IPAC 2018
Old Town Alexandria, VA
July 29, 2018 – August 1, 2018

Adverse Impact and Utility: A Comparison of Top-Down and Banding Selection Protocols

Frank Igou, Ph.D - Louisiana Tech University
Reagan Girardot, M.S. - Louisiana Tech University
Mallory Wright - Louisiana Tech University
Zollie Saxon - Louisiana Tech University

Introduction

Utility, specifically value-added return, is the usually primary consideration when using an assessment process to select or promote job candidates.

An assessment process should return value to an organization through identifying the job candidates who more likely to be the higher performers.

Less than Perfect Prediction

Assessment processes do not make perfect predictions on an individual candidate performance, but can make reasonably accurate predictions about a group of candidates selected over time.

This is conceptually similar to the actuarial prediction that insurance companies.

More Probabilistic than Deterministic

Although valid assessment processes improve the overall quality of the candidates selected, such processes are more probabilistic than deterministic.

For example, is it a given that the person who obtains a score of 92 going to be a higher performer than the person who obtains a score 90?

More Probabilistic than Deterministic

Because of the less than perfect individual predictions, many organizations use banding methods based on error of measurement.

These methods define a range of scores in which as scores are considered statistically equivalent.

Reported Demographic Differences

Previous research has reported lower mean scores and in some cases smaller variances for EEO protected groups

African-Americans and Hispanics compared to white examinees; women on standardized mathematics tests)

Some have reported these score differences as great as one standard deviation below the mean test score of referent groups.

(e.g., McKinney & Collins, 1991, Gottfredson, 1986).

Secondary Considerations

A secondary consideration may be workforce diversity.

Many organizations strive to create a diverse workforce which demographically represents individuals in the relevant labor market, constituents, stakeholders or customers.

Demographic Differences

Previous research has reported lower mean scores and in some cases smaller variances for EEO protected groups

African-Americans and Hispanics compared to white examinees; women on standardized mathematics tests)

Some have reported these score differences as great as one standard deviation below the mean test score of referent groups.

(e.g., McKinney & Collins, 1991, Gottfredson, 1986).

Demographic Differences

Some believe these reflect “real world” differences among groups in innate intelligence, developmental opportunities, etc.

Some believe the testing processes are biased. For example, different groups use language differently – verbal loading for exam

Errors in Prediction

"False misses or erroneous rejections due to error in prediction may reduce employment opportunities for minority group members and can perpetuate the effects of past discrimination on job candidates from lower scoring minority groups."

(Murphy, 1994; Hartigan & Wigdor 1989).

Score Use Methods Examined

- Strict top-down selection (TD)
- Top-down within groups (TDW)
- Fixed bands-random selection (FR)
- Fixed bands minority preference with top-down selection (FP)
- Fixed bands, minority preference with random selection (FR)
- Sliding bands, random selection (SR)
- Sliding bands, minority preference with top-down selection (SP)
- Sliding bands, minority preference with random selection (SPR)

The Standard Error of Difference

Cascio, Zeddeck, Outtz and Goldstein (1994) suggested the use of the standard error of difference (SED) as that the proper statistic for determining whether two scores are reliably different. Anecdotally speaking, it appears to be the most commonly accepted statistic for creating bands.

Banding Types

Traditional

Bands determined based on expert opinion, tradition, trend analysis, etc.

90 -100 = A
80 – 89 = B
70 – 79 = C
60 – 69 = D
Below 60 – Don't Ask

Rules of 3,5 and 10
All Qualified
"Naturally Occurring" Breaks in the Score Distribution

SED

(Standard Error of the Difference)

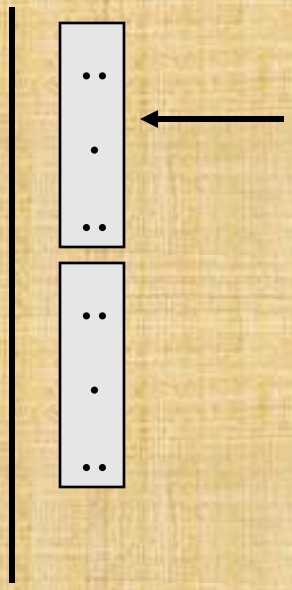
Using tests of statistical significance to determine test bands considered equal

Consideration of the SEM of the test (standard deviation, test reliability, and level of confidence desired)

SED Banding Types

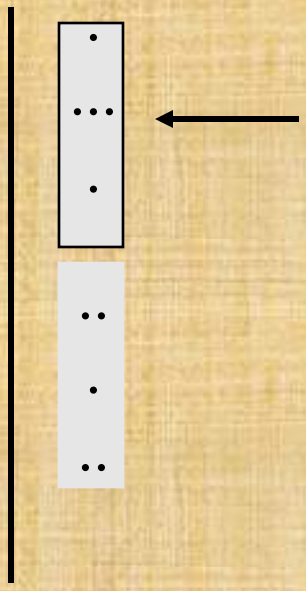
The top scores are used to anchor the bands

Fixed



All are selected from a band (the band is "exhausted) before moving to the next band.

