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. Compliance with these legal requirements is determined, in part, through the use of statistics. The goal is to allow all persons to apply and then select the best from among them. The two most important qualities of a good measuring device are reliability and validity. Test characteristics, such as reliability and validity, exist whether or not they have been computed. Knowledge of the test characteristics adds value to the selection process. The computed statistics provide valuable information to the test giver and test developer and this information assists in making decisions relative to the selection process. Without this information, there is no basis for making crucial decisions that impact on 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. DMRS 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 DMRS. To this end, DMRS is prepared to provide the necessary consulting services and training for agency staff. We also believe that each agency with significant staffing delegation should have at least one individual on staff who provides the necessary interpretation.

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. A basic familiarity with elementary statistics is presumed. DMRS is available to provide training and consultation as needed.

Sec. 202.020 Statutory and Rule Authority

1. "All examinations, 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 administrator." s. 230.16(4), Wis. Stats.
2. "The administrator shall establish criteria for evaluating applicant qualifications and shall require the same or equivalent examination for all applicants competing for eligibility on a register . . . ." s. ER-MRS 6.05(1), Wis. Adm. Code.

3. "All examinations 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 on the test 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 on the test are related to skills and knowledges which differentiate among levels of job performance if the examination results are to be used as a basis for ranking candidates; (d) Sufficiently reliable to comply with appropriate standards for test 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. A more complete explanation can be found in an elementary statistics text or DMRS 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 test or assessment device 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. **Correlation**: A measure of the strength of relation between two raters or variables, most frequently obtained by computing a Pearson Product-Moment correlation or Pearson r.

4. **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 or coefficient Alpha.

5. **Score range**: A rough measure of spread or dispersion in scores, generally obtained by subtracting the lowest observed score from the highest observed score.

6. **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; it conveys important information about the quality of the test.

7. **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;" used to raise or lower passing points and set score bands.

8. **Subtest**: One of several test parts or components or assessment instruments used in combination.

9. **Validity**: The extent or degree to which a measurement instrument is accurate for the intended purposes, as distinguished from reliability which is a measure of consistency. The ultimate purpose of any assessment is validity or accuracy of measurement.

10. **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 usually thought of as a statistic. However, the State of Wisconsin currently has a system in place for conducting a minimal level of job analysis for staffing purposes. The high importance job content questionnaire (HIJC) provides a minimal level of job analysis information based on the ratings of job experts, along with the position description or identification of tasks and knowledge, skills and abilities (KSAs). At a minimum, this form of job analysis, position description with KSAs and HIJC, needs to be completed for every recruitment and staffing activity. The process is more thoroughly described in Chapter 168—High Importance Job Content Ratings (HIJC) of the *Wisconsin Human Resources Handbook*, found on the Office of State Employment Relations website at the following link: http://oser.state.wi.us/docview.asp?docid=1363.

2. While many circumstances warrant a more in-depth job analysis (multiple position classification, for example), the expertise for conducting a job analysis that might include development of a special questionnaire or job analysis instrumentation may not reside within an agency. In these cases, it is important that agency staff are aware that consultation with DMRS may be warranted and a more in-depth job analysis conducted.

Sec. 202.050 Validity Evidence

1. At a minimum, all test development activities need to be documented with the job analysis data as indicated above, and with a detailed examination plan. This, too, is not usually thought of as a statistic but the exam plan is an essential link in the content validation chain of evidence. The examination plan identifies and details the examination dimensions that are to be tested as well as the testing mechanism for each dimension. The checklist to facilitate the exam planning process can be found on the OSER website at the following link: http://oser.state.wi.us/docview.asp?docid=1197. This evidence is supporting documentation of the validity of the test. This type of evidence is generally referred to as content validity evidence. It should be noted that this documentation would not withstand scrutiny without additional statistical data, including the reliability evidence that is generally obtained through the analysis of the results. The standard content validity model is outlined more completely in Attachment #2.

