the frequencies spread out over the range? Standard deviation is used to compare the (28)_____ of different groups.

John, an enterprising college student, was researching resorts to pick a spot for his spring break. He found one resort, Sunny Sands, where the mean age of single females was 21.7 years old. At the Palm Paradise, the mean age of single females was 22.0 years. Apparently there was not much difference in the ages of the two groups. Sunny Sands had a high standard deviation (spread of scores), and Palm Paradise had a small standard deviation (grouped around the mean). Sadly, John selected Sunny Sands resort for his spring break. When he arrived, he found that the single female population consisted of six grandmothers age 62 and higher and their eight sub-teen granddaughters.

Percentiles: Percentile scores are expressed in terms of the percentage of persons in the (29)_____ sample who fall below a given raw score. For example, if 28 percent of the persons obtain fewer than 15 problems correct on an arithmetic test, then a raw score of 15 corresponds to the 28th percentile. With percentiles we start counting at the bottom, so that the smaller the number, the (30)_____ the person's test performance. A person at the 84th percentile did as well as or better than (31)_____ percent of the standardization sample.

The 50th percentile corresponds to the (32)_____.

Standard Scores: One type of derived score, found by linear transformation of the raw scores is called z-scores. Standard scores or z-scores are the difference between the individual's raw score and the mean of the standardized sample.

Computation:

$z = \frac{X-M}{SD}$ standard score	M = 60 mean	SD = 5 standard deviation
	Ann's score $X_1 = 65$	Betty's score $X_2 = 58$
	$z_1 = \frac{65 - 60}{5}$	$z_2 = \frac{58 - 60}{5}$
	$z_1 = 1.0$	$z_2 = \frac{-2}{5}$
		$z_2 = -0.40$

28. variability 29. standardization 30. worse or poorer 31. 84 32. median

The (33)_____ procedure yields derived scores that have a negative sign for three subjects who score below the mean. The total range of most groups is not much more than three SDs above and below the (34)_____. Because the range is limited, z-scores must be reported to at least one decimal place in order to provide sufficient differentiation among individuals.

⇒ Remember, one of the reasons for transforming raw scores into any derived scale is to render scores on different tests comparable. On an easy test, most of the class might score 85 out of 100 correct. On a difficult test, most of the group might score 27 out of 45 correct. Knowing a person's z-score tells us how he/she compared to the rest of the group on both tests.

Pass Points: A conventional method of setting pass points (**not to be confused with the ADS procedure**) also uses the standard deviation. The pass point equals $\frac{1}{2}$ a standard deviation below the mean. Peggy Sue has a score of 67 on a test she must pass to be a cheer leader. The class average was 72 and the standard deviation was 8. Peggy Sue is now a (35)_____.

Compensatory Tests: If an exam plan has more than one test or has sections that are scored separately, and one high score can compensate for a low score on another section, the exam is a (36)_____ model. In a compensatory model high scores and low scores “average out” in calculating the total score.

Non-compensatory Tests: In non-compensatory (multiple-hurdle) tests, each test or section must be passed. One score does **not** compensate for another. Non-compensatory tests require passing scores on all of the parts regardless of the (37)_____ score.

33. z-score	34. mean	33. fan in the seats	34. compensatory
35. total	36. frequency	37. deviation	

Name: _____

Department or Institution: _____

Date: _____

TEST SCORING QUIZ

1. Define the following terms in your own words:

frequency distribution

ranking

standard deviation

2. A standard deviation of 5.1 is calculated from the test results of Group A's. Group B's standard deviation is 3.8. Which group is most likely to have the greatest range? Why?
3. What are the only two accepted methods of test scoring in the personnel system? Which is the preferred method and why?
4. Are test norms important and if so why?
5. Referring to the bell curve in Figure 1 describe the characteristics of it. (e.g., mean, standard deviation, etc.) Explain the significance of each characteristic.

