Reliability and Validity of Angoff Ratings

J. Anthony Bayless
Henry Busciglio

Personnel Research and Assessment Division
Office of Human Resources Management

U.S. Customs and Border Protection

Standard Setting

- Process to establish a performance standard, cut score, or passing score
- Process not purely technical or empirical
- Process involves value judgments (*Standards for Educational and Psychological Testing*)
- Various methods of standard setting, for example:
  - Contrasting Groups and Borderline Groups (Livingston & Zieky, 1982)
  - Angoff (1971)
  - Ebel (1972)
  - Nedelsky (1954)

Angoff Procedure

- SMEs are administered the test
- SMEs estimate the proportion of “minimally qualified” or “minimally competent” examinees who would answer each item correctly
- Average Angoff rating is calculated for each item
- Grand average of the Angoff ratings across items is calculated to represent the recommended performance standard (or cut score)

Promotional Assessments

- Career Experience Inventory
- Critical Thinking Skills
- In-Basket Job Simulation
- Managerial Writing Skills
- Job Knowledge Test

Job Knowledge Test

- 80 items for each occupation’s (IEA and DO) test
- Multiple-choice items with four response options
- Dichotomously scored items
- Power tests

Research Interest

- How good are SMEs at conceptualizing and consistently applying a hypothetical construct of "minimally qualified" examinees?
  - Specifically, how reliable are the SME estimates?
  - Specifically, how valid are the SME estimates?

Methodology – Angoff

IEA SMEs	DO SMEs
n=5 (Time 1 + Time 2)	n=8
No group discussion	Group discussion

Methodology - Study

- Two post hoc studies, one per occupation
  - DO sample (N=259 examinees)
  - IEA sample (N=318 examinees)
- Assessed interjudge reliability via internal consistency estimate of reliability
- Assessed validity via correlation of average Angoff rating and actual (observed) item difficulty index for a “minimally qualified” group of examinees

Results - Reliability

- DO Sample (72 scored items, 8 SMEs)
  - Alpha = .863, no removable SMEs
  - Item-total correlations from .582 to .680

- IEA Sample (70 usable items, 5 SMEs)
  - Initial Alpha = .429, with 2 removable SMEs
  - Final Alpha = .547, using 3 SMEs
  - Item-total correlations from .364 to .422
  - We used both 5- and 3-SME groups for further analyses.

Results - Validity

- Validity - agreement between SMEs' Angoff estimates and actual p-values among group of "minimally qualified" test takers.
- "Minimally qualified" defined two ways:
  - Candidates scoring close to 50th percentile
  - Candidates getting 70% of items correct
- Used both correlations and t-tests to assess validity

Results – Validity (Corr.)

- For DO sample, correlations were:
  - .591** for 50th percentile group
  - .479** for 70% correct group

- For IEA sample, correlations (for 5- and 3-SME groups, respectively) were:
  - .311** and .243* for 50th percentile group
  - .282* and .183 for 70% correct group

** p<.01. *p<.05.

Results – Validity (T-tests)

- Agreement – magnitude of mean differences between the Angoff ratings for each item and the corresponding p-value among minimally qualified test takers.
- Used paired-samples t-tests
- For DO sample:
  - Grand average Angoff rating = .6310
  - Average p-value for 50th percentile group = .6315
    - t = 0.025, df = 71, p = .980
  - Average p-value for 70% correct group = .6906
    - t = 2.750, df = 71, p = .008

Results – Validity (T-tests)

For IEA sample:

- Grand average Angoff ratings
    - 5-SME = .7716
    - 3-SME = .7710
- Average p-values
    - 50th percentile group = .6810
    - 70% correct group = .6980

Results – Validity (T-tests)

For IEA sample, continued:

- Comparisons:
  - 1: 50th perc p-values compared to 5-SME Angoffs
    - t = -3.233, p = .002
  - 2: 70% corr p-values compared to 5-SME Angoffs
    - t = -2.685, p = .009
  - 3: 50th perc p-values compared to 3-SME Angoffs
    - t = -3.148, p = .002
  - 4: 70% corr p-values compared to 3-SME Angoffs
    - t = -2.587, p = .012

Results – Validity (T-tests)

IEA T-Test Comparisons		
	50th Percentile *p*-values	70% Correct *p*-values

Avg. Angoffs for 5 SMEs	t = -3.233
	p = .002
		t = -2.685
		p = .009

Avg. Angoffs for 3 SMEs	t = -3.148
	p = .002
		t = -2.587
		p = .012

Results – Summary

- DO SMEs gave reasonably reliable and valid estimates of actual p-values, especially for test takers at the 50th percentile.
- IEA SMEs gave less reliable and valid estimates by exhibiting less interrater agreement, demonstrating less insight into the relative difficulty of items, and overestimating p-values.
- The notably superior performance of the DO SMEs is reasonable given the differences between the procedures used to obtain Angoff estimates from the two groups.

Limitations of Current Study

- Post hoc studies
- Did not retain initial round of Angoff ratings prior to group discussions during second round

How Does This Help You?

- The more SMEs, the merrier!
- Group discussion is critical
- SMEs need to be experienced and representative of occupational workforce

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (pp. 508-600). Washington, DC: American Council on Education.

Cizek, G.J. (2001). Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Associates.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational measurement: Issues and practice,* 23(4), 31-50.

References (continued)

Ebel, R.L. (1972). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Goodwin, L.D. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline examinees. *Applied measurement in education, 12(1),* 13-28.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and psychological measurement, 14,* 3-19.