Chapter 6 Standardized Testing: Introduction

Standardized testing (ST) is prevalent and pervasive in education, training, and most aspects of modern life. ST drives educational programming, teaching, and decision-making to an unprecedented extent. Due to its centrality, ST must be understood and its mysteries mastered.

A standardized test is one that is administered under "standard" conditions, often proctored, with scores interpreted in a "standard" or consistent manner. Many tests produced by governments, commercial test publishers, professional organizations, etc. are standardized. Test administration procedures, conditions, time limits, and scoring procedures are strictly observed. These tests may be either group or individually administered. Score interpretation may be either norm- or criterion-referenced.

Achievement tests measure an examinee's relevant current level of knowledge and/or skill. An Aptitude (or ability) test measures estimated potential. Interest inventories (e.g., strong Interest Inventory) broadly measure an examinee's vocational or academic interests. Attitude scales measure an examinee's disposition concerning an issue or object of interest, e.g., social beliefs, love, marriage, values, etc. Personality scales or indexes measure personality type, behavioral tendencies, or mental health, e.g., Myers-Briggs Personality Type Indicator. Accordingly, we will first examine ST applications; next, we will study achievement and aptitude ST. Finally, we will explore ST score interpretation.

I. Standardized Testing Applications: Individual or Programmatic

A. Individual Applications

1. Screening typically involves a group administered test, whose purpose it is to identify weak or superior examinee performance on relevant knowledge, skills, or attitudes. Examinees are typically sorted into groups either for more testing, training, counseling, or any combination of the three. Diagnostic testing is to identify specific educational weaknesses (e.g., reading enabling skills) or a counseling or psychiatric treatment need.

2. A standardized test may be to predict academic success or achievement. Examples are: the ACT, SAT, GRE, MAT, GMAT, MCAT, or LSAT.

B. Programmatic Applications

1. Using achievement tests to evaluate instructional programs is complicated.
   (a) If a standardized achievement test is used for this purpose, it is important that the test publisher provide
      (1) A full content description (so an alignment study can be done);
      (2) An administration manual;
      (3) Suitable normative data, if using a NRT interpretation;
      (4) Suitable criterion referenced data, if using a CRT interpretation; and
      (5) A manual which reports documented validity and reliability indices.

(b) We should also be mindful that test scores may change due to
  (1) Introduction of a new test or new test version;
  (2) Differences between time of instruction for the examinee group and the norm group (e.g., fall or spring norms);
  (3) Changes in student or trainee and/or faculty demographics; and
  (4) Changes in a school's or organization's service areas.

(c) Avoid using the school as the level of analysis due to (b) (3) & (4). If the intention is to assess students then it is they who should be assessed.

(d) If the purpose is to improve instruction, we will typically:
  (1) Employ a criterion-referenced interpretation,
  (2) Use subtest scores when making decisions about a group, and
  (3) Use total test score (to reduce the possibility of misclassification) to place a student for remedial instruction.

2. In a minimum competency testing, examinees are assessed over specific knowledge and skills.
   a. Specific cut scores are established to determine acceptable performance levels. The driver's license examination or high school graduation tests are examples.
   b. In public and private education, such tests determine which students receive remedial instruction, are promoted, or receive a degree. States and provinces have an absolute legal right to require and specify what must be learned and can test that learning. Schools are required to provide instruction over that content and skills. The Texas Assessment of Academic Skills, the Florida Comprehensive Achievement Test (FCAT), and professional competency tests are prime examples of this use of standardized tests.
   c. When setting a "cut score":
      (1) Anyone of several accepted methods can be used to set the passing score (Crocker & Algina, 1986).
      (2) Whatever the passing score, some professional judgment is going to be required.
      (3) When setting a cut or passing score, select a committee of experts and distinguished community members to offer advice on the decision.

II. Aptitude and Achievement Standardized Testing

A. Introduction

1. <u>Purpose of Achievement and Aptitude Testing</u>
   a. An achievement test measures an examinee's current knowledge or skill performance level with respect to a defined body of knowledge or skills.
      (1) When scores are compared to national or local norms, an examinee's relative position with respect to that norming group is revealed.

(2) Uses include:
(a) Setting minimum standards for promotion and graduation;
(b) Exempting students from instruction (e.g., course challenge examination—if the challenge examination is passed, then the examinee is exempt from the course);
(c) Diagnosing learning or other problems; and
(d) Evaluating effectiveness of an educational program or instruction
(3) When combined with aptitude tests, achievement tests can be used to determine eligibility for learning disability services.

b. Ability tests measure an examinee's capacity to learn new knowledge and/or skills. These tests help students or trainees form realistic academic and vocational interests; determine specialized instructional services eligibility; and are used to help make admission decisions, etc., based on test scores.