Sliding



The top of a band slides down after the highest score is removed from the top; the band is then "re-anchored."

Calculating the Standard Error of Difference

In its simplest computational formula, the *SED* is simply the product of the *SEM* and the square root of 2 (approximately 1.414).

The width of these score band may be calculated as:

$$\text{Band Width} = C \times sd_x \times (1 - r_{xx})^{1/2} \times 1.414$$

Proponents assert that using bands of scores reduces adverse impact while preserving the validity of selection procedures.

Bandwidth

In the previous slide, r_{xx} refers to the reliability of the test, and the term "C" refers to the normal deviate that corresponds to the desired level of confidence. For example, a C value of 1.96 corresponds to a 95% confidence interval. Thus, if one wanted to establish bands that were 95% confidence intervals, one would set the bandwidth at approximately 2 *SEDs.*

Data Examined

Data for the examples presented today are from two entry-level police selection multiple-choice selection tests from for two medium sized cities (250 – 600k people).

The data presented here do not reflect the actual selection outcomes. The use of the tests were on-going at the time of data collection.

Data Set 1	Total	Pass	Mean	SD
All Examinees	688	435	69.81	9.85
Whites	346	256	71.77	10.12
African-Americans	258	164	67.93	9.23
All Other Groups	84	63	72.48	9.41
Data Set 2	**Total**	**Pass**	**Mean**	**SD**
All Examinees	532	436	72.12	11.37
Whites	245	191	71.06	10..20
African-Americans	223	181	72.19	11.48
All Other Groups	64	54	73.71	11.14

To simplify this presentation, only results comparing African-American and white examinees outcomes are presented.

Selection Ratios

For illustration purposes a 20% selection ratio was applied.

Realistically, neither agency would be likely to hire 20% of the examinees during a calendar year, however both agencies may need to refer 20% or more for subsequent steps in the selection process to make a sufficient number of hires.

Adverse Impact Examined Using the EEOC 4/5ths Rule

"The agencies have adopted a rule of thumb under which they will generally consider a selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5ths) or eighty percent (80%) of the selection rate for the group with the highest selection rate as a substantially different rate of selection." EEOC, 1979

The 4/5ths Rule is the most commonly used test for adverse impact.

Adverse Impact Using the Fisher's Exact Test

Tests of statistical significance are also allowed to examine for adverse impact (Hazelwood School District v. United States, 1977)

The EEOC mentions the 4/5ths Rule specifically because of it's simplicity of calculation – it doesn't require software or even technology for that matter

The Fisher's Exact Probability Test

The Fisher's Exact Test is a non-parametric test used to analyze differences in proportions. It is valid for all sample sizes unlike Chi-Square

This test is contained in most commercially available statistical software including SAS or SPSS, however there are free versions of this test available online including a free Excel template available at: http://adverseimpact.org/CalculatingAdverseImpact/StatisticalTests.htm

Applying to the Eight Score Use Methods to Data Set 1

Group	TD	TDG	FR	FP	FPR	SR	SP	SPR
White	84	70	80	77	79	80	71	77
African-American	37	51	41	44	42	41	50	44
4/5ths Rule	.68	.98	.77	.78	.76	.70	.88	.78
Fisher's Test	.03*	.92	.15	.26	.13	.09	.69	.26

Adverse impact indicated by using the 4/5ths Rule is highlighted in **red**
*Denotes adverse impact applying Fisher's Exact Probability Test

Findings of adverse impact depend upon the operational definition used

Applying to the Eight Score Use Methods to Data Set 2

Group	TD	TDG	FR	FP	FPR	SR	SP	SPR
White	51	45	50	49	49	49	47	48
African-American	36	42	37	38	38	38	40	39
4/5ths Rule	.74	.98	.81	.85	.85	.85	.94	.89
Fisher's Test	.15	.91	.34	.48	.48	.48	.81	.63

Adverse impact indicated by using the 4/5ths Rule is highlighted in **red**
*Denotes adverse impact applying Fisher's Exact Probability Test

Only Strict Top-down produced adverse impact in this example – it may have been different if other selection ratios were used

An Exception for Small Numbers

"Generally, it is inappropriate to require validity evidence or to take enforcement action where the number of persons and the difference in selection rates are so small that the selection of one different person for one job would shift the result from adverse impact against one group to a situation in which that group has a higher selection rate than the other group."

