Wisconsin Human Resources Handbook

Chapter 202

Statistical and Reliability Analysis

Sec. 202.010 Introduction

State civil service law requires that applicant assessment be job-related, objective, fair, reliable and valid. Reliability and validity ensure the fairness, utility, and relevance of assessments. Compliance with these legal requirements is determined, in part, through the use of statistics. Statistics provide valuable information on the quality of the assessment and are informative for decision-making relative to the selection process. Without this information, there is no basis for making crucial decisions that impact state hiring activities and the lives of job seekers.

The purpose of this chapter is to describe some of the statistical requirements and the minimum level of computation and interpretation needed. The Bureau of Merit Recruitment and Selection (BMRS) recognizes that not every Human Resource (HR) Specialist need be an expert in tests and measurements. In fact, many HR Specialists will not need the level of statistical sophistication reflected in this chapter to do their work. It is important, though, that this expertise is available either within an agency or within BMRS.

This chapter is not intended as a replacement for a basic course or text in elementary statistics; that resource is available at many technical college and university campuses throughout the state. However, a basic familiarity with elementary statistics is presumed. Human resources staff overseeing selection procedures are expected to have the capability of completing the necessary calculations and interpreting them. BMRS is available to provide training and consultation as needed.

Sec. 202.020 Statutory and Rule Authority

1. "All selection criteria, including minimum training and experience requirements, for positions in the classified service shall be job-related in compliance with appropriate validation standards and shall be subject to the approval of the director. All relevant experience, whether paid or unpaid, shall satisfy experience requirements." s. 230.16(4), Wis. Stats.

2. "The director shall establish criteria for evaluating applicant qualifications and shall require the same or equivalent competitive procedure for all applicants competing for eligibility on a register . . . ." s. ER-MRS 6.05(1), Wis. Adm. Code.

3. "The competitive procedures shall be: (a) Based on information from job analysis, position analysis or other equivalent information documenting actual job tasks to be performed or skills and knowledges required to perform job tasks, or both; (b) Developed in such a manner as to establish the relationship between skills and knowledges required for the successful performance in the competitive procedure and skills and knowledges required for successful performance on the job; (c) Supported by data documenting that the skills and knowledges required for successful job performance in the competitive procedure are related to skills and knowledges which differentiate among levels of job performance; (d) Sufficiently reliable to comply with appropriate standards for validation; and (e) Objectively rated or scored." s. ER-MRS 6.05(3), Wis. Adm. Code.

Sec. 202.030 Definitions

The following are definitions of terms used in this chapter. The mathematical formulas associated with some of these terms are found in Attachment #1. Visuals of statistical concepts are found in Attachment #2. A more complete explanation can be found in an elementary statistics text or BMRS sponsored training.

1. **Arithmetic average or mean**: The most common measure of central tendency, computed by totaling or summing scores and dividing by the number of scores; yields a measure of the difficulty level of the assessment and the overall quality of the candidate group.

2. **Coefficient alpha**: A general reliability coefficient based on the variance of scores on individual items or test parts or raters.

3. **Confidence interval:** A range of values that are believed to contain, with a certain probability, the true value in the population

4. **Correlation**: A measure of the direction and strength of relationship between two raters or variables, most frequently obtained by computing a Pearson Product-Moment correlation or Pearson r. Correlations can range from -1 (as one variable changes, the other changes in the opposite direction by the same amount), through 0 (change in one variable does not correspond with change in the other) to 1 (as one variable changes, the other changes in the same direction by the same amount).

5. **Reliability**: The extent to which the assessment device or instrument produces a consistent, trustworthy, dependable result. Reliability is necessary, but not sufficient, to produce validity and may be calculated in a variety of ways including rater reliability or agreement coefficients (e.g., resume screens) or coefficient Alpha (e.g., multiple-choice exams, OIQ assessments).

6. **Score range**: A rough measure of spread or dispersion in scores, generally obtained by subtracting the lowest observed score from the highest observed score. Score range represents the score distribution, which affects the central tendency and reliability statistics.

~~7.~~ **Skew:** A measure of the symmetry of the score distribution. Symmetrical or normal score distributions have a skew of approximately 0. When the majority of scores are low and the tail of the distribution points toward higher scores, the value of the skew is positive. Conversely, when the majority of scores are high and the tail points toward lower scores, the skew is negative.

8. **Standard deviation**: The most widely used measure of variability or spread in a set of scores; the average of how much a score differs from the mean.

9. **Standard error of measurement (SEM)**: An indicator of the amount of error in a particular individual's obtained score; the difference between an obtained score on an assessment and the "true score;" may be used to set the confidence interval to adjust the passing point. As rating consistency decreases, SEM increases.

10. **Validity**: The extent or degree to which evidence and theory support a process or decision. In the context of selection, validity refers to evidence that a competitive procedure accurately assesses what it was intended to. Different types of validity correspond with different sources for this evidence.

11. **Variance**: Another common measure of the spread or variability in scores, obtained by squaring the standard deviation.

Formulas for calculating these and other statistics are located in Attachment #1. As noted later on, software is readily available to make these calculations. The key is knowing which formula is appropriate and how to interpret the result.