2. Content validity evidence is not always sufficient, however. In the case of classifications that require considerable formal training on the job, other validity evidence, such as predictive or concurrent validity evidence, may need to be determined. Generally, this type of empirical study should be completed in consultation with, or performed by, DMRS. Again, 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 examinations. These include determining the number of candidates, and frequency counts of gender, ethnicity, veterans status, and disability status. Also, this may include part score and total raw score as well as civil service score and rank, where the scoring is numerical. A typical file structure that captures the essential information is illustrated below and is easily setup in WiscJobs or outside of WiscJobs (in MS Exel or Access or a standard commercial statistical package such as SAS, SPSS, SYSTAT, Minitab, etc.). Most software routines contain statistical formulas that eliminate manual calculations. For brevity purposes, an accommodation for part scores is not included in the illustration below but provision for part-scores would usually be incorporated into the design.

	TYPICAL FILE STRUCTURE											
								SCORES				
Appl ID	Name	SSN	Gender	Ethnic	VP	HDP	R1	R2	R3	RAW TOT	CSS	RANK
Struct.xls												

KEY:

Appl ID = Applicant identifier or number
Name = Applicant name
SSN = Applicant Social Security Number
Gender = Male (M) or Female (F)
Ethnic = Race/Ethnic Code
VP = Veterans Preference Points (10-15-20)
HDP = disability or handicap status
R1, R2, R3 = Rater 1, 2, 3 raw scores
Raw Tot = Total Raw Score or R1+R2+R3
CSS = Civil Service Score
Rank = numeric score rank

Sec. 202.070 Means and Standard Deviations

1. If the examination scoring is number based, computation of arithmetic means and standard deviations on scores obtained through the testing process is possible and required. As mentioned in section 202.060 of this handbook chapter, WiscJobs and readily available commercial statistics packages incorporate these formulas and greatly reduce the time and effort required to make the required calculations. A typical data set for an instrument based on rater judgments (i.e., Achievement History Questionnaire, Oral, Essay, Application Materials Review and similar devices) is illustrated on the next page. This scenario shows examinee raw scores summed across five 1-9 rating scales for each of three raters (R1, R2, R3) and 10 applicants.

2. If "qualitative ratings" (e.g., pass/fail, eligible/not eligible) are obtained, then means and standard deviations are less obvious. These can be computed by attaching a numerical value to the scores (that is, pass = 1 and fail = 0) and then computing the means and standard deviations. The purpose of doing this is primarily to facilitate determining the reliability and standard error of measurement for the test (discussed in section 202.080 of this handbook chapter).

3. In the special case of multiple-choice tests, this statistical information is routinely provided by any capable item analysis package available at most university campuses and large scale testing environments.

4. In the case of Objective Inventory Questionnaires (OIQs), arithmetic averages and standard deviations can be calculated by OIQ test part or subtest component.

5. Delegated agency staff are expected to have the capability of completing the necessary calculations and interpreting them as well as being able to use these statistics in other calculations such as the standard error of measurement. (See section 202.080 of this handbook chapter.) DMRS makes training and consultation available as needed and requested to the limit of available resources.

Situation: AHQ/Essay/Oral Bd Scores Obtained for 10 Examinees								
	Five 1-9 Scales where "4" = passing							
	APPL ID	**R1**	**R2**	**R3**	**Raw Tot**			
	1	24	26	30	80			
	2	29	26	33	88			
	3	24	25	31	80			
	4	30	25	29	84			
	5	30	27	32	89		Btwn Rater Correl	
	6	26	25	30	81		R1 v R2 =	0.7853
	7	21	22	28	71		R1 v R3 =	0.8326
	8	38	42	42	122		R2 v R3=	0.9466
	9	26	23	27	76			
	10	36	28	35	99			
Solve:								
	Avg =	28.40	26.90	31.70	87.00			
	Std Dev	5.38	5.59	4.32	14.51			
	Var =	28.93	31.21	18.68	210.44			
	Alpha Rel =		0.95					
	SEM =		3.24					
	1.96 SEM=		6.35					
	2.58 SEM=		8.36					
	CSS for a person with a raw score of 76 =					76.40		
		Assume:	100% = 135					
			70% = 60					
			Recip = 0.40					
File=descsol.xls								

