“Statistical Banding”: Statistically Inappropriate?

Dr. Philip Bobko
Professor of Management and Psychology Gettysburg College
President, Building Personnel Systems

Dr. Philip L. Roth
Professor of Management
Clemson University
President, Roth, Roth, & Associates

Overview

- Statistically-based banding is used extensively in public sector organizations to denote ranges of scores that are allegedly “equal” with respect to selection
- There are major criticisms, yet banding persists
- Regarding diversity, even banding theorists demonstrate it doesn’t accomplish this goal without subgroup preferences in bands
- We present other, new/unstated problems with banding
- Individuals within a band <u>are</u> indeed different
- Bands, as currently mis-constructed, are too wide and mis-label too many individuals as equal with respect to selection

Some formulas

- [1] SEM = $\sigma_X \sqrt{1-rel}$

- SD associated with the difference in two scores is

  [2] SED = SEM = $\sqrt{2}\left(\sigma_X \sqrt{1-rel}\right)$

- Thus, the bandwidth is typically

  [3] Bandwidth = 1.96 $\sqrt{2}\left(\sigma_X \sqrt{1-rel}\right)$

Current procedures

- If two scores differ by more than the bandwidth value, they are said to be “statistically reliably different”
- In Cascio et al. (1991), scores ranged from 22 to 96 and all scores between 96 and 84 in the first band
- Current banding formulas imply that IQs of 110 (75th percentile) and 123 (94th percentile) are essentially equivalent
- Or, any score between 720 and 790 on the SAT math test would be considered equivalent to a score of 800

Scores in a band ARE different

- Bands are computed around a single observation, but organizations are concerned about aggregate utility across hiring decisions
- The issue is: are top scoring individuals statistically better, on average, than others in a band? The answer can be "Yes."
- Consider Cascio et al. (1991). The top score was 96. We computed mean score for the n=534 in first band. Using a conservative assumption about standard deviations, the difference is "significant" ($t = -14.70$, $p<.05$)
- On average, scores in the first band are less than the highest scoring individuals (scores of 96)
- Further details available in Bobko & Roth, *International Journal of Selection and Assessment* (2004)

Bands are too wide

- To try to recover from a logical inconsistency, bands are constructed from the "highest available score" (see Cascio et al., 1995)

- However, standard errors of measurement (SEMs) depend upon the underlying level of the test taker – well-known in classical test theory and item response theory (and in APA Standards)

- Smaller SEMs are associated with high/top scorers; yet banding uses a SEM which better reflects test takers (and scores) at the middle of the distribution (often not who you want to select)

- This is intuitively true and obvious from the binomial; really good people will get most things correct and have relatively less implicit variation in their scores

Demonstrating bands are too wide

- Used Math Knowledge (MK) Form 9A test from the ASVAB
- Computed bands using the traditional (incorrect) unconditional SEM approach
- Computed bands using conditional SEMs (using the binomial error approach and an IRT approach)
- With current banding approach, 30% of the individuals would be labeled equivalent; using more correct approach, 14% would be labeled equivalent
- Thus, bands as currently mis-calculated are at least 50% too wide

More simulations

- We modeled what would happen if the test was more difficult; bands would be reduced in size by about 36%
- We modeled what would happen if the test was easier; bands would be reduced in size by about 63%
- More details are in Bobko, Roth, & Nicewander (*Organizational Research Methods*, 2005)

Summary

- On average, individuals in bands are statistically different (lower) than the top available score
- As currently calculated, bands are too wide – by a factor of 36% to 63%; too many individuals are mislabeled as “equal”
- Why use banding, given that banding does not increase diversity unless sub-group preferences are used within bands
- Utility loss can also be substantial (22% in Cascio et al.’s original data)
- We suggest a back-to-basics approach combined with top-down selection – defended in court (thorough job analysis, systematic test development procedures, involvement of multiple constituencies, other sound science practices)