Sec. 202.040 Job Analysis

1. Job analysis is not a statistic. However, it influences the statistics and results of the selection process. Job analysis is the process of identifying the knowledge, skills, abilities, and experience necessary to perform a job. Job analysis is the foundation for determining what criteria to assess in the selection process. Subsequently, assessment results represent the extent to which applicants meet the selection criteria. Statistics summarize the results of the assessment across the applicant pool. For more information, refer to Chapter 176.060-Competitive Selection Procedures, Job Analysis of the *Wisconsin Human Resources Handbook.*

Sec. 202.050 Validity Evidence

1. Validity evidence is essential for an effective and legally defensible selection process. Validity evidence demonstrates that the competitive procedure is job-related and meets the intended purpose. Different types of validity correspond with different sources for this evidence. For instance, face validity refers to whether the assessment appears to be job-related. Applicants' perceptions of the assessment can provide information about face validity.

2. At a minimum, all competitive procedures need to demonstrate content validity. Content validity refers to the extent to which the criteria assessed represents the position. Evidence of content validity comes from the job analysis and is documented on the Selection Assessment Strategy (SAS) form (DOA form 15536). The SAS form provides documentation on the criteria assessed and how each criterion assessed corresponds with the position description. This form be found on the DPM website at the following link: https://dpm.wi.gov/Documents/DPM%20FORMS/DOA-15536%20Selection%20Assessment%20Strategy.docx. Content validity is outlined more completely in Attachment #3 and *Wisconsin Human Resources Handbook* Chapter 176.

3. Content validity evidence is not always sufficient, however. Complex competitive procedures (e.g., large applicant pool, continuous recruitments, major changes to a selection process) may warrant a validation study of the predictive validity of the competitive procedure. Predictive validity refers to the relationship between the competitive procedure and a job-related outcome (e.g., performance, absenteeism, turnover). This validation study provides statistical evidence on the job relevance and utility of the competitive procedure. Generally, this type of empirical study should be completed in consultation with or performed by BMRS. What is required of agency HR staff is a recognition that there are instances where the minimum validity evidence is not adequate.

Sec. 202.060 Routine Statistics

There are some routine statistics that are required for all assessments. These include determining the number of candidates, the passing point, the number of passing candidates, and frequency counts of gender, ethnicity, veterans status, and disability status. The following table provides information on which statistics are expected for each of the main assessment types in State civil service.

	Rated Assessments (e.g., Resume Screen, Training and Experience)	Rated Assessments Pass/Fail Scale	OIQ, Multiple-Choice Test
Pass Rate*	X	X	X
Rater Agreement	X	X	-
Mean and Standard Deviation*	X	-	X
Applicant Score*	X	-	X
Correlations Between Raters	X	-	-
Reliability	X	-	X
Standard Error of Measurement	X	-	X
Adverse Impact Analysis	X	X	X

*Calculate this statistic at both the rater or subsection level and overall level

Sec. 202.070 Score Distribution

1. Reviewing the distribution of raw scores is a valuable first step. This step is an opportunity to identify and correct data entry mistakes and infer what the statistics will be.

   Signs of data entry mistakes may include:
   - Scores that are out of range for the assessment method
   - Negative civil service scores or civil service scores greater than 100
   - Failing civil service scores for all applicants
   - Negative reliability or reliability estimates greater than 1.00
   - Missing scores or incomplete scores or sets of ratings

2. This initial review also provides information on what the statistics will be. For example, if the majority of scores are at either the low end or the high end of the distribution, this will lower the variance and standard deviation. When reviewing raw scores on resume screen assessments, check for agreement between raters to identify indicators of inconsistent application of benchmarks, such as a lack of category agreement or scores being apart more than two points.

3. When there is a small number of applicants (less than 30 individuals), statistical analysis should be interpreted cautiously because one or two scores could substantially affect the statistics, and thus yield a distorted view of the assessment results. One possible solution is to combine small samples over time for a larger sample and more stable results.

4. While there are few universal hard-and-fast rules suitable for all situations, careful review and professional judgement is warranted when the circumstances above occur. Contact BMRS for advice and consultation as needed.

Sec. 202.080 Means and Standard Deviations

Means or average scores represent a typical score on the assessment and summarize the applicant pool's performance. When there are a similar number of high and low scores, the mean is typically at the middle of the minimum and maximum scores. If the average score is relatively low, there may be a lack of minimally qualified applicants and more recruiting may be in order before continuing, or the standards may have been set unrealistically high and the assessment and/or benchmarks may need to be modified.

Standard deviations represent the variability in competitive procedure scores. A low standard deviation indicates that applicant scores are close to the mean, whereas a high standard deviation indicates more variability and distance from the mean. Generally, a standard deviation of approximately 1.5 would be expected for nine-point resume screen assessments and approximately 0.5 for three-point resume screen assessments.

1. If the assessment scoring is number based (e.g., a resume screen assessment using a three or nine-point scale), computation of means and standard deviations is necessary.

2. If the assessment uses a pass/fail scoring method, then means and standard deviations are not applicable. Calculating the pass rate (number of applicants passed divided by the total number of applicants) is more appropriate for summarizing the results of the assessment than the mean and standard deviation.

3. In the case of Objective Inventory Questionnaires (OIQs), means and standard deviations should be calculated by OIQ subsection.

4. Multiple-choice assessments also require means and standard deviations to represent the typical score. If the assessment has subsections, it may also be useful to calculate the mean and standard deviation for each subsection.