2. Skills Measured: Generally
a. Achievement tests measure recently acquired specific knowledge and skills taught in school, staff development classes or other venues of formal instruction. Achievement tests should not be used to estimate a student's future learning potential.
b. Aptitude tests measure skills and knowledge learned over a long time, e.g., several years.
(1) These skills are:
(a) Usually learned informally, such as vocabulary;
(b) Less affected by a particular classroom or school experience; and
(c) Affected by cultural bias, cultural factors, and physical impairments or disability.
(2) If measuring learned skills acquired over time, aptitude scores can change significantly due to measurement error or recent student learning.
(3) Since any educational or psychological test must measure observable behavior, what a student has previously learned is a good predictor of subsequent learning.
(4) Aptitude tests provide a long-range prediction of an examinee's learning potential in a variety of subjects.

3. Validity & Reliability
a. Content validity is relevant to both aptitude and achievement tests.
b. Both achievement and aptitude tests require internal consistency reliability, as do any related subtests.
(1) Aptitude tests require test-retest reliability due to their predictive purpose.
(2) Don't assume test-retest reliability, if an aptitude test has predictive validity. Each type of reliability must be independently established.

c. Since aptitude tests often rely on an underlying psychological construct, e.g., intelligence, ensure that an intelligence test's characteristics are consistent with an accepted theory of intelligence.

B. Achievement Tests: Types & Uses

1. Diagnostic Tests: These are designed to detect underlying problems that prevent examinees from learning, and are used primarily to detect which enabling knowledge and/or skills are lacking. In essence, these tests are a small series or battery of achievement tests. For example, reading, mathematics, and writing are skills which are based upon other discrete knowledge and skills. To write, one must know an alphabet, be able to form letters, etc.
   a. Diagnostic achievement tests must:
      (1) Identify each critical enabling skill and accurately assess its component elements; and
      (2) Provide subset scores which accurately reflect examinee performance on each critical enabling skill.
   b. Each subtest score must be highly reliable, but the correlation coefficient between subtest scores should be low. Low correlations between subtest scores indicate that each subtest measures a discrete or separate skill. If subtests are moderately or highly correlated, then the subtests measure a general ability (construct) rather than specific prerequisite skills; this is not desirable.
   c. Each critical skill is measured by a single subtest, which when combined form a reading readiness or diagnostic test. These tests are usually orally and/or individually administered. Each diagnostic subtest should have its own validity and reliability indices.
   d. For example, reading readiness subtests include auditory discrimination, visual discrimination, vocabulary, comprehension, and copying and drawing. Commercially available readiness tests include the Metropolitan Readiness Tests or Murphy-Durrell Reading Readiness Analysis.

2. An Achievement Battery is a series of academic or professional subtests each measuring distinct knowledge or skills. These are used to monitor performance of individual examinees, schools, organizations, etc. Interpretation is based on national norms or other measures of position. Scores are usually reported for each subtest.
   a. Test batteries tend to be both time and cost efficient and draw on a common norm group. Thus, if a common norm group is used for the achievement battery, an examinee's performance across subject areas can be measured and compared.
   b. Test battery subtests tend to be shorter than single-subject tests with reliability coefficients range between 0.75 and 0.90.
   c. Because achievement batteries don't identify examinee strengths and weaknesses very thoroughly, they should not be used for diagnostic purposes.

d. Many test publishers are now providing both norm- and criterion-referenced interpretation. Commercially available achievement (battery) tests include the:
   (1) California Achievement Test <http://www.setontesting.com/cat/>
   (2) Terra Nova 3 Complete Battery
   < http://www.ctb.com/>
   (3) Iowa Tests of Basic Skills
   <http://www.riverpub.com/products/itbs/index.html>
   (4) Metropolitan Achievement Test (MAT8) \
   < http://www.pearsonassessments.com/>
   (5) Stanford Achievement Tests (10th ed.)
   <http://www.pearsonassessments.com/>

e. Major Commercial Test Publishers include:
   (1) Educational Testing Service <http://www.ets.org/tests_products/>
   (2) ACT http://www.act.org/products/k-12-act-test/

f. Many state competency tests are designed for criterion-referenced interpretation. Examples include the:
   (1) Florida Comprehensive Assessment Test (FCAT)
   <http://fcat.fldoe.org/fcat/>
   (2) The State of Texas Assessments of Academic Readiness (STAAR)
   <http://www.tea.state.tx.us/student.assessment/>
   (3) California Assessment of Student Performance and Progress
   http://www.cde.ca.gov/ta/tg/ca/
   (3) Many US states are shifting to or adding end of course examinations (EOC examinations). To locate EOC examinations, Google the state department of education (or equivalent) and then search the website for the EOC examinations. EOC's in Florida are found at
   http://fcat.fldoe.org/eoc/