APPENDIX A

TEST TYPES

TEST TYPES

TEST TYPE	WHEN MAY BE APPROPRIATE	ADVANTAGES	DISADVANTAGES
Assessment Center	<ol style="list-style-type: none"> 1. Fairly small # of qualified applicants 2. High level, professional/technical, management positions 3. Impact of positions is extremely important/critical 4. Consequences of error are high in terms of money or peoples' lives 	<ol style="list-style-type: none"> 1. Content validity (i.e., job-relatedness) is usually very good because exercises closely approximate actual job situations 2. Typically shows high criterion-related validity 3. Reliability is usually very high because numerous observations are made on each applicant 	<ol style="list-style-type: none"> 1. Requires substantial time to develop 2. Generally requires that each applicant devote a substantial amount of time (e.g., 6-8 hours) for testing purposes. In addition to the assessment time, assessors must spend a minimum of 1 day in training before any ratings are made. 3. Very expensive to administer 4. Appropriate for only a limited number of upper level positions and lower level positions where the consequences of substandard performance are great
Application Review	<ol style="list-style-type: none"> 1. High level, professional/technical, management positions 2. Impact of positions is extremely important/critical 	<ol style="list-style-type: none"> 1. Reduces the size of applicant pool by self-screening 2. Brings "best qualified" to more costly final phase(s) 3. Applicants not required to assemble for testing 	<ol style="list-style-type: none"> 1. Time consuming, must allow time for applicant to complete and return supplemental form and requires a specialist to review and verify information 2. Penalizes applicants who have poor written communication skills
Background Check	<ol style="list-style-type: none"> 1. Job analysis identifies areas where previous behavior would be related to job performance. 	<ol style="list-style-type: none"> 1. Identifies and documents prior behaviors that may affect success on the job. 2. Addresses issues of negligent hiring. 	<ol style="list-style-type: none"> 1. Costly for large number of applicants
Canvassing Letter	<ol style="list-style-type: none"> 1. Experience, education or conditions of employment (i.e., working in isolated area for long periods of time), information that is not available on application form 2. Large # of applicants that meet minimum qualifications but few may have additional education, experience or willingness to work under "special conditions" 	<ol style="list-style-type: none"> 1. Reduce size of applicant pool to only individuals meeting desired criteria 2. Applicants not required to assemble for testing 	<ol style="list-style-type: none"> 1. Time consuming: must allow applicants time to respond

TEST TYPE	WHEN MAY BE APPROPRIATE	ADVANTAGES	DISADVANTAGES
Departmental Promotional Rating (DPR)	<ol style="list-style-type: none"> 1. The total number of raters is small. Ideally, only one supervisor or a group of supervisors rate all applicants for a given position. 2. Applicants tend to come from homogeneous classes. 3. Past performance in certain tasks is critical to the performance of the major duties of the job for which the exam is being conducted. 	<ol style="list-style-type: none"> 1. Past performance within a department is used to predict future performance. 2. Some appointing authorities prefer using supervisors' ratings in determining promotional decisions within the department. 	<ol style="list-style-type: none"> 1. Cannot be used for open competitive exams 2. May perpetuate an existing underutilization situation 3. In most cases, one or more of the conditions listed in the first column cannot be met.
Driving Record Check	<ol style="list-style-type: none"> 1. Job analysis indicates Driving is important KSAP for job. 	<ol style="list-style-type: none"> 1. Reduces the size of applicant pool 2. Only applicants who pass this as first stage continue in process. 	<ol style="list-style-type: none"> 1. Time consuming to complete
Drug Test	<ol style="list-style-type: none"> 1. In accordance with agency policy and ADA 	<ol style="list-style-type: none"> 1. Reduces negligent hiring liability 	<ol style="list-style-type: none"> 1. Costly 2. Requires careful attention to ADA and other laws and regulations
Oral Exam	<ol style="list-style-type: none"> 1. Fairly small # of qualified applicants 2. Job analysis reveals oral communication is significant factor. 3. High level, professional/technical, management positions 	<ol style="list-style-type: none"> 1. Can assess behaviors that are critical to successful job performance that can only be assessed in a face-to-face situation (e.g., oral communication, interpersonal interaction, etc.) 	<ol style="list-style-type: none"> 1. Expensive and time-consuming to administer 2. Difficult to ensure rating standards are being followed 3. Test security is an issue. 4. Relatively high appeal rate. 5. Criterion-related validity may be very low unless great care is taken to develop good questions and keys. 6. May measure speaking ability rather than intended KSAPs
Panel Assessment Device	<ol style="list-style-type: none"> 1. Fairly small # of qualified applicants 2. Job analysis reveals significant factors that are measured by different types of devices. 3. High level, professional/technical management positions 	Refer to Advantages of each device included in PAD.	Refer to Disadvantages of each device included in PAD.