Sec. 202.090 Reliability and Standard Error of Measurement

Reliability and standard error of measurement are key statistics for demonstrating validity evidence. Reliability represents the consistency or stability of the assessment. Without reliability, there can be no validity because there is a lack of consistency in what is being measured and thus a lack of evidence that the competitive procedure assesses candidate qualifications in a systematic way.

The appropriate method for computing reliability depends on the type of assessment.

1. Resume screen assessments (three or nine-point scale) require computing the Coefficient Alpha. Pearson Product-Moment correlation (r) between individual rater pairs is also useful, especially where Alpha reliability is marginal or worse. An intraclass correlation (ICC) can be useful for estimating interrater reliability when three or more raters are used. A Coefficient Alpha of 0.90 and correlations of 0.85 or higher may indicate consistent application of the benchmarks. If the coefficient alpha and correlation values are lower than this, review of raw scores for consistent application of benchmarks may be warranted.

2. For OIQs, computing reliability is more complex. While a Coefficient Alpha may be calculated for the overall instrument based on part or subpart scores, this method is crude and provides an underestimate of true reliability for the OIQ. A better procedure is to calculate a split-half correlation coefficient (such as a correlation between scores on even numbered items vs. scores on odd numbered items across the candidates) being careful to make sure that the two halves or parts are carefully matched in terms of content (for instance, matched in terms of experience and education content as well as number of items included in each half). See Attachment #4 for an example of the split-half procedure. Another option is to calculate a Coefficient Alpha for each individual OIQ subsection.

3. For multiple-choice tests, standard item analysis typically includes a Cronbach's Alpha coefficient. A Cronbach's Alpha of approximately 0.70 may indicate sufficient reliability. The intent of reliability in multiple-choice exams is to have consistency in what is being measured, while also allowing for some variability and complexity. If the reliability coefficient is less than 0.60 for either an OIQ or multiple-choice test, further review and revisions to the content before future administrations is recommended.

4. The Standard Error of Measurement (SEM) is easily calculated once the reliability estimate has been determined. It is calculated with the standard deviation and reliability coefficient. It is used to establish a confidence interval, or band of values, within which we expect the true passing point to be. The 68% confidence interval of the passing point is calculated by subtracting one SEM from the passing point.

5. The SEM is useful for adjusting passing points. Adjusting the passing point may be appropriate when there is a high SEM (a lack of consistency). If there is a lack of rater agreement in a resume screen assessment and reconvening the panel is not feasible, adjusting the passing point by one SEM may be appropriate. Certification bands can also be established using other statistical techniques (quartiles, deciles, fixed number of points, and so forth). The statutory requirement in determining the number of names to be certified is to ". . . use statistical methods and personnel management principles that are designed to maximize the number of certified names that are appropriate for filling the specific position vacancy." s. 230.25(1), Wis. Stats.

6. BMRS is available to provide advice and training on the various ways to compute reliability and adjust the passing point.

Sec. 202.100 Passing Point Determination

1. All assessments are required to have a passing point or a reasonable minimum standard. The passing point is dependent upon the nature of the assessment. Passing points must be set at a level that is reasonable, rational and consistent with normal expectations of acceptable job proficiency. Passing points frequently involve the judgements of job experts whose qualifications to make the required judgments must be beyond question and well documented. Several factors are generally considered in setting or adjusting a passing point on a case-by-case basis.

    a. Consequence of hiring mistakes and wrong decisions
    b. Assessment reliability and SEM
    d. Affirmative action and the 4/5 rules for adverse impact, or other means to determine adverse impact
    e. Candidate availability
    f. Number of vacancies

2. Persons or agencies generally unfamiliar with setting passing points should seek assistance before proceeding. The importance of setting passing points carefully and judiciously cannot be overemphasized. Passing points involve legal risk. If inappropriately set, they can result in unqualified persons continuing in the hiring process or qualified persons being rejected. BMRS is available to provide consultation and expert assistance prior to finalizing the results of an assessment wherever this assistance is needed.

Sec. 202.110 Rater Agreement Considerations

1. To the extent that a person may be denied a future opportunity due to a single judgment by one rater or denied an opportunity by virtue of needing one or two more points on a numeric scale of one kind or another is a critically important consideration.

2. Consistent application of the benchmarks and rating scale throughout the assessment is essential. Thorough rater calibration helps ensure consistent application of the benchmarks. The requirements for rater agreement depend on the type of scale and number of raters used. Using two raters for resume screen assessments requires agreement at the passing point. When using three or more raters for resume screen assessments, majority rules. Therefore, the numeric passing point may need to be adjusted to correspond with the passing point where the majority agrees.

3. HR Staff overseeing selection procedures should review raw scores to ensure consistent and logical application of the benchmarks and rating scale. Indicators of rater disagreement include:
    - Raw scores being apart more than two points
    - Disagreement on the rating category of an applicant
    - Disagreement on whether an applicant should pass

    See Examples 1 and 2 in this section for more information on checking for rater agreement.

4. If raw scores indicate a lack of rater agreement, reconvening the rating panel to discuss rating differences and reaching consensus is strongly encouraged. Reconvening the panel is also an opportunity for the HR staff to seek feedback from the panel on the clarity and utility of the benchmarks for future use.