C. Aptitude Tests: Types and Uses

1. Intelligence Tests
   a. These tests vary widely as to the specific examinee skills and tasks which are measured. An individual's intelligence is determined by his or her innate or genetic intellect, the effects of environment (e.g., home, exposure to cultural and community events, etc.), and the effects of schooling (e.g., high performance expectations, rigorous curriculum, etc.)

   b. Intelligence tests are generally constructed as follows:
      (1) Verbal subtests assess an examinee's proficiency with a spoken and/or written language.
      (2) Quantitative subtests assess an examinee's proficiency in solving arithmetic or mathematical problems.
      (3) Nonverbal subtests assess an examinee's skills with patterns, images, etc.
   c. Scores are typically reported as deviation IQ scores, stanines, or percentile ranks, which cannot be substituted for one another.

d. Since there are different conceptions of IQ, a test publisher should state explicitly and fully what is or are the test's underlying theory of intelligence constructs. An IQ test's construct validity must be firmly established.
   (1) IQ scores are based on learned behaviors and/or skills and are not static.
   (2) IQ scores are influenced by:
      (a) Innate ability and the effects of schooling and environment,
      (b) Opportunity to learn the skills measured by the specific test, and
      (c) Other prior learning.
   (3) Intelligence tests are usually individually administered and provide an overall measure of aptitude.

e. Examples of individually administered intelligence tests are the
   (1) Stanford-Binet IV
   (2) Wechsler Adult Intelligence Scale (WAIS-R), 16+ years
   (3) Wechsler Intelligence Scale for Children (WISC-R), ages 6-16
   (4) Wechsler Preschool and Primary Scale of Intelligence (WPPSI), ages 4-6

2. Scholastic Aptitude Tests
   a. These types of standardized tests are often confused with IQ tests; both are used to predict examinee ability or aptitude. IQ tests are based on a specific theoretical construct; whereas, most scholastic aptitude tests are based upon a specified knowledge or skill domain.
   b. Typically, scholastic aptitude tests are employed to predict success in training or school. Most scholastic aptitude tests are
      (1) Developed in a series;
      (2) Composed of separate versions for differing age, developmental, cultural groups; and
      (3) Offered in alternate forms (e.g., L, M, etc.).
   c. Predictive validity is required and must be established. Predictive validity is established by correlating test scores with an external criterion such as a student's grade point average (GPA).
   d. Validity coefficients range between .4 to .6 and rise when combined with other measures, e.g., similar scholastic aptitude tests.
   e. Examples of group administered scholastic aptitude tests are the:
      (1) Otis-Lennon School Ability Test (8th ed.) <http://www.pearsonassessments.com/>
      (2) Scholastic Aptitude Test <http://sat.collegeboard.org/home?affiliateId=nav&bannerId=h-satex>
      (3) The ACT < http://www.act.org/products/k-12-act-test/>

III. Standardized Test Score Interpretation

A. Introduction

1. The Two Primary Interpretations: Norm- and Criterion-Referenced
   a. The Norm-Referenced (NRT) Interpretation
      (1) The purpose of this approach is to determine an examinee's performance standing against (i.e., compared to) a norm group. To be a valid comparison, the norm group must be well-defined and representative of the population from which the examinee is drawn. For example, to interpret a test which measures critical thinking skills of college freshmen requires a norm or comparison group of college freshmen who are demographically and academically similar to the group being tested.
      (2) One commonly uses a percentile rank table or chart (Table 6.2) to make comparisons. Such comparisons are relative in their interpretation, i.e., a change in the norm group will most likely change the examinee's performance standing, e.g., a drop or increase from the 67^{th} percentile to or from the 78^{th} percentile. Percentile ranks are not interval level scores; so, they can't be added, subtracted, multiplied or divided. A percentile is a ranked ordinal category with a numerical name.
      (3) Norm referenced test scores are not very useful for curricular or instructional decisions as the NRT test content is often somewhat different than a school's or training program's curriculum and/or what was actually taught.
      (4) A norm-referenced (NRT) test score interpretation approach, uses the Z-score, T-score, Stanine, Percentile Rank, Normal Curve Equivalent, Grade Equivalent Scores, and/or scale scores.

