Using Difficulty Anchored Rating Scales in Setting Cut Scores: A New Angoff Modification

Calvin C. Hoffman,
LA County Sheriff's Department
and Alliant International University

C. Chy Tashima, LA County Sheriff's Department

Gypsi Luck, California State University, San Bernardino

Presented to IPAC, July 20, 2010

Overview

- Acknowledgements
- Summary of Angoff cut score method
- Angoff use
- Concerns regarding Angoff method
- Some common Angoff modifications
- Project (Sergeant exam)
  - Difficulty anchored rating scale
- Findings
- Practice implications

Acknowledgements

- Difficulty anchored rating scale is based on suggestions of Jerry Kehoe (personal communication, 2006)
- Without the efforts of the subject matter experts who supported this project, none of this work would have been possible
- An earlier version of this research was presented at the 2007 SIOP conference and 2007 PTC-SC meeting
- A more complete version of this paper is in press, *International Journal of Selection & Assessment*

Summary of Angoff Method

- Angoff method dates to 1971
- Judges estimate the proportion of *minimally qualified* persons who would answer each item correctly
- After judges perform ratings, proportions for each SME are summed to arrive at the Minimum Passing Level for each judge (MPL)
- Average MPL across judges = overall cut score (Ricker, 2003)

Angoff in Practice

- Angoff is most frequently used cut score method (Ricker, 2003)
- Angoff method:
  - simple to implement
  - easy to explain to lay audience
  - produces stable cut score estimates over time
- Plake, Impara, and Wilson (2000) reported inter-rater reliability of judges *across* years was as high as intra-rater reliability *within* years

Concerns with Angoff Method

- Number of judges (*U.S. v. South Carolina*, 1978)
- Choice of judges (see Berk, 1986)
- Training of judges (Hambleton, 2001)
- Extent to which judges represent multiple perspectives or constituencies (Busch & Jaeger, 1990)
- Berk (1996) - identifying borderline performers is a "nearly impossible cognitive task" (p. 216)
- Angoff method places a heavy cognitive demand on raters (see Impara & Plake, 1997; Berk, 1996; Shephard, 1995)

Some Angoff Modifications

- Iterative rating and feedback process
- Normative feedback modification
- Revise rating judgment into Yes/No decision
- Use item response theory (IRT) in setting cut scores
- Combinations of methods

Bowers & Shindoll (1989)

- Conducted a study comparing four different cut score methods (standard Angoff and normative feedback)
- 200-item professional certification test
- Five expert raters
- Compare findings for standard Angoff and normative feedback modification
- Normative feedback removes expert judgment

Bowers & Shindoll (1989) Results

Comparison	Standard	Normative
r Angoff rating with M Angoff	.55 to .78	.94 to .98
r Angoff rating with item *p*	.13 to .32	.92 to .97
r M Angoff rating with item *p*	.32	.99

Normative Feedback Modification

- Can be viewed as removing expert judgment and replacing it with item analysis results
- Requires knowledge of item characteristics **before** conducting Angoff ratings
  - Not feasible in this setting
  - Civil Service Rules require publication of cut score **before** test is administered

Hurtz & Auerbach (2003) Meta-analysis

- Examined multiple Angoff modifications:
  - Discuss minimal competence
  - Iterative feedback modification
  - Normative data modification
  - All possible two-way interactions
- Some Angoff modification results in higher reliability
- Some Angoff modifications or combinations of modifications resulted in **higher** cutoff scores than expected, an undesirable outcome where adverse impact is a concern (Ployhart & Holtz, 2008)

Current Application

- Promotional exam (Deputies to Sergeant)
- High stakes/high visibility test (litigious setting)
- Conducted detailed job analysis
- Exam components:
  - Written multiple-choice job knowledge test
  - Appraisal of Promotability
  - Structured panel interview
- Written test had *Reference* & *Recall* sections
- Results discussed here are part of *Recall* section