See Examples beginning on next page.

Example 1: Reviewing Scores on a Nine-Point Resume Screen Assessment with Three Raters

Scale
0 = no scoreable response
1-3 points = less than acceptable (LTA)
4-6 points = acceptable (A)
7-9 points = more than acceptable (MTA)
Passing Point: 4

App #	Rater 1	Rater 2	Rater 3	Total Score	Average Score	Raw Score Consistency	Category - Decision
1	9	9	9	27	9	Yes	MTA - Pass
2	9	8	9	26	8.67	Yes	MTA - Pass
3	8	7	7	22	7.33	Yes	MTA - Pass
4	7	4	6	17	5.67	No	? - Pass
5	7	3	6	16	5.33	No	? - Pass
6	4	4	3	11	3.67	Yes	? - Pass
7	4	3	2	9	3.00	Yes	? - Fail
8	2	1	1	4	1.33	Yes	LTA - Fail
9	1	0	0	1	0.33	Yes	LTA - Fail
10	0	0	0	0	0	Yes	LTA - Fail

Ratings are far apart (more than two points) on Applicants 4 and 5 which indicates inconsistent application of the benchmarks.

Category disagreement on Applicants 4, 5, 6, and 7.

Per WHRH 176.080, majority rules when using 3 or more raters. Therefore, Applicants 5 and 6 would pass and Applicant 7 would fail.

Checking for Rater Agreement

- Raw Score Consistency – Are any of the raw scores more than two points apart for an applicant?
  - 80% - Out of 10 applicants, scores were within two points for 8 of them.
  - Reconvening the panel to discuss the applicants where scores were more than two points apart is recommended.
- Category Agreement – Did the raters agree on the applicant's rating category? (see rating scale above)
  - 60% - Raters agreed on the applicant's rating category for 6 out of 10 applicants.
  - Indicates inconsistent application of the benchmarks.
  - Discuss these category differences when reconvening the panel.
- Passing Point Agreement – Did the raters agree on whether the applicant should pass?
  - 70% - Raters agreed on whether the applicant should pass for 7 out of 10 applicants.
  - Because the assessment uses three raters, majority rules on the passing point.
  - May need to adjust the numeric passing point from 4 to 3.60 to allow Applicant 6 to pass

Next steps

- Reconvene the panel to discuss Applicants 4 through 7
- Document any rating changes
- Change the numeric passing point if needed.

Example 2: Reviewing Scores on a Three-Point Resume Screen Assessment with Two Raters

Scale
0 = no scoreable response
1 point = less than acceptable (LTA)
2 points = acceptable (A)
3 points = more than acceptable (MTA)
Passing Point: 2

App #	Rater 1	Rater 2	Total Score	Average Score	Raw Score Consistency	Category-Decision
1	3	3	6	3	Yes	MTA - Pass
2	3	2	5	2.5	Yes	? - Pass
3	2	2	4	2	Yes	A - Pass
4	2	1	3	1.5	Yes	? - ?
5	1	1	2	1	Yes	LTA-Fail
6	0	0	0	0	Yes	LTA-Fail

Disagreement on Applicant 2's category.

Disagreement on whether Applicant 4 should pass.

Checking for Rater Agreement

- Raw Score Consistency – Are any of the raw scores two or more points apart for an applicant?
  - 100% - Scores were within two points for all 6 of applicants.
  - Although the raw score consistency is good, a three-point scale and two raters requires a closer look at category agreement and decision agreement.
- Category Agreement – Did the raters agree on the applicant's rating category? (see rating scale above)
  - 67% - Raters agreed on the applicant's rating category for 4 out of 6 applicants.
  - The disagreement on Applicant 2 indicates disagreement about the benchmarks. This disagreement affects the applicant's average score and could affect whether they are on the certification list.
  - The disagreement on Applicant 4 indicates disagreement about the passing point.
- Passing Point Agreement – Did the raters agree on whether the applicant should pass?
  - 83% - Raters agreed on whether the applicant should pass for 5 out of 6 applicants.
  - Because the assessment uses two raters, both raters need to agree on the passing point.

Next steps

- Reconvene the panel to reach agreement on Applicant 4 and possibly Applicant 2
- Document any rating changes

Sec. 202.120 Adverse Impact Analysis

1. The purpose of adverse impact (or disparate impact) analysis is to determine whether a competitive procedure disproportionately excludes applicants in a particular group based on race, ethnicity, sex, or another basis protected by equal employment opportunity legislation. Adverse impact analysis is a comparison of the selection rates of each group (e.g., ethnic minorities vs. whites, females vs. males).

2. The Uniform Guidelines on Employee Selection Procedures provide guidance on the four-fifths rule. Per the four-fifths rule, there is evidence of adverse impact when the selection rate of one group is less than 80% of the comparison group's selection rate. A brief example is in Attachment #4. Statistical significance tests (e.g., Fisher's Exact Test, Corrected Chi-Square, z-test) are also useful for adverse impact analysis for large samples. Statistical significance indicates that the difference between groups is likely not due to chance. If the number of people in each group is small, it could result in an inaccurate indication of adverse impact.

3. If adverse impact is indicated by either the 4/5 rule or statistical significance testing, a review of the passing point, selection criteria, and assessment method is necessary. Consider whether 1) the selection criteria are valid and 2) there is an alternative assessment method that has less adverse impact and would be equally effective for selection purposes.