   b. The Criterion-Referenced (CRT) Interpretation
      (1) An examinee's performance is compared to an identified and formally constructed content and/or skill domain (i.e., curriculum with specified learning objectives or performance standards).
      (2) Because criterion-referenced tests are based upon specified learning objectives or performance standards, content sampling (i.e., number and depth of test items) is deeper than the broad, shallow content sampling (a few, general test items) used in the NRT approach.
      (3) The meaning of the score is derived from the learning objectives and/or performance standards upon which the CRT test is based. It is assumed that a high score means that the examinee knows more of the material and/or can perform more of the skills expected than an examinee earning a lower score.
      (4) The CRT approach is most useful for instructional and/or curricular decision-making, as test items are selected due to their ability to describe the assessment domain of interest. It is common for examinees to score 80% or more of items correct. CRT tests are used

as achievement tests in classrooms and minimum competency or licensure examinations.

(5) Criterion-referenced test (CRT) score interpretation relies primarily on mastery classification (a cut or passing score), points earned/points possible, or percentage correct scores. Standard or scale scores might be used. With a CRT approach, we make an inference regarding how much about the content or skill domain (i.e., curriculum) the examinee knows or can perform.

2. Definitions**:** Derived, Standard, Linear, Normalized, and Scale Scores
   a. A derived score is any score, other than an examinee's raw score.
   b. A standard score is a derived score based on standard deviation units between a raw score and the mean of its score distribution. This process involves the transformation of a raw score to a measurement scale with a predetermined mean and standard deviation. Examples are z-scores, T-scores, and deviation IQ scores.
   c. Linear standard scores preserve the absolute position of a raw score relative to its mean when transformed to a "new" measurement scale with a predetermined mean and standard deviation. The shape of the original raw score distribution is exactly retained. Examples include z-scores and T-scores. Since the absolute position of each raw score is retained, mathematical operations (i.e., adding, dividing, etc.) can be performed on either linear or normalized standard scores.
   d. Normalized standard scores (e.g., stanines and NCE's) create a score distribution which approximates the standard normal curve, even if the raw score distribution didn't. Because normalized standard scores are based on percentile rank (PR) values that closely correspond to cumulative standard normal curve percentages, a standard score distribution is produced that approximates the normal curve and is said to be normally distributed. The computations are very complicated to perform by hand so they are computed using a computer software program.
   e. Scale scores are either linear or normalized standard scores (transformed to the standard normal curve). They are used to compare scores from different forms, editions, and/or levels of a test. There are several scale score systems. Computations are performed using computer software.
      (1) Scale scores are increasingly used in reporting results of national, state, or provincial standardized achievement tests.
      (2) Scale scores are used to describe organizational, school, state, or provincial level examinee performance. The statistical properties of these scores allow for longitudinal performance tracking, as well as direct comparisons between classes, schools, or organizations.
      (3) Another reason scale scores are now widely used is the need to create equi-difficult forms of the same test. It is virtually impossible to create absolutely equivalent forms of the same test; raw scores can be

statistically adjusted (i.e., transformed to a scale score) so that performance is represented as if the test forms were equidifficult.

(4) An example of a scaled score model is the Florida Comprehensive Achievement Test (FCAT). To view examples, visit http://fcat.fldoe.org/fcatUnderstandReports.asp

B. Selected Linear-Standard Scores, Based on a Mean and Standard Deviation

1. The z-score
   a. The z-score is the basis for all other linear standard scores. Half of z-scores are negative and half positive. They are usually carried out to one or two decimal places. The z-score has a mean of 0.00 and a standard deviation of 1.00.
   b. Formula 6.1 (Spatz, 2011, p. 72): z-score

$$Z = \frac{X - \bar{x}}{s}$$

X = a specified raw score
$\bar{x}$ or μ= group mean
s = group standard deviation

   c. Example: An examinee scored 59 on a test with μ = 51 and s = 6.
      (1) z = 59-51 = 8/6 = 1.33 or 1.33 standard deviations above the mean
      (2) The examinee scored at the 91st percentile or scored as well or better than 91% of the group, assuming that the test score distribution approximated the standard normal curve. To arrive at the 91% estimate, consult Appendix 6.1, where we see that a z-score of 1.33 yields a proportion of the area under the SNC of .40824 above the mean of z = 0.0. Next, we add .40824 + .50 (the proportion of the area under the SNC below the mean of z = 0.0) to yield .90824 x 100 or the 91st percentile.