TEST TYPE	WHEN MAY BE APPROPRIATE	ADVANTAGES	DISADVANTAGES
Performance Exam	<ol style="list-style-type: none"> 1. Manual skills (e.g., typing, data entry, equipment repair, etc.) important to position 2. Skilled trades positions 3. Interpersonal skills that can best be assessed in a role-play situation 	<ol style="list-style-type: none"> 1. Provides actual "work sample" for evaluation, making it fairly easy to demonstrate its content validity 2. Many performance exams are less likely to penalize applicants who write or read poorly. 	<ol style="list-style-type: none"> 1. May be time-consuming and/or expensive to develop and/or administer 2. Few jobs are easily simulated in test situations. 3. Difficult to ensure that raters are applying similar standards when behaviors are complex
Polygraph	<ol style="list-style-type: none"> 1. When job analysis reveals work-related factors that may be assessed this way 	<ol style="list-style-type: none"> 1. May be used to substantiate or refute information acquired from other sources 	<ol style="list-style-type: none"> 1. Costly 2. Appeal-prone 3. Very complex issues related to legality
Physical Performance/Agility	<ol style="list-style-type: none"> 1. When job analysis reveals specific requirements critical to job 	<ol style="list-style-type: none"> 1. May be used to identify applicants early in process who may not be able to perform all job duties 	<ol style="list-style-type: none"> 1. Must attend to ADA, other laws to avoid discrimination complaints 2. Costly
Psychological	<ol style="list-style-type: none"> 1. Should not be used 		
T&E - Type A (Application)	<ol style="list-style-type: none"> 1. Large # of qualified applicants, few openings 2. No written exam available 3. Turnaround time is critical 4. Factors or tasks can be adequately measured from application form 5. Applicants who meet minimum qualifications (MQs) are widely distributed with respect to range of their job qualification ("high", "medium", "low") 6. Consequences of error and level of position are not high 7. Applicants are notified of exam type and have opportunity to provide complete information on application or supplemental form. 	<ol style="list-style-type: none"> 1. Can be done in short amount of time 2. Applicants are not required to assemble for testing. 3. Efficient method of "screening" applicant pool 4. Brings "best qualified" to more costly final phase 5. Actual screening may be done by personnel specialists. 	<ol style="list-style-type: none"> 1. Applicants may object to lack of opportunity to complete, if screened out. 2. Penalizes applicants who fail to fill out application thoroughly 3. Applicants may inflate their training and/or experience levels, leading to reduced validity. 4. Does not measure "quality" of experience and other valid predictors of job success 5. May result in taking "over-qualified" applicants only, to final phase 6. Is only a "rough cut" – imprecise

TEST TYPE	WHEN MAY BE APPROPRIATE	ADVANTAGES	DISADVANTAGES
T&E - Type C (Checklist)	<ol style="list-style-type: none"> 1. Large # of qualified applicants, few openings (good screening device) 2. Job analysis reveals finite # of easily identified tasks 3. Measures of "willingness" factors (e.g., work shift; unpleasant environment) important 4. No written exam available 5. Writing ability not a significant factor 6. Skilled/semi-skilled labor, craft, trades positions 	<ol style="list-style-type: none"> 1. SMEs are not required for rating applicants' responses. 2. Applicants are not required to assemble for testing. 3. Relatively quick developmental time provided a comprehensive listing of task statements exists 4. Applicant has indications of his/her suitability for job and may screen self out. 5. Does not penalize applicants who write, read, or communicate poorly (may obtain assistance) 6. Can be designed to minimize adverse impact on protected classes and disabled applicants 7. Relatively easy to demonstrate job-relatedness 8. Exam security is not a problem. 	<ol style="list-style-type: none"> 1. Applicants may inflate ratings, leading to reduced validity. 2. Does not measure quality of experience/training 3. Turnaround time may be slower if T&Es are sent by mail 4. Requires time consuming verification checks
T&E - Type N (Narrative)	<ol style="list-style-type: none"> 1. Fairly small # of qualified applicants 2. Job analysis reveals writing ability is important to the position. 3. MQs require at least some work experience. 4. High level, professional/technical management 5. Can be used as "screener" and also as basis for T&E interview 	<ol style="list-style-type: none"> 1. Can be used to measure breadth, depth, and quality of experience 2. Can be used as a measure of writing ability 3. Can be designed to minimize adverse impact on protected classes 4. Requires no assembling of applicants for testing purposes 5. Because applicants are not required to assemble for testing purposes, Type N T&E is often the only appropriate exam type when recruitment is nationwide (residency waiver granted) for selected higher level positions. 6. Relatively easy to demonstrate job-relatedness 7. Test security not generally an issue 	<ol style="list-style-type: none"> 1. Time-consuming for raters as well as applicants 2. May measure writing ability rather than intended KSAPs 3. Inflation bias may be a problem.