KEY:

R1, R2, R3 = Total raw score for Rater 1, 2, 3
Raw Tot = Grand total raw score summed across raters
Appl ID = Applicant identifier or number
Avg = arithmetic average or mean
Std Dev = S or Standard deviation
Var = S-squared or variance
Alpha Rel = Coefficient Alpha Reliability
SEM = Standard Error of Measurement
CSS = Civil Service Score
Recip = Reciprocal

Betwn Rater Correl = Pearson r correlation between pairs of raters (R1 v R2, etc.)

Interpretive Aids and Standards

- Average scores-Examine arithmetic or average scores to ascertain the overall quality of the applicant pool; low scores may reflect a mediocre or even poor quality applicant pool or weaknesses in certain content areas (if the low scores are restricted to particular test parts, components, or questions). More recruiting may be in order before continuing. Consistently high or low scores may reflect a test that is either too hard or too easy, assuming the candidates are qualified. Combined with restricted variance, high or low scores may also be indicators of common rater errors such as strictness or leniency in rating environments, calling for better training or scoring guides.

- Score spread or variance-Without some variability, scores tend to pile up on one end of the score range or in the middle. As a result, the opportunity to distinguish between candidates is lost. Also, restricted variance reduces reliability. As a very rough rule of thumb, standard deviations should approximate 1/6 of the possible score range (low to high). In practice, real world assessment often falls short, sometimes far short, of this standard. Furthermore, in any testing environment, the test part (component, rater, etc.) that has the largest variance has the most impact on the final scores. For example, a part (rater, etc.) that has twice the variance of any other part (rater, etc.) carries twice the weight in determining final scores. If this isn't the intended result, then adjustments need to be made; for instance, include more items or questions in the more important content area(s), better benchmarks, better rater training, etc.

- Reliability estimates (r, Alpha, split-half, etc.)-These estimates should reach at least moderate levels, for example, +.70 or better, especially for total score. Extremely high reliability (high .90s) may not be useful either, since inordinately high estimates may indicate that only something very simple and not very complex is being measured or that raters may be biasing each others' ratings rather than exercising independent judgement. Generally, however, we applaud a higher reliability and trust the results. All things equal, raters whose judgements correlate most highly with one another carry more impact on the final result than raters who correlate less well.

Warning Flags or Indicators

Several indicators are cause for concern with ratings or other assessment devices. While there are few universal hard-and-fast rules suitable for all situations, careful review and professional judgement is warranted when the circumstances below occur. Contact your agency measurement professional or DMRS Human Resources Consultant for advice and consultation as needed. Warning indicators:

- Small sample size-many recruiting efforts currently produce small numbers of applicants. There is nothing inherently wrong or flawed with results based on small sample sizes. However, statistical analysis based on small sample sizes, for instance, less than 30 individuals, should be interpreted cautiously since trends may be affected heavily by one or two scores. For instance, if we have 30 people in a room and average their salaries where 29 earn less than $50,000 a year and the 30th earns $1,000,000 per year, the average salary will yield a distorted view of the earnings of the typical person in the room. Or if the shift in just one person from passing to failing reverses our conclusions with respect to adverse impact, the results should be cautiously interpreted. One possible solution for objective procedures, such as multiple-choice or objective inventories, is to combine small samples over time for a larger sample and more stable results.

- Markedly different average or mean scores for individual raters—on the order of 1.0-1.5 standard deviation difference between rater pairs.