2. T-score
   a. First a z-score is computed. Then a T-score with a mean of 50 and a standard deviation of 10 is applied. T-scores are whole numbers and are never negative.
   b. Formula 6.2: T = 10z + 50
   c. Example: An examinee scored 59 on a test with μ = 51 and s = 6.
      (1) z = 59-51 = 8/6 = 1.33 or 1.33 standard deviations above the mean
      (2) T = 10(1.33) + 50 = 63.3 or 63
      (3) The examinee scored as well or better than 90% of the group.
   d. Be careful and avoid confusing the linear standard T-score with the T-scaled score. While the two "T's "are identical in a standard normal distribution, they most likely are very different in a skewed score distribution.

3. Deviation IQ Scores
   a. This IQ score is used by the *Wechsler Preschool and Primary Scale of*

Intelligence-Revised (WPPSI-R), Wechsler Intelligence Scale for Children-III (WISC-III) and the Wechsler Adults Intelligence Scale-III (WAIS-III). Dr. Wechsler uses a mean of 100 and a standard deviation of 15. Some IQ tests (e.g., Stanford-Binet) have a standard deviation of 16.

b. Formula 6.3: IQ = 15z + 100

c. Remember that different intelligence tests measure different abilities, so be very familiar with what the IQ test is measuring for interpretation purposes.

C. Selected Standard Scores, Based on Rank within Groups

1. There are systems of standard scores based on rank. Lyman (1998, p. 101) writes, "they are based on the number of people with scores higher (or lower) than a specified score value." The percentile rank is the most common.

2. Percentile Rank (PR)
   a. Characteristics:
      (1) The PR ranges from 1 to 99 and is expressed only in whole numbers.
      (2) The numerical value of a percentile rank changes more rapidly near the center than the lower or upper ends of a distribution of raw scores.
      (3) Percentile ranks change inconsistently as more examinees tend to earn middle scores versus low or high scores. See Appendix 6.2.
   b. The advantages are that PR scores are very easy to explain and are easily understood. A percentile is a ranked ordinal category with a numerical name.
   c. One can't average percentile ranks as inter-percentile distances are not equal. Percentile ranks are not interval level scores; so, they can't be added, subtracted, multiplied or divided. On short tests, a raw score difference of 2 or 3 points can yield a 20 or 30 PR difference.
   d. It is easy to confuse a PR with percentage correct which is based on the percentage of test items answered correctly.
   e. Formula 6.4: Percentile Rank Estimation

$$\frac{S_B + .5\left(S_{AT}\right)}{N}$$

S_B = the number of students below a specified score

$.5\left(S_{AT}\right)$ = half of the students at the specified score

N = the number of students

   f. Example: You have been asked to estimate percentile ranks for employees from a customer service staff development program. There were 62 junior level employees who participated in the 8 hour training course. Scores are:

Table 6.1
Percentile Rank Data Table for Rank Estimation Example

Raw Score														
94	93	92	91	87	86	84	83	82	81	78	74	73	72	71
Examinees Scoring each Raw Score														
1	2	3	3	5	6	10	9	7	5	2	2	3	3	1

(1) Calculate an estimated percentile rank for a score of 87.

$$\frac{S_B + .5(S_{AT})}{N} = \frac{48 + 2.5}{62} = \frac{50.5}{62} = .8145$$

A raw score of 87 yields an estimated percentile rank of 82. Eighty-two percent of the examinees scored at or lower than 87.

(2) Calculate an estimated percentile rank for a score of 93.

$$\frac{S_B + .5(S_{AT})}{N} = \frac{59 + 1}{62} = \frac{60}{62} = .9677$$

A raw score of 93 yields an estimated percentile rank of 97. Ninety-seven percent of the examinees scored at or lower than 93.

3. Normalized standard scores (e.g., stanines and NCE's) create a score distribution which approximates the standard normal curve, even if the raw score distribution didn't. Norm groups are very large samples and that as sample size increases the distribution of scores increasingly approximates the standard normal curve.
   a. Stanine
      (1) A stanine is a normalized standard score commonly used in education.
      (2) Stanine range from one to nine with a $\bar{x}$ = 5 & s = 2; except for stanines 1 and 9, the others are exactly one-half standard deviation in width. Stanine values are presented in Table 6.2.
      (3) Its advantages are the same as any normalized standard score. Disadvantages are that the scores are coarse; and reflect status, not growth.

   b. Normal Curve Equivalent (NCE) s
      (1) NCE cores are an educationally related normalized standard score.
      (2) Characteristics are:
         (a) $\bar{X}$ = 50 and s = 21.06 with a range from 1 to 99;
         (c) NCE values & PR's are equal at 1, 50, and 99; and

(d) Equal changes in an examinee's raw score results in an equal change in his or her NCE score, but this is not the case for percentile ranks.