TEST TYPE	WHEN MAY BE APPROPRIATE	ADVANTAGES	DISADVANTAGES
Thought & Strategy Paper	<ol style="list-style-type: none"> 1. Fairly small # of qualified applicants 2. Job analysis reveals writing ability is important to the position. 3. MQs require at least some work experience. 4. High level, professional/technical management 	<ol style="list-style-type: none"> 1. Can be used to measure breadth, depth, and quality of experience 2. Can be used as a measure of writing ability 3. Can be designed to minimize adverse impact on protected classes 4. Requires no assembling of applicants for testing purposes 5. Because applicants are not required to assemble for testing purposes, a Thought & Strategy Paper is often the only appropriate exam type when recruitment is nationwide (residency waiver granted) for selected higher level positions. 6. Relatively easy to demonstrate job-relatedness 7. Test security not generally an issue 	<ol style="list-style-type: none"> 1. Time-consuming for raters as well as applicants 2. May measure writing ability rather than intended KSAPs 3. Inflation bias may be a problem
Written Essay	<ol style="list-style-type: none"> 1. Fairly small # of qualified applicants 2. Job analysis reveals writing ability is significant factor. 3. High level, professional/technical, management positions 	<ol style="list-style-type: none"> 1. Can be used as a measure of writing ability 2. Relatively easy to demonstrate job-relatedness 	<ol style="list-style-type: none"> 1. Requires substantial time to develop 2. May measure writing ability rather than intended KSAPs 3. Usually requires assembling of applicants for testing purposes 4. Test security is an issue when knowledge-based questions are used. 5. Difficult to ensure rating standards are being followed
Written Exercise	<ol style="list-style-type: none"> 1. Large # of qualified applicants 2. No written exam available 	<ol style="list-style-type: none"> 1. Practical and cost effective 2. Easy to administer and score 	<ol style="list-style-type: none"> 1. Difficult to sample full content domain 2. Requires hand scoring 3. Requires a panel of raters to score <p>Difficult to establish scoring criteria</p>

<i>TEST TYPE</i>	<i>WHEN MAY BE APPROPRIATE</i>	<i>ADVANTAGES</i>	<i>DISADVANTAGES</i>
Written Multiple Choice Exam	<ol style="list-style-type: none"> 1. Large # of qualified applicants (good screening device) 2. Written exams available for many classes in exam bank 3. Most of the critical job elements are knowledges 4. Subject matter for the field is not rapidly changing or developing 	<ol style="list-style-type: none"> 1. Reliability and validity are generally very high 2. Very good at assessing factual knowledge 3. Exams available for many classes 4. SMEs are not required for rating applicants' responses. 5. Administration and scoring time are minimized. 	<ol style="list-style-type: none"> 1. Requires substantial time to develop 2. Requires assembling of applicants for testing purposes 3. Test security is an extremely important issue. 4. Can have adverse impact on protected classes

APPENDIX B

READABILITY MEASURES

READABILITY MEASURES

There are a number of methods that may be used to assess the reading difficulty level of any document (test). Some of the common methods are: SMOG Readability, Raygor Readability Estimate, and Fry. Graph Reading Level Index, and Flesch-Kincaid Grade Level Index. Any or all of these may be used. They will not produce precisely the same results, but will generally place reading difficulty at about the same levels.

Many of the word processing software products have a readability measurement tool as part of the software which uses one of these methods to determine readability. Microsoft Word uses the Flesch-Kincaid Grade Level Index determining the Flesch Reading Ease score and the Flesch-Kincaid Grade Level score.

The following references can provide information about readability formulae.

Dale, Edgar and Jeanne S. Chall, "A Formula for Predicting Readability." "Education Research Bulletin", Vol. 27, Jan. 21, 1948.

Flesch, R., "How To Test Readability". New York, Harper and Brothers, 1951.

Fry, Edward, "A Readability Formula That Saves Time." "Journal of Reading", Vol. 11, No. 7 (April 1968), p. 512-16, 575-78.

McLaughlin, G. Harry, "SMOG Grading: A New Readability Formula." "Journal of Reading", Vol. 12, No. 8 May (1969), p. 639-46.

Scully, Sarah V. and Joan Doyle, "E.M.P.O.W.E.R.: Evaluate Materials To Promote Optimal Use of WIC Education Experiences". Massachusetts WIC Program, Department of Public Health, April 1985.

U.S. Department of Agriculture, Food and Nutrition Service, "The Idea Book: Sharing Nutrition Education Experiences". FNS-234, Sept. 1981.

"Readability Testing in Cancer Communications". Reprinted June 1981 by the Office of Cancer Communications, National Cancer Institute, Bethesda, MD.