Sec. 202.130 Item Analysis

1. Item analysis is required for multiple-choice tests and OIQs, where items are typically scored one or zero, correct or incorrect response, respectively. This analysis consists of a review of each multiple-choice item (or objective test item) with respect to item difficulty and discrimination. These two indices determine the quality of the test item and help with decisions regarding the removal or modification of the item.

2. Item difficulty refers to the percentage of applicants who selected the correct answer. Lower percentages indicate greater item difficulty. Item difficulty informs whether the item is too easy or too difficult.

3. Item discrimination refers to the relationship/correlation between answering the question correctly and overall exam score. Item discrimination helps identify which items are valuable for differentiating applicants.

4. Item analysis can also help with identifying distracters or answer choices that seem to be working poorly or not working as well as they might, hence, providing valuable information about how to improve items. If the item analysis indicates that a low percentage (less than 5%) of applicants are selecting a distracter answer choice, it could indicate that the distracter is working poorly, and that item difficulty could be improved with more plausible distracters.

   In general, the ideal items will be of intermediate difficulty (proportion answering correct .50) and have strongly positive correlations with subtest scores for the keyed answer (in the .20s or .30s and beyond, but less is also acceptable). Incorrect answer choices will generally correlate negatively with subtest scores, but not always.

5. Agencies using multiple-choice (objective) tests are expected to effectively obtain and interpret these statistics (and other relevant statistics) with BMRS training and consultation as part of the assessment development and refinement process.

Sec. 202.140 Multiple Assessments

Routine statistics (reliability, pass rates, adverse impact analysis) should be computed separately for each assessment. For example, a selection process that consists of a resume screen, an objective-inventory questionnaire, and essay questions should be treated as three separate assessments and statistics for each assessment should be computed and evaluated separately. This is done to determine the quality of each assessment without confounding the information by treating the entire selection process as one assessment.

Sec. 202.150 Civil Service Score Conversion

1. The conversion from raw score to civil service score is analogous to going from one scale (such as Centigrade) to another scale (such as Fahrenheit). Raw scores on most assessments are on a scale different from the civil service scale of 70 to 100. Civil service scores provide information on applicant performance across the various types of assessments.

2. Procedures and details for converting assessment raw scores to civil service scores are found in Chapter 204—Assessment Scoring and Register Establishment of the *Wisconsin Human Resources Handbook*. A brief example is also found in Attachment #5.

Sec. 202.160 Documentation

1. Once the statistical properties of the assessment procedure have been determined (reliability established, passing point set, adverse impact determination made, and so forth), what remains is to document these findings, judgments, and conclusions. The Assessment Score Analysis form (DOA-15514) is used for this purpose

2. The Human Resources Specialist responsible for the transaction must complete the required form. record the essential elements listed on the form and obtain a signed endorsement of another Human Resources Specialist or supervisor having responsibility in such matters.

3. Once completed, the form is retained with the examination materials associated with the recruitment. It becomes essential documentation of the statistical performance or quality of the assessment procedure and the reasons and rationale followed in establishing the passing point.

Sec. 202.170 Administrative Information

This chapter was published in June 2002. In March 2003, the electronic links were updated and an administrative section added.

In March 2005, the chapter was updated to change the *Staffing Plan Summary & Approval* form to the *Exam Score Analysis* form.

In August 2018, Chapter 202 underwent a review and update pursuant to changes introduced by 2015 Wisconsin Act 55 and by 2015 Wisconsin Act 150. In July 2015, the Office of State Employment Relations was eliminated and the functions were transferred into the newly created Department of Administration, Division of Personnel Management. This chapter was updated to reflect the changes in terminology that resulted from the organizational restructuring. This chapter was also updated to generally change terminology from examinations to competitive procedures, to delete obsolete information, and to provide new information regarding the use of resume screen assessments. All attachments were updated and Attachment #2, Visuals of Stat. Concepts was newly created.

Attachment #1

Formulas Commonly Used in Civil Service

n = number of scores/applicants/raters/items
Σ = "sum of"
Xbar = refers to an X with a line over it

Mean: (Xbar): *the average of a set of data*

X = score
n = total number of scores/applicants

$$\overline{X} = \frac{\sum X}{n}$$

Standard Deviation: (s_x): average of how much a score deviates (differs) from the mean

X = score
Xbar = mean
n = total number of scores/applicants

$$s_x = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}}$$

Variance: (s^2): range/distribution among the scores (standard deviation squared)

X = score
Xbar = mean
n = total number of scores/applicants

$$s^2 = (s_x)^2 = \frac{\sum \left(X - \overline{X}\right)^2}{n-1}$$

Standard Error of Measurement: (SEM): difference between an obtained score and a true score (allows us to talk about assessment scores with confidence intervals)

s_x = standard deviation of assessment
r = reliability of assessment

$$SEM = s_x\sqrt{1-r}$$

Confidence Intervals: (CI):
68% CI:(1.000)(SEM)
95% CI (one tailed): (1.645)(SEM)
99% CI (one tailed): (2.330)(SEM)
95% CI (two tailed): (1.960)(SEM)
99% CI (two tailed): (2.580)(SEM)