(3) Advantages are the same as any normalized standard score.

(4) The disadvantage is that NCE's were designed for research and evaluation purposes and are not recommend for use in test score interpretation.

Table 6.2
Stanines

SNC Area	Stanine	Cumulative SNC Area
Highest 4%	9	97-100%
Next 7%	8	90-96%
Next 12%	7	78-89%
Next 17%	6	61-77%
Next 20%	5	41-60%
Next 17%	4	24-40%
Next 12 %	3	12-23%
Next 7%	2	5-11%
Lowest 4%	1	0-4%

Note. SNC means area under the standard normal curve.
Stanines aren't scores; they are ranked ordinal categories.

D. Grade Equivalent Scores

1. Grade Equivalent (GE) scores are used almost exclusively in education.
   a. Popham (2000, p. 189) defines grade equivalent scores as "score-reporting estimates of how a student's performance relates to the [median] performance of students in a given grade and month of the school year."
      (1) The school year is 10 months (September = 0 and June = 10).
      (2) A student entering the 3rd grade would have a grade equivalent score of 3.0. Assuming performance is at the median level, he or she will have a grade equivalent score of 3.3 in December (Sept = 0, Oct = 1, Nov = 2, Dec = 3).
   b. Grade Equivalent Score Characteristics
      (1) Since test publishers don't have the resources to test the very large numbers of students at each grade level, most grade equivalent scores are set using interpolated values, i.e., most grade equivalent scores are not empirically set.
      (2) Even if overall student performance at a school improves, 50% of student's will be above and below the median level of performance.
      (3) Since student learning is not linear, the use of points to the right of the decimal, really don't equate student performance to a given month during the school year. A one or two raw point difference can mean several months' difference in a student's grade equivalent score.
      (4) GE scores are not useful for comparing an examinee's performance on two different tests.

(5) Generally, above-average students gain more than one grade equivalent each year, average students tend to gain one grade equivalent, and most below-average students gain less than one grade equivalent per year.

c. Advantages and Disadvantages
   (a) The advantage of GE scores is that one is able to tell if an examinee scores at, above, or below the median (midpoint) of other students' performance.
   (b) Disadvantages are that GE scores tend to be easily misinterpreted and should not be used for determining grade placement or service eligibility for learning disabled students.

E. Score Band Interpretation (Kubiszyn & Borich, 1996, pp. 320-328)

1. Brief Standard Error of Measurement Review
   a. The standard error of measurement is the standard deviation of error scores within a test. It is an estimated value.
   b. The mean of the error score distribution is zero.
   c. The distribution of the error scores approximates the SNC.
   d. Remember that even the best tests will still have some measurement error. A perfectly reliable test has $s_m = 0.00$; whereas a completely unreliable test has $s_m = s$. The standard error equals the standard deviation.
   e. The standard deviation summarizes the variability of raw scores, while the standard error of measurement is the variability of the error scores.

2. Interpreting Scores Bands
   a. Think back to when we constructed a score band for an individual test score. Now we will do so for several test scores over different content, but drawn from the **same** test battery with a common norm group.
   b. Sally has completed a year-end achievement test. Her raw scores on the subtests of the test battery score bands are found in Table 6.3.

Table 6.3
Score Band Interpretation

Subtests	Scores	68% Band	95% Band
Reading	106	103-109	100-112
Writing	109	106-112	103-115
Social Studies	93	90-96	87-99
Science	91	88-94	85-97
Math	100	97-103	94-106

$\bar{x} = 100$, $s = 10$, & $s_m = 3$

   c. Those score bands that overlap are most likely to represent differences which occurred by chance. So we say, in such circumstances, that levels of achievement are similar across the affected subjects. Where there is an obvious difference between score bands, we conclude that levels of achievement are dissimilar and not due to chance.

(1) Typically two bands are constructed: 68% and 95
(2) The 95% interval is more conservative than the 68% interval and is most often used when making decisions about individuals. Kubiszyn and Borich (1996, p. 338) offer the following advice: "let differences at the 68% level [be] a signal to you [teacher or administrator]; let differences at the 95 percent level be a signal to [the] school and parents."