APPENDIX C

SECURITY AGREEMENT



**TEST USE AND SECURITY
AGREEMENT**
DIVISION OF HUMAN RESOURCES
COLORADO DEPARTMENT OF PERSONNEL &
ADMINISTRATION

The parties to this agreement are:
Workforce Planning & Development
Division of Human Resources
Colorado Department of Personnel & Administration
1313 Sherman, 1st Floor, Denver, CO 80203-2245

and:

Agency Name:

Address:

The agency (hereinafter known as the recipient) and agency official (hereinafter known as principal signer) accept and agree to the terms of this agreement in exchange for permission to obtain test materials from the Division of Human Resources (hereinafter known as the supplier) in the Colorado Department of Personnel & Administration.

This agreement is intended to protect the mutual interests of all private or public agencies and officials who use test materials obtained from the supplier, as well as the interests of persons who take such tests, in order that no person may gain special advantage by having improper access to the material. The supplier requires that all recipients who desire to obtain and use confidential testing materials execute this agreement and fulfill its terms as a condition for making its test materials available.

Terms and Conditions

The principal signer of this agreement for the recipient accepts personally and on behalf of the above named agency, the continuing responsibility for carrying out the terms and conditions of the Test Use and Administration Manual. The principal signer for the recipient further agrees that all necessary administrative steps will be taken to assure that staff members, special consultants or others who may have access to the test materials supplied will be informed of this agreement and required to comply with it.

Test materials obtained from the supplier will be used only for the official purposes of the recipient in testing candidates for employment and promotion, test research, and development. Under no circumstances will supplied test material be made available to prospective job seekers or other unauthorized persons for purposes of study, copying or publication. No official, staff member, test consultant or other agent of the recipient will loan, give, sell, or otherwise make available any of the supplier's testing material to any agency or person who is not specifically authorized by the supplier to have access to such material, nor will they knowingly permit others to do so.

All supplied testing material in the possession of the recipient will be handled and stored in a manner that will prevent unauthorized persons from having access to it.

The recipient agrees that the supplier shall not be held responsible for any liability incurred by the recipient in any action arising out of the use of test material provided under this agreement.

Signer

It is understood and agreed that the principal signer of this agreement for the recipient is the person who is officially responsible for requesting test material from the supplier; will be the one to whom such material is sent; and will be regarded by the supplier as having responsibility for carrying out the terms of this agreement. The principal signer for the recipient must have successfully completed the Selection Personnel Certification Program.

The recipient has the option of identifying designated alternate signers to the principal signer. Alternate signer(s) will share in the full responsibility with the principal signer. Any designated alternate signers to the principal signer of this agreement must have successfully completed the Selection Personnel Certification Program.

The principal signer for the recipient may identify additional staff members who are authorized to pick up, deliver, and have custody of test materials provided by the supplier. The principal signer shall certify that each additional staff member agrees to the terms and conditions outlined in the Test Use and Administration Manual.

When the principal signer in this agreement is no longer a member of the agency, the recipient shall immediately appoint a successor to the principal signer who meets the principal signer requirements and is acceptable to the supplier. A new Test Use and Security Agreement Signature Page shall be completed by the successor immediately and provided to the parties of this agreement.

When authorized designated alternate signers or authorized staff members designated by the principal signer are no longer members of the agency, a new Test Use and Security Agreement Signature Page and/or Authorized Staff Member Signature Page shall be completed immediately and provided to the parties to this agreement.

Termination of Agreement

In the event that officials of the recipient find that they are no longer able to guarantee fulfillment of this agreement, the principal signer or other official making such determination will notify the supplier to that effect in writing. The principal signer or other official will return any and all testing material obtained from the supplier or produced directly or indirectly from the supplier's materials. The supplier reserves the right to terminate this agreement, or to withhold access to its testing materials, if it has reason to believe that the terms of the agreement are not being fulfilled by the principal signer or other officials of the recipient.

Test Use and Security Agreement Principal Signer and Alternate Signer Signature Page

On behalf of the agency I represent, I accept and agree to comply with the terms and conditions of this agreement and the Test Use and Administration Manual. It is further understood and agreed that when the principal signer or authorized designated alternate signer(s) designated in this agreement are no longer members of the agency, a new Test Use and Security Agreement Signature Page shall be completed immediately and provided to the parties to this agreement.

PRINCIPAL SIGNER FOR RECIPIENT

The following named individual has been delegated as the principal signer of this agreement. The named individual is authorized to order, receive, and have custody of test materials from the supplier.

Name _____
Title _____
Agency _____
Address _____
Date _____
Phone _____ Signature _____

AUTHORIZED DESIGNATED ALTERNATE SIGNERS (Optional)

The following named individual has been designated as an alternate signer of this agreement. The named individual is authorized to order, receive, and have custody of test materials from the supplier.

(1) Name _____
Title _____
Signature _____ Date _____

(2) Name _____
Title _____
Signature _____ Date _____