- Markedly different variance estimates (S-squared) for individual raters—on the order of 1.5X to 2X difference between any rater pair. One solution to mean and variance differences is to standardize or transform individual rater raw scores to a common scale such as a T-score (mean of 50 and standard deviation of 10).

- Low reliability—for instance, below +.50 for individual rater pairs or Alpha (split-half, KR-20, etc.) reliability less than +.70 for overall or total score.

- Insufficient score variability—scores that all pile up in the same area of the score distribution, for instance, all high or all low or all near the middle.
- Test parts or components that are the most important exhibit less variance, perhaps considerably less, variance, than the parts or components of reduced importance.
- Any other result that appears suspect, for instance:
  - Negative civil service scores or civil service scores greater than 100
  - Failing civil service scores for everyone
  - Negative reliability or reliability estimates greater than 1.00
  - Missing scores or incomplete scores or sets of ratings

Sec. 202.080 Reliability and Standard Error of Measurement

Reliability and standard error of measurement are the key statistics for providing evidence that the examination is valid. Without reliability, a statistic that reflects the consistency or stability of the testing device, there can be no validity. Obviously, an unreliable, hence inaccurate, test leads to wrong decisions about the quality of the candidates and provides no evidence that the candidates can perform the job. There are various ways of computing reliability.

1. Computing reliability in rating environments (that is, Achievement History Questionnaire or AHQ, Application Material Review or AMR, Essay, Oral, or Work Sample) requires computing Coefficient Alpha for total score, at a minimum. Pearson Product-Moment correlation (r) between individual rater pairs is also useful, especially where Alpha reliability is marginal or worse.

2. In the special case of multiple-choice tests, standard item analysis routines capably generate a value for Kuder-Richardson estimates (KR-20 or KR-21) as well as Coefficient Alpha.

3. In the special case of OIQs, the problem of computing reliability is more complex. While a Coefficient Alpha may be calculated for the overall instrument based on part or subtest 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 #3 for an example of the split-half procedure. Another option is to calculate a Coefficient Alpha for each individual OIQ or test part using individual item variance, but this approach is impractical without scanned answer sheets and considerable computer support.

4. The Standard Error of Measurement (SEM) is easily calculated once the reliability estimate has been determined. It is used to establish a confidence interval, or band of values, within which we expect the persons true score or ability to fall. By convention, confidence intervals are most usually expressed as 1.96 or 2.58 SEMs, the 5% and 1% confidence intervals, respectively.

5. The SEM is useful in adjusting passing points. (See section 202.090 of this handbook chapter; see also the situational example in section 202.070, above.) For instance, it is not unusual to lower a raw score passing point by 1.96 SEM or 2.58 SEM, corresponding to the 5% and 1% confidence intervals, respectively. More rarely, passing points can also be raised using SEM. The SEM can also be used to develop score bands for use in certification. This is an example of making banding decisions based on the statistical attributes (i.e., reliability) of the examination. The use of SEM is one of the more sophisticated approaches to establishing passing point reductions and band-width certification; as such, use of SEM is widely accepted. 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. DMRS is available to provide advice and training on the various ways to compute reliability and establish score bands.

Sec. 202.090 Passing Point Determination

1. All tests are required to have a passing point or a reasonable minimum standard. The passing point is dependent upon the nature of the test and for some tests, such as multiple-choice, statistical calculations based on job expert opinions are required. 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. Test reliability and SEM
    c. Candidate test performance
    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 can be the subject of costly litigation, for one thing. For another, if inappropriately set, they can result in unqualified persons continuing in the hiring process or qualified persons being rejected. DMRS is available to provide consultation and expert assistance prior to finalizing the results of an exam wherever this assistance is needed.