3. Interpreting Sally's Score Bands
   a. 68% Score Band (We are 68% confident that Sally's true score on each subtest lies between the interval's lower limit and upper limit)
      (1) Levels of achievement are similar for reading, writing, and math.
      (2) Levels of achievement are similar for social studies and science, but different from reading, writing, and math.
      (3) Thus, Sally probably would benefit from remediation in both social studies and science provided by her teacher.
   b. 95% Score Band (We are 95% confident that Sally's true score on each subtest lies between the interval's lower limit and upper limit)
      (1) Given our interpretation guidelines we would conclude that there are similar levels of achievement across subjects, except for social studies and science.
      (2) We would most likely recommend that Sally receive additional instruction to improve her achievement in social studies and science.

Application Exercises

1. Bill had a score of 76 on the history test and 87 on the math test. Sally had a 75 on the history test and an 84 on the math test. History test $\bar{X}$ = 71 and δ = 3.6. Math test: $\bar{X}$ = 83 and δ = 3.1. For the history test n = 57 with 23 scoring lower than Bill. For the math test, n = 85 with 72 scoring lower than Bill. For history 45 scored lower than Sally and 53 scored lower in math than Sally. (Don't compare Sally to Bill.)
   (a) On which test did Bill perform better, using the z-score?
   (b) On which test did Sally perform better, using the z-score?
   (c) Explain how you know this.

2. (a) Convert Bill's two test scores to T-scores and (b) explain what each test score means in your own words, (c) Why would someone want to use T-scores? (Hint: 2 reasons.)

3. (a) Convert Sally's two test scores to stanines and (b) explain what each means in your own words.

4. A group of students has taken a math test. The measures of central tendency are: mode: 84; mean: 83; median: 84. The standard error of measurement is 3.0 points. Fully interpret this standard error to three standard error units. Mike earned 77 points.

Answers

1a) Bill performed better on the history test (z=1.39) than on the math test (z=1.29).
1b) Sally performed better on the history test (z=1.11) than on the math test (z= 0.32).
1c) Z-scores convert raw scores to a common metric or scale. Thus, it is possible to compare examinee performance across different content areas using tests which have differing means and standard deviations. The z-score has a mean of "0" and a standard deviation of "1". Higher positive z-scores suggest higher achievement levels.

2a) Bill's history T-score is 64 and his math T-score is 63.
2b) Since a T-score has a mean of 50 and a standard deviation of 10, a T-score of 64 suggests that Bill scored at about the 92nd percentile. A T-score of 64 equates to a z = 1.39. By consulting Appendix 6.1, we see that a positive z-score of 1.3 equals .41774. Since the mean of the z-distribution is zero half of the area under the curve is below zero. Thus, we add .50 + .41774 = .91774 or approximately the 92nd percentile. –A T-score of 63 equals a z-score of 1.29. Again consulting Appendix 6.1, we see that his score approximates the 90th percentile (.50 + .40147).

2c) The T-score is always expressed as a whole number and is never negative.
3a) Sally's approximate percentile rank for history is the 87th percentile (z = 1.11 or .36650 + .50) or 7'th stanine. Her approximate percentile rank for math is the 51st percentile (z= .32 or .01197 + .50) or 5'th.
3b) Sally scored above average on both tests.
4) We are 99.7% confident that the examinee's true score falls between 74 and 92 points.

References

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart, & Winston.

Kubiszyn, T. & Borich, G. (1996). *Educational testing and measurement*. New York, NY: Harper Collins College Publishers.

Lyman, H. B. (1998). *Test scores and what they mean* (6th ed.). Needham Heights, MA: Allyn & Bacon.

Popham, W. J. (2000). *Modern educational measurement* (3rd ed.). Needham Heights, MA: Allyn & Bacon.

Spatz, C. (2011). *Basic statistics: Tales of distributions* (10th ed.). Belmont, CA: Wadsworth/CENGAGE Learning.