Coefficient Alpha: (α): provides an actual estimate of reliability (internal consistency). You should always calculate this statistic even if other reliability estimates are appropriate. (Nunnally, 1994).

n = number of raters or number of items
$s^2_{(items/parts)}$ = variance of items or parts
$s^2_{(total)}$ = variance of total assessment

$$\propto = \frac{n}{n-1}\left(1-\frac{\sum s^2_{items\ /\ parts}}{s^2_{total}}\right)$$

Pearson's Correlation Coefficient *(r)*: indicates the strength and direction of the relationship between two continuous variables. Calculated by the covariance divided by the product of the standard deviations of the variables

$$r_{xy} = \frac{Cov(x,y)}{s_x s_y}$$

Spearman-Brown Formula: used for split-half reliability

$$\boldsymbol{r_{12} = \frac{2r_{12}}{1 + r_{12}}}$$

$\boldsymbol{r_{12}}$**=** The correlation between scores on two halves of a measure (e.g., even/odd numbered items, experience/training OIQ items)

Civil Service Score: (CSS):

Civil Service Score or CSS = Reciprocal R (Applicant's Total Raw Score - Passing Point) + 70

Civil Service Scores range from 0 to 100, with 70 as the passing point:

70_____ (30 pt)_____100
Pass Max

But assessments have any number of points available, for example a resume screen with 3 raters using a nine-point scale has the following point spread (assume a rater score of 4 is passing):

12_____ (15 pt) _____27
Pass Max

- The ratio between the two scales in the present instance is 30:15 or 2; this is called the Reciprocal.
- Maximum possible score or 100% = 27 pts
- Unadjusted passing score or 70% = 12 pts

Civil service score = Reciprocal (Raw Score - Pass Point) + 70

where reciprocal = 30/15 = 2 in the present instance

Example: If raw score for applicant = 19, passing point = 12

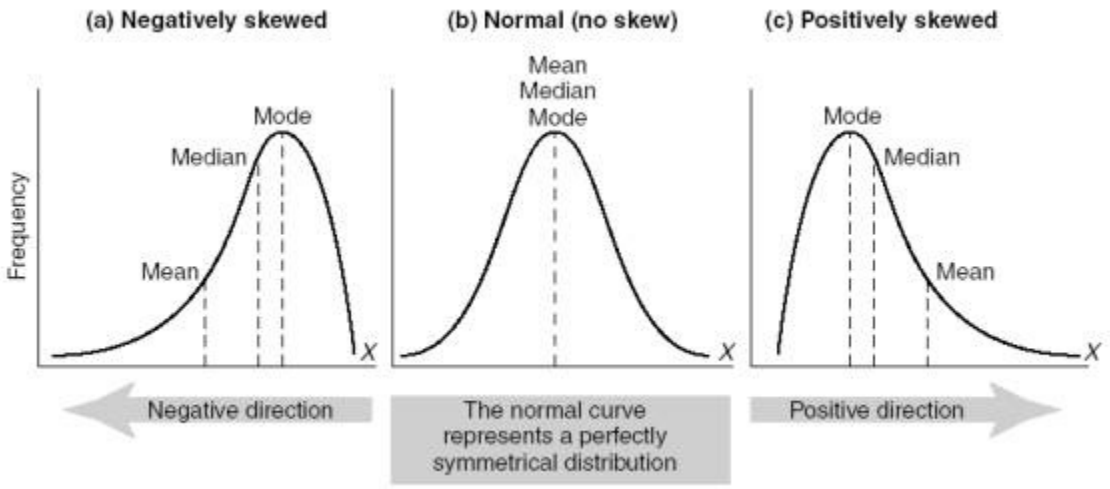
Civil service score = 2(19 - 12) + 70 = 84.00

Please see the *Wisconsin Human Resources Handbook* Chapter 204: Assessment Scoring and Register Establishment, for a more detailed explanation of the raw score to civil service score conversion process and examples.