From the Uniform Employee Selection Guidelines Interpretation and Clarification (Questions and Answers, Question 21)

Utility

One of the more common method for assessing utility is to examine the mean z score of the groups selected

It may important to periodically examine the mean z score of candidates selected as an eligibility list is used, especially as time passes

As the mean score gets closer to z = 1.00, the more utility is being lost. To state the obvious, the point in using assessment is to try to avoid z – 1.00.

Z scores may not be easily understood if you are communicated test utility to someone without a background in testing.

Utility in Dollar Units

Brogden-Cronbach-Gleser Utility Formula

Expected Gain ($) = **(N) (T) (SDy) (rxy) (Zx) - (NT) (C)**

Where:

N = Number Selected

T = Tenure or predicted time in job is selected

SDy = Standard deviation of job performance in dollars

Rxy = Validity coeffient

Zx = Mean z score of applicants selected

NT = Number Tested

C = the cost of testing per applicant

Brogden-Cronbach-Gleser Utility Analysis

The following parameters were used:

- Number tested
  - Data Set 1 = 604 (Removing groups not in the analyses)
  - Data Set 2 = 532 (Removing groups not in the analyses)
- Tenure = 5 years
- SDy =$23,670 (.4 * $59.176; starting salary from O*NET)
- Rxy = .375 (Typically reported for public safety exams)
- Zx = obtain mean Z values for each of the different score use protocols
- C = $30 (includes test development or transportability study costs; other consideration may be facilities costs for agencies who don't own their testing site; this may be a very conservative estimate)

Utility Gain/Loss Data Set 1

	TD	TDG	FR	FP	FPR	SR	SP	SPR
Mean Z Score	1.74	1.38	1.66	1.63	1.59	1.54	1.59	1.42
Total BCG Utility	$9,325,908	$7,392661	$8,896,297	$8.735,194	$8,520,359	$8,251,857	$8,320,389	$7,607,460
BCG Utility per Year	$1,865,181	$1,478,532	$1,779,259	$1,747,039	$1,704,078	$1,650,376	$1,704,077	$1,521,493
BCG Utility per Hire per year	$15,415	$12,219	$14,705	$14,438	$14,038	$13,639	$14,083	$12,574

As typically found, the largest losses in utility occur with Top-down Within-group Selection

The difference between Strict Top-down and Sliding bands, Minority Preference, Top-down Selection Within Bands is relatively small

In practical terms, how big is this difference?

Utility Gain/Loss Data Set 2

	TD	TDG	FR	FP	FPR	SR	SP	SPR
Mean Z Score	1.58	1.18	1.41	1.38	1.37	1.33	1.43	1.33
Total BCG Utility	$6,086,607	$4,542,139	$5,430,208	$5,314,373	$5,314,373	$5,121,314	$5,468,820	$5,121,314
BCG Utility per Year	$1,217,321	$908,428	$108,604	$1,062,875	$1,062,875	$1024,263	$1,093,764	$1024,263
BCG Utility per Hire per year	$10,060	$7508	$8976	$8784	$8704	$8465	$9039	$8465

Again, the largest losses in utility occur with Top-down Within-groups Selection, and relatively small difference between Strict Top-down and Sliding bands, Minority Preference, Top-down Selection Within Bands.

Post hoc inspection of the data revealed a few significant outliers in the white examinee group

The Legal Status of Banding

Can be used with secondary criteria. For example, additional licensure, training or experience

If EEO Protect Class status is secondary criteria, it must be part of consent decree, court order, or voluntary AA plans (as long as selection from bands not based solely on minority preference)

Transparency may be important

May need to explain to examinees and employees on rationale of banding and why is it being used (e.g., what it is, how it works)

If used legally for diversity purposes may need to explain what Title VII or Affirmative Action and not simply some type of artificial quota system

Have a written policy describing all selection/promotional procedures

Any Questions???

Thank you for you kind attention!

There was a lot of information presented here and certainly a lot more that could have been presented due to time constraints.

Please address any correspondence to:
Frank Igou, Ph.D.
Associate Professor of Industrial-Organizational Psychology
116B Woodard Hall
Ruston, Louisiana 71270
(318) 257-5455
figou@latech.edu
frankigou@gmail.com