Appendix 6.1 Area under the Normal Curve

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41308	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670
2.0	0.47725	0.47778	0.47831	0.47882	0.47932	0.47982	0.48030	0.48077	0.48124	0.48169
2.1	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461	0.48500	0.48537	0.48574
2.2	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.48840	0.48870	0.48899
2.3	0.48928	0.48956	0.48983	0.49010	0.49036	0.49061	0.49086	0.49111	0.49134	0.49158
2.4	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305	0.49324	0.49343	0.49361
2.5	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477	0.49492	0.49506	0.49520
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.9	0.49813	0.49819	0.49825	0.49831	0.49836	0.49841	0.49846	0.49851	0.49856	0.49861
3.0	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889	0.49893	0.49896	0.49900
3.1	0.49903	0.49906	0.49910	0.49913	0.49916	0.49918	0.49921	0.49924	0.49926	0.49929
3.2	0.49931	0.49934	0.49936	0.49938	0.49940	0.49942	0.49944	0.49946	0.49948	0.49950
3.3	0.49952	0.49953	0.49955	0.49957	0.49958	0.49960	0.49961	0.49962	0.49964	0.49965
3.4	0.49966	0.49968	0.49969	0.49970	0.49971	0.49972	0.49973	0.49974	0.49975	0.49976
3.5	0.49977	0.49978	0.49978	0.49979	0.49980	0.49981	0.49981	0.49982	0.49983	0.49983
3.6	0.49984	0.49985	0.49985	0.49986	0.49986	0.49987	0.49987	0.49988	0.49988	0.49989
3.7	0.49989	0.49990	0.49990	0.49990	0.49991	0.49991	0.49992	0.49992	0.49992	0.49992
3.8	0.49993	0.49993	0.49993	0.49994	0.49994	0.49994	0.49994	0.49995	0.49995	0.49995
3.9	0.49995	0.49995	0.49996	0.49996	0.49996	0.49996	0.49996	0.49996	0.49997	0.49997
4.0	0.49997	0.49997	0.49997	0.49997	0.49997	0.49997	0.49998	0.49998	0.49998	0.49998

Retrieved from: *NIST/SEMATECH e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/eda/section3/eda3671.htm.

Appendix 6.2 Percentile Ranks, z-scores, IQ scores & T-scores							
PR	z-score	IQ score	T-score	PR	z-score	IQ score	T-score
99	2.33	135	73	49	-.003	100	50
98	2.05	131	71	48	-0.05	99	49
97	1.88	128	69	47	-0.07	99	49
96	1.75	126	68	46	-0.10	98	49
95	1.64	125	66	45	-0.12	98	49
94	1.55	123	66	44	-0.15	98	48
93	1.48	122	65	43	-0.18	97	48
92	1.41	121	64	42	-0.20	97	48
91	1.34	120	63	41	-0.23	96	48
90	1.28	119	63	40	-0.25	96	47
89	1.22	118	62	39	-0.28	96	47
88	1.18	118	62	38	-0.31	95	47
87	1.13	117	61	37	-0.33	95	47
86	1.08	116	61	36	-0.36	95	46
85	1.04	116	60	35	-0.39	94	46
84	0.99	115	60	34	-0.41	94	46
83	0.95	114	60	33	-0.44	93	46
82	0.91	114	59	32	-0.47	93	45
81	0.88	113	59	31	-0.49	93	45
80	0.84	113	58	30	-0.52	92	45
79	0.80	112	58	29	-0.55	92	44
78	0.77	112	58	28	-0.58	91	44
77	0.74	111	57	27	-0.61	90	44
76	0.71	111	57	26	-0.64	90	44
75	0.67	110	57	25	-0.67	90	43
74	0.64	110	56	24	-0.71	89	43
73	0.61	110	56	23	-0.74	89	43
72	0.58	109	56	22	-0.77	88	42
71	0.55	108	56	21	-0.80	88	42
70	0.52	108	55	20	-0.88	87	42
69	0.49	107	55	19	-0.91	87	41
68	0.47	107	55	18	-0.94	86	41
67	0.44	107	54	17	-0.95	86	40
66	0.41	106	54	16	-0.99	85	40
65	0.39	106	54	15	-1.04	84	40
64	0.36	105	54	14	-1.08	84	39
63	0.33	105	53	13	-1.13	83	39
62	0.31	105	53	12	-1.18	82	38
61	0.28	104	53	11	-1.22	82	38
60	0.25	104	53	10	-1.28	81	37
59	0.23	104	52	9	-1.34	80	37
58	0.20	103	52	8	-1.41	79	36
57	0.18	103	52	7	-1.48	78	35
56	0.15	102	52	6	-1.55	77	34
55	0.12	102	51	5	-1.64	75	34
54	0.10	102	51	4	-1.75	74	32
53	0.07	101	51	3	-1.88	72	31
52	0.05	101	51	2	-2.05	69	29
51	0.03	100	50	1	-2.33	65	27
50	0.00	100	50	0	--------	--------	---------
Retrieved (and adapted) from http://www.psychology.edu/Links/prtable.htm. Differences between cell values in similar tables are due to rounding.							