Attachment #2

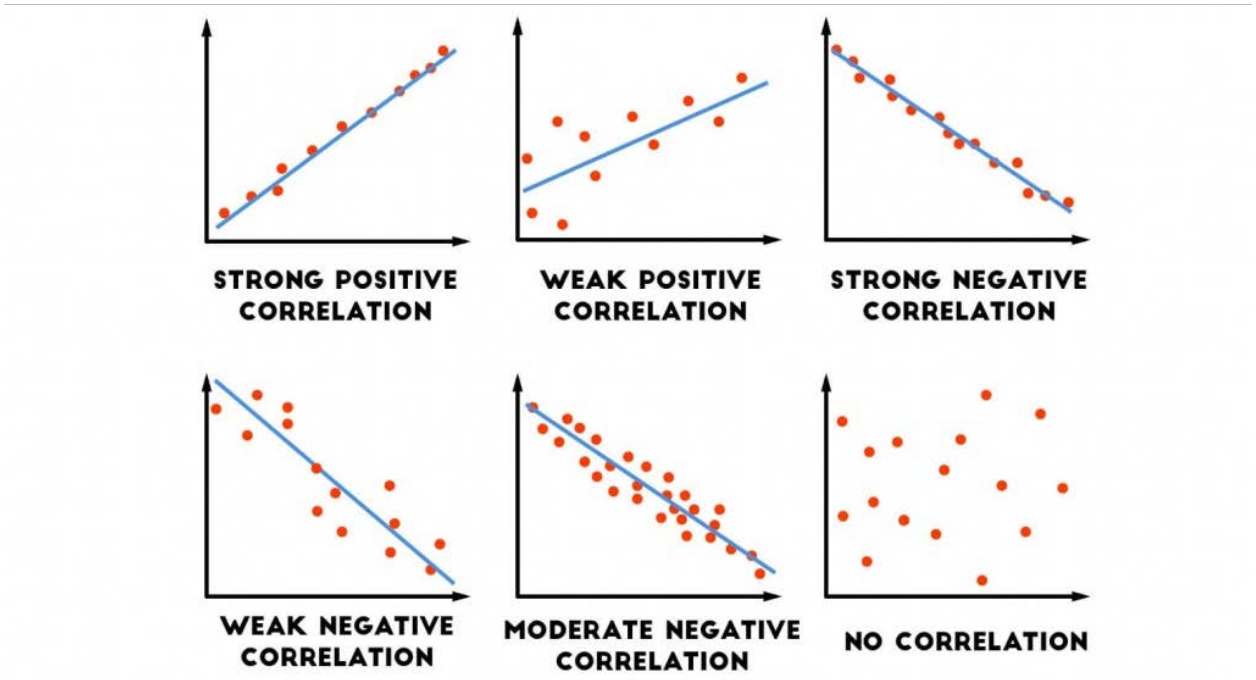
Visuals of Statistical Concepts

Central tendency and Skew



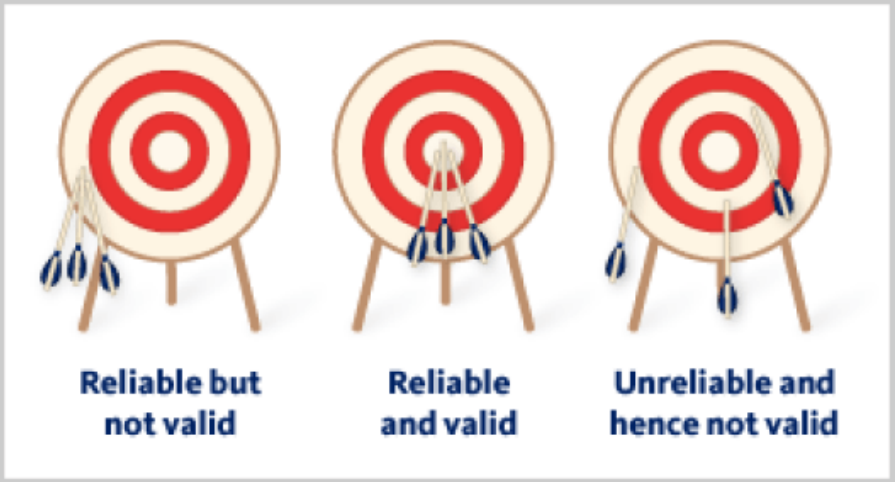
Source - http://kineticmaths.com/images/2/28/Skew.jpg

Correlations



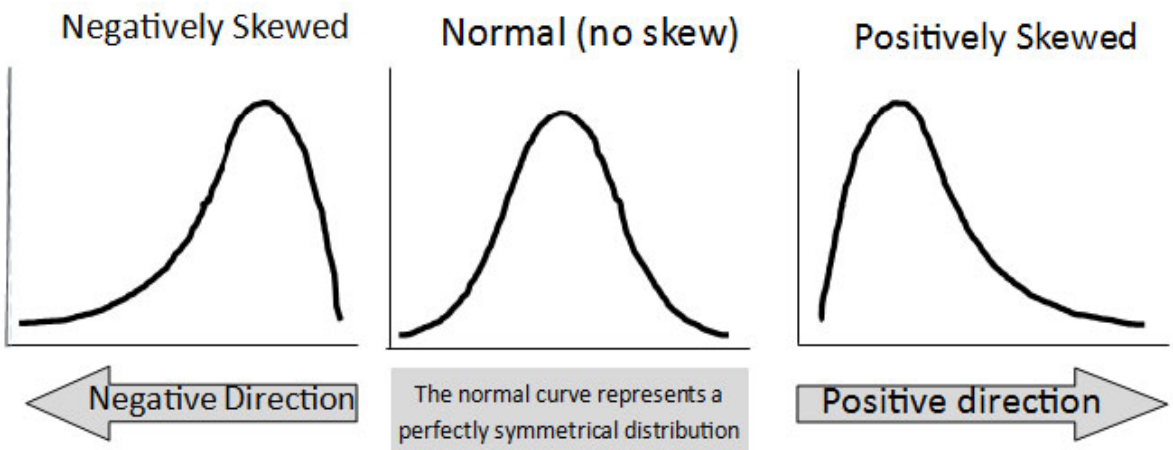
Source - http://cdn.pythagorasandthat.co.uk/wp-content/uploads/2014/07/correlation-1-1024x675.jpg