Sec. 202.100 Rater Consensus in Pass/Fail Settings

1. Rater judgements that place individuals at or near the passing point warrant especially careful consideration, whether the exam is pass-fail or number based. See the example in section 202.100(2) of this handbook chapter. 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. Fine distinctions should generally be avoided in judgmental situations. Wherever possible, except in cases of the most critical jobs and job elements and large numbers of applicants, the benefit of the doubt should usually go to the applicant in terms of further consideration and advancement to the next step in the process.

2. As an illustration, in the data below we see that the raters are in universal agreement with respect to seven of the ten applicants: three are unanimously acceptable (#1, 2, 3), four are unanimously unacceptable (#7, 8, 9, 10). However, the discrepant decisions with regard to candidates # 4 , 5, and 6 are troublesome and need careful consideration in terms of such important factors as reliability of measurement, adverse impact, the need for qualified candidates, and the anticipated number of vacancies.

	Pass-Fail Rater Consensus				**Same Data – Numerical**		
App	R1	R2	R3		R1	R2	R3
1	P	P	P		1	1	1
2	P	P	P		1	1	1
3	P	P	P		1	1	1
4	**P**	**F**	**P**		**1**	**0**	**1**
5	**P**	**F**	**P**		**1**	**0**	**1**
6	**P**	**F**	**F**		**1**	**0**	**0**
7	F	F	F		0	0	0
8	F	F	F		0	0	0
9	F	F	F		0	0	0
10	F	F	F		0	0	0
	Pass=1			Avg=	0.60	0.30	0.50
	Fail = 0			Std Dev=	0.52	0.48	0.53
				Var =	0.27	0.23	0.28
				R1 v R2 =	0.53		
				R1 v R3 =	0.82		
	Consensus.xls			R2 v R3 =	0.65		

Sec. 202.110 Adverse Impact Analysis

1. Results of any examination need to be analyzed to determine if there is adverse impact. Adverse impact analysis is a review of the passing rates of the majority and minority groups, e.g., ethnic minorities vs. whites and females vs. males. Typically, a significance testing is not completed, although with large samples it is the preferred approach. The 4/5 rule is used in the majority of cases. This long-standing practice involves the comparison of 80% of the passing rate of the majority group (white) with the passing rate of the ethnic minority group; also the male vs. female comparison. If this comparison results in a greater passing rate for the majority group, then there is statistical evidence that the test may have adverse impact. Subsequently, a review of the passing point is necessary and a decision regarding the placement of the passing point must be made. The purpose of the review is to determine if the passing point is appropriately set for job-related reasons or if the passing point needs to be modified in order to address the adverse impact. DMRS expects all agencies to complete an adverse impact analysis using the 4/5 rule for small sample sizes. DMRS uses a special passing point sheet to facilitate this process; it is available to agencies with staffing delegation.

2. For a large sample, the 4/5ths rule can also be used, but a statistical significance test may be more appropriate. An example of an adverse impact analysis as well as a brief treatment of passing points and the process of converting raw scores to civil service scores can be found in Attachment #4.

Sec. 202.120 Item Analysis

1. Item analysis is required only for multiple-choice tests, 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 discrimination and difficulty level. These two indices determine the

quality of the test item and help with decisions regarding the removal or modification of the item. 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. 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.

2. Agencies with staffing delegation for multiple-choice (objective) tests are expected to effectively obtain and interpret these statistics (and other relevant statistics) as part of the qualitative review process. DMRS is available to provide training and consultation as needed.

Sec. 202.130 Subtest Properties

Many of the statistics cited above are also computed for each subtest, test part, or component. For example, if an assessment procedure contains an Application Materials Review (AMR) followed by an Oral Examination, then those statistics that are relevant for the test type should also be computed for each component. That is, routine statistics (reliability and standard error of measurement) should be computed for the AMR and Oral Examination, separately. If the assessment procedure consists of a multiple-choice test and consists of subtests, then statistics should be computed for each of these subtests. Formulas for obtaining overall reliability of the entire assessment procedure are available (Attachment #1) and should be used to combine the separate reliabilities. For example, assessment procedures that contain achievement history questions, objective-inventory questions, and essay questions should be treated as three separate subtests and statistics computed and evaluated for each component separately using the formula for three parts or subtests. This is done to determine the quality of each subtest without confounding the information by treating the entire selection procedure as one test.

