Applied Issues in Statistical Banding

Maury Buster, Ph.D.

Alabama State Personnel Department

Purpose

- Without taking a position on statistical banding, address a number of operational issues associated with it.
- Specifically, review such issues as:
  - Standard error of measurement (SEM) v. standard error of difference (SED)
  - Choice of reliability measure
  - Composite reliability
- Describe federal court case where rank-ordered scoring was successfully defended.

SEM v. SED

- A common question among those applying statistical bands is, “Do I use the SEM or the SED, and ‘Why’?”
- The answer is simpler than you might think, but is dependent upon what you want to know.

SEM v. SED

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

$$SED = \sqrt{2}\ SEM$$

Where:

σ_x = standard deviation of test scores

r = measure of reliability

NOTE: The final SED bandwidth is found by taking the product of the SED and a value from the normal distribution consistent with a predetermined level of confidence (e.g., one-tail $Z_{.05}$=1.65, two-tailed $Z_{.05}$=1.96).

SEM v. SED

An analyst can seek one of two things:

- Interval of likely/possible true scores around a given individual's score (SEM).
- Test of the significance between two individuals' scores (SED).

SEM v. SED

From an I/O textbook:

"Using the principle of the standard error of measurement, a method has been proposed for establishing bands of scores to replace individual scores. Using this approach, all candidate scores within a band are considered "equal" with respect to the attribute being measured if they fall within some specified number of SEMs of each other (usually 2 SEMs). It is assumed that any within band differences are really just differences due to the unreliability of the measure.

SEM v. SED

Examples:

If You Are Currently Using	Equivalent SED Level of Confidence 1-Tail	2-Tail
SEM	76%	52%
1.65 x SEM	88%	76%
1.96 x SEM	92%	83%

Choice of Reliability Measure

- There are several forms of reliability commonly referred to:
  - Internal consistency
  - Alternate-form
  - Test-retest
  - Inter-rater
- But, be careful, all r's are not created equal!

Choice of Reliability Measure

Standards 2.5

“ A reliability coefficient or a standard error of measurement based on one approach should not be interpreted as interchangeable with another derived by a different technique unless their implicit definitions of measurement error are equivalent.

Comment: Internal consistency, alternate-form, test-retest, and generalizability coefficients should not be considered equivalent, as each may incorporate a unique definition of measurement error.”

Choice of Reliability Measure

Recall:

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

- As r decreases, the bandwidth increases.
- Choice of r really does matter.

Choice of Reliability Measure

Actual occurrence:

For a given exam, three measures of reliability were calculated, two resulting in overall estimates.

Test-retest r = 0.64

Internal consistency (α) r = 0.84

Inter-rater was calculated by item

Once bands were calculated, 358 of the 698 (51%) examinees were in Band 1.

Choice of Reliability Measure

- Consider a scenario where, two-person panels provide initial ratings on a number of items. Afterwards, perfect consensus is required on the final ratings.
  - Thus, by definition, the operational inter-rater reliability is **1.00**.
- Had consensus been in place in the previous scenario, we would have been looking at a bandwidth of zero, and therefore no bands (versus the proposed band with 358 names).

Composite Reliability

- When creating composite scores from multiple components there are additional issues to consider.
  - How do we combine the r from multiple components to get one composite reliability estimate?
  - How do we estimate reliability for multiple-choice tests with multiple sections?

Composite Reliability

- In both instances it's not uncommon to see an agency use the a) internal consistency, b) alternate-form or, c) test-retest method to estimate r, based upon the composite/overall score.
- There is a formula for calculating the reliability of composite scores.

Composite Reliability

Composite reliability:

$$r_c = 1 - \left(\frac{\Sigma w_j^2 \sigma_j^2 - \Sigma w_j^2 \sigma_j^2 r_j}{\Sigma w_j^2 \sigma_j^2 + 2 \Sigma w_j \sigma_j w_k \sigma_k r_{jk}} \right)$$

Where:

w_j = weight of component j

σ_j = standard deviation of component j

r_j = reliability of a component j

r_{jk} = correlation between components j and k

Federal Court Case

§14C(9) *Uniform Guidelines* reads:

> "If a user can show, by a job analysis or otherwise, that a higher score on a content valid selection procedure is likely to result in better job performance, the results may be used to rank persons who score above minimum levels. Where a selection procedure supported solely or primarily by content validity is used to rank job candidates, the selection procedure should measure those aspects of performance which differentiate among levels of job performance."

Federal Court Case

1987 SIOP Principles for the Validation and Use of Personnel Selection Procedures reads:

"Interpretation of content-oriented selection procedures depends on the measurement properties of the given procedure. If a selection instrument measures a substantial and important part of the job reliably, and provides adequate discrimination in the score ranges involved, persons may be ranked on the basis of its results."

Federal Court Case

- The plaintiffs argued that the tests were not sufficiently valid to be used for the purpose of rank-ordering.
- The court found that the exams in question were highly content valid, so the defendants' remaining burden was to show “that a higher score on a content valid selection procedure is likely to result in better job performance.”

Federal Court Case

- The judge stated, “Whether there has been a sufficient demonstration that an exam may be used on a ranking basis is a matter that is within the bounds of acceptable professional practice.”
- The State collected ratings from SMEs regarding the exams’ ability to differentiate.

Federal Court Case

- The Court ruled:
  - ". . . the defendants have met their burden of showing that a candidate who has a higher score on these exams is likely to exhibit better job performance."

    Based upon:

    - "These exams are highly content valid--reflecting quite closely the content of the underlying jobs--and the SMEs have evaluated the exam exercises to ensure that they distinguish between different levels of job performance."
    - Testimony from three defendant I/O Psychologists. Specifically, testimony that there is an adequate variation in exam scores.

Federal Court Case

- The Court further ruled:
  - The plaintiffs ". . . have not undertaken to show that banded scoring is as content valid as ranking, or that it would have less adverse impact than ranking."