Reliability and Validity



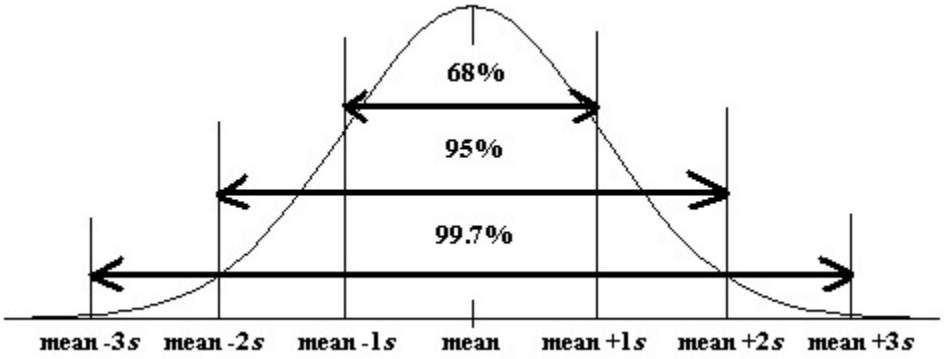
Source - http://strongbyscience.net/2017/03/23/reliable-versus-valid-beware-compare/

Skew



Source - https://www.managedfuturesinvesting.com/images/default-source/default-album/measure-of-skewness.jpg?sfvrsn=0

Standard deviation and confidence intervals



Source - https://newonlinecourses.science.psu.edu/stat200/sites/stat200/files/inline-images/emp_rule.jpg

Attachment #3

CONTENT VALIDITY

SELECT JOB EXPERTS

Experts must have direct, recent, first-hand knowledge of job
Generally supervisors and senior job incumbents

CONDUCT JOB ANALYSIS

Description of job tasks & activities
List of knowledge, skills, abilities required to perform job
Rate Position Description (minimum standard)

DEVELOP SELECTION ASSESSMENT STRATEGY

Use job analysis information to develop list of criteria/areas to be assessed
Select assessment methods
Identify criteria "trained for" upon entry and after hire to differentiate required and preferred qualifications

DEVELOP ASSESSMENT CONTENT

Develop assessment instructions, benchmarks, scoring guides
Conduct final review and make adjustments to content

ADMINISTER ASSESSMENT AND EVALUATE QUALITY OF RESULTS

Check for rater agreement, if applicable
Assess reliability and other measurement properties
Determine adverse impact

Attachment #4

Sample OIQ Analysis

Given: An Objective Inventory Questionnaire (OIQ) covering three areas (A, B, C) and six numbered task statements (or items) per area. Each task statement consists of an education score and an experience score. All items use a 1-4 scale. The results for Applicant #1 are reflected, below: a score is developed for odd numbered items (in bold) vs. a score for even numbered items. An odd item score vs. an even item score is subsequently developed for each applicant, then the two halves (even vs. odd item score) across applicants is analyzed using standard statistical methods (arithmetic mean, standard deviation, variance, split-half correlation), shown below.

continued

		Sample OIQ Reliability: Split-Half					
			(Items on 1-4 Scale)				
Applicant #1							
		Educ Score		Exper Score		Total	
Area A	**Item 1**	**4**		**2**		**6**	
	2	2		2		4	
	3	**3**		**1**		**4**	
	4	3		1		4	
	5	**2**		**2**		**4**	
	6	4		3		7	
Area B	**Item 7**	**2**		**1**		**3**	
	8	2		2		4	
	9	**1**		**1**		**2**	
	10	1		2		3	
	11	**2**		**2**		**4**	
	12	3		3		6	
Area C	**Item 13**	**4**		**4**		**8**	
	14	4		4		8	
	15	**3**		**3**		**6**	
	16	4		3		7	
	17	**3**		**3**		**6**	
	18	4		4		8	
Score Odd Items (1,3,5,7,etc) =				**43**			
Score Even Items (2,4,6,8,etc.) =				**51**			
		Appl	**Odd**	**Even**			
		1	**43**	**51**			
		2	40	45			
		3	120	110			
		4	85	95			
		5	74	64			
		6	133	121			
		7	115	120			
		8	94	87			
		9	56	62			
		10	72	84			
		Avg	83	84			
		Std. Dev	32	28			
		Var	1044	774			
		Split-Half r	0.96			Sample OIQ.xls	

Attachment #5

4/5th Rule

STEPS IN DETERMINING ADVERSE IMPACT
(4/5 OR 80% RULE)

1. Calculate the pass rate for each group (ethnic minority vs. white, male vs. female) by dividing the number of persons passing in that group by the number examined.

2. In each comparison, observe which group has the highest passing rate.

3. Observe whether the pass rate for the protected group is substantially less (i.e., less than 4/5 or 80%) than the pass rate for the unprotected group. If it is, adverse impact is indicated. Exercise caution in interpreting results based on small numbers, for instance, a few dozen cases where the shift in one or two people can change the conclusion.

Examples

Ethnic minority vs. white

Evaluated	Pass	Pass Rate
80 white	48	48/80 or 60%
40 minority	12	12/40 or 30%

4/5 of 60% is 48%. Since the pass rate for ethnic minority is less than 4/5 or 80% of the pass rate for whites, adverse impact against minorities is indicated.

Male vs. female

Evaluated	Pass	Pass Rate
55 male	28	28/55 or 51%
65 female	31	31/65 or 48%

4/5 of 51% is 40.8%. Since the pass rate for females is within 4/5 or 80% of the pass rate for males, adverse impact for females is not indicated.