Sec. 202.140 Civil Service Score Conversion

1. For numerically scored tests, the legislative intent is to add veterans preference points (currently 10, 15, or 20 points) to a 30-point civil service scale, where 70 = passing and 100 = maximum civil service score. 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 tests are on a scale different from the civil service scale of 70 to 100. For instance, a multiple-choice test may have 200 items with 80 items considered passing, or a range of 200 – 80 = 120 point range rather than the civil service scale of 100-70 = 30. Adding veterans preference points (10, 15, or 20 points) to a 120 point scale defeats the intent of the legislature; hence, the civil service score conversion is necessitated.

2. Procedures and details for converting test raw scores to civil service scores and adding veterans preference points are found in Chapter 204—Examination Scoring and Register Establishment of the *Wisconsin Human Resources Handbook* located on the OSER website at the following link: http://oser.state.wi.us/docview.asp?docid=1387. A brief example is also found in Attachment #4.

Sec. 202.150 Exam Score Analysis

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 Exam Score Analysis form (OSER-MRS-198 (Rev 2/05) is used for this purpose and can be found on the OSER website at http://oser.state.wi.us/docview.asp?docid=4606. The DMRS Human Resources Consultant responsible for the transaction must complete the required form, known informally as the "passing point sheet." Record the essential elements listed on the form, including passing point rationale. Then obtain a signed endorsement of another DMRS Human Resources Consultant or supervisor having responsibility in such matters. Once completed, the form is retained with the

examination materials associated with the staffing transaction. 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. 202.160 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.

Attachment #1

Formulas Commonly Used in Civil Service

n = number of scores/examinees/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/examinees

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

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

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

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

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

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

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

continued

Standard Error of the Mean: (SE): difference between the mean of a random sample of sample means and the population
s_x = standard deviation of test
n = number of scores/examinees

$$SE = \frac{s_x}{\sqrt{(n)}}$$

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

s_x = standard deviation of test
r = reliability of test

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

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 test

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

Kuder-Richardson 20: (KR_{20}): reliability estimate for dichotomous items (Eligible/Not Eligible)

n = number of items
s^2 = variance of test
pq = proportion correct x proportion incorrect
p = x/n q = 1-p

$$r_{20} = \frac{n}{n-1}\left(1 - \frac{\sum pq}{s^2}\right)$$

continued

Kuder- Richardson 21: (KR_{21}): nearly the same as the KR20, but this is a simpler formula. However, it is NOT as precise as the KR_{20} formula.

n = number of items
Xbar = mean of test
s^2 = variance of test

$$r_{21} = \frac{n}{n-1}\left(1 - \frac{\overline{X}\left(1 - \frac{\overline{X}}{n}\right)}{s^2}\right)$$

Total Test Reliability: (r_{tt}):

Two Parts or Two Tests:

s_1^2 = variance of part/test 1 r_1 = reliability of part/test 1
s_2^2 = variance of part/test 2 r_2 = reliability of part/test 2

$$r_{tt} = \frac{s_1^2(r_1) + s_2^2(r_2) + 2r_{1,2}s_1s_2}{s_1^2 + s_2^2 + 2r_{1,2}s_1s_2}$$

Three Parts or Three Tests:

s_1^2 = variance of part/test 1 r_1 = reliability of part/test 1
s_2^2 = variance of part/test 2 r_2 = reliability of part/test 2
s_3^2 = variance of part/test 3 r_3 = reliability of part/test 3

$$r_{tt} = \frac{s_1^2(r_1) + s_2^2(r_2) + s_3^2(r_3) + 2r_{1,2}s_1s_2 + 2r_{1,3}s_1s_3 + 2r_{2,3}s_ss_3}{s_1^2 + s_2^2 + 2r_{1,2}s_1s_2 + 2r_{1,3}s_1s_3 + 2r_{2,3}s_ss_3}$$