This Study

- Implemented a simple modification to the Angoff normative feedback method
- Needed a way to provide normative feedback while (hopefully) retaining expert judgment
- Solution was simple: use items from previous tests to provide normative information on relative difficulty
  - Rather than estimate difficulty in a vacuum, provides SMEs with an external reference

Difficulty Anchored Scale

- Based on 2004 item analysis results, we selected 9 items with *p*-values between .20 and .97
- Items were presented on a two-page document as a rating scale
- Arranged in order from easiest (high *p*-value) to hardest (low p-value) items
- No items on scale were being used on current test
- Simple and elegant way to provide normative feedback

Scale Format

P-value	Complete item text accompanies *p*
.97	
.83	
.74	
.52	
.42	
.34	
.20	

Angoff Procedure In This Study

- SME panel consisted of 10 Sergeants and Lieutenants
- Provided brief training session on Angoff and use of rating scale
- Discussed concept of minimum competence
- Practice ratings with feedback
- SMEs rated 116 items (102 retained)
  - 30 items were slightly revised and reused from 2004
  - Provided immediate ‘validity’ test since *p*-value estimates were available (>1800 candidates in 2004)

Results

- Reliability of Angoff ratings:
  - Reliability corrected for 10 raters (using Spearman-Brown prophecy formula) = .73
- 'Validity' of mean Angoff ratings for predicting empirical p-values:
  - .65 for items as presented in 2004 test
  - .73 for same items slightly edited and reused in 2006
- Correlation between actual p-values for 2004 and 2006 = .83 (stability over time)
- 'Validity' estimates were corrected (attenuation)

‘Validity’ Corrected for Attenuation

	Observed Correlation	Corrected Correlation
M Angoff with 2004 *p*-values	.65	.71
M Angoff with 2006 *p*-values	.73	.80

Individual Raters

- Individual raters varied in *reliability*:
  - Correlation of individual raters with mean of all raters ranged from -.09 to .79
  - Average ‘rater-total’ correlation = .53
- Individual raters varied in *validity*:
  - Validity ranged from .10 to .51
  - Validity of average of all raters was much higher (.63)
  - Dropping **“least valid”** raters led to a **decrease** in validity for average of remaining SMEs!
- Moral – be cautious when dropping raters from Angoff process!

Comparing B & S vs. Our Findings

Comparison	B & S (Standard)	B & S (Normative)	This Study
r Angoff rating with M Angoff	.55 to .78	.94 to .98	-.09 to .79 (median .58)
r Angoff rating with item *p*	.13 to .32	.92 to .97	.10 to .55
r M Angoff rating with item *p*	.32	.99	.73

Clarification

- We have used standard psychometric terms regarding reliability and validity
- The reliability aspect is straightforward and needs no further discussion
- The validity aspect merits further attention
- "Validity" here is the correlation between Angoff ratings of item difficulty and actual item difficulty – it is **not** equivalent to the concept of "validating" a cut score
- While some authors describe 'validating' cut scores, it is **not** possible to validate a cut score (SIOP *Principles*, 2003; Kehoe & Olson, 2005)

Practice Implications – Current Study

- Taube (1997) argues against dropping judges from Angoff panels (representation issue)
- In our study, even least valid raters made unique contributions to validity of panel's ratings
- Critics of the normative information Angoff modification argue that providing too much normative information to raters can remove the judgmental aspect of the Angoff rating task (Garrido & Payne, 1991; Wheeler, 1991)

Practice Implications – Current Study

- The difficulty-anchored rating scale provided raters with no normative information *on current items*
- Scale should make the Angoff rating process less complex for raters
- The rating scale as packaged did not "give away" information about any item in the current examination

Difficulty Anchored Rating Scale

- Easy to develop
- Provides normative information without sharing item results
- Produced high reliability and “validity”
- Efficient use of rater time
- We encourage others to try this simple Angoff modification

Thank You