Spearman-Brown Prophecy: (r_{SB}): estimates expected increase in test reliability due to increase in length

$$r_{SB} = \frac{n(r)}{1+(n-1)r}$$

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) continued

z-score: a numerical transformation used to standardize data (scores)

$$z = \frac{X - \overline{X}}{s_x}$$

T-score: a numerical transformation used to convert scores into easily understandable numbers

$$T = 10(z) + 50$$

Civil Service Score: (CSS): Converts a raw score into the 100 point scale

R = reciprocal = 30/(maximum # of points – passing point)
pp = passing point

$$CSS = R\ (score - pp) + 70$$

*NOTE: "n" has been simplified for this formula sheet. In some instances the proper notation should be "N," depending on whether or not the information you are evaluating is sample driven or population driven. However, for simplicity's sake when dealing with a state civil service exam, "n" should suffice.

Attachment #2

	CONTENT VALIDATION	

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
Obtain HIJC ratings (minimum standard)

DEVELOP EXAM PLAN AND STRATEGY

Use job analysis information to develop list of dimensions/areas to be tested
Select exam approach(es) from available alternatives (written, oral, AHQ etc.)
Eliminate content "trained for" after hire

DEVELOP EXAMINATION CONTENT

Develop exam items, questions or other content to "get at" important areas
Develop benchmarks, scoring guides, answer keys
Conduct final review and make adjustments to content

ADMINISTER EXAM AND EVALUATE QUALITY OF RESULTS

Assess reliability and other measurement properties
Determine adverse impact
Set job related passing point

Attachment #3

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 #4

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

Examined	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 isless than 4/5 or 80% of the pass rate for whites, adverse impact against minorities is indicated.

Male vs. female

Examined	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.

80%rule –

continued

FACTORS AFFECTING PASSING POINTS

The need for judgment: must be rational and job related – See sec. 202.090 of this handbook chapter.

Test quality: Reliability and the Standard Error of Measurement
1.96 SEM
2.58 SEM

Candidate performance

Consequence of hiring mistakes and standards of job performance
Risk of false reject – for example, rejecting a person for a basic-skills job that most anyone could perform
Risk of false accept – for example, hiring a pilot who proves to be a performance failure

Affirmative action, adverse impact and the 4/5 Rule

Number of vacancies and candidates available

SUMMARY: CRITERIA FOR A GOOD PASSING POINT

- must be reasonable, rational and job related consistent with expectations of acceptable job proficiency
- based on judgment of job experts
- based on a variety of relevant factors and professional judgment; factors include:

  -hiring mistakes and consequences: risk of false accepts and false rejects

  -test reliability and SEM

  -candidate test performance

  -AA, adverse impact

  -candidate availability and number of vacancies

continued

BASIS OF CIVIL SERVICE SCORE CONVERSION

General formula for civil service score conversion:

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

****

Civil Service law requires 30 point scale, 70 to 100, for addition of vets points:

70__(30 pt)__100
Pass Max

But most tests have any number of points available, for example a 3 rater, 5 question oral with each question on a 1-9 scale has the following point spread (assume "4" is passing):

60_____(75 pt)_____135
Pass Max

Notice:
The ratio between the two scales in the present instance is 30:75 or 0.40; this is called the Reciprocal.
Maximum possible score or 100% = 135 pts
Unadjusted passing score or 70% = 60 pts

Civil service score transformation is a linear transformation and exactly like going from Centigrade to Fahrenheit temperature or vice-versa:

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

where reciprocal = 30/75 = 0.40 in the present instance

Example: If raw score for applicant = 90, passing point = 60

Civil service score = 0.40 (90 - 60) + 70 = 82.00

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