Evaluating a Strategy for Setting Cut Scores on a Computer Adaptive Test

Deanna L. Morgan, The College Board

Chad W. Buckendahl, Alpine Testing Solutions

CollegeBoard
inspiring minds™

Introduction

- Use of CATs have steadily increased
  - Currently planning underway for implementation of Common Core Standards and the Smarter Balance Consortium is focused around a model using CAT for a large part of the assessment system.
- Setting cut scores with a CAT can be challenging
  - Potentially infinite number of test forms
  - Administration characteristics may be unfamiliar to panelists

“Wainer Method”

- Modification of Angoff (1971) dichotomous judgment method that leverages a CAT algorithm as experienced by the examinees.
- Strategy was conceptualized by Lunz (2000) and Sireci and Clauser (2001) but limited empirical evidence of successful application is available.
- O’Neill, Tannenbaum, and Tiffen (2005) applied the methodology to TOEFL in the context of the nursing licensure examination program.

Current Study

- Applied “Wainer Method” to a fixed-length CAT as one part of a two part study for a multiple-choice writing assessment. Only the part of the study pertaining to the CAT will be included.
- Three sources of validity evidence as described in Kane (1994, 2001) will be used to organize the presentation.
  - Procedural, Internal, External

Procedural

- Panelists were solicited by the state higher education coordinating board to represent multiple stakeholder groups.
  - Included high school teachers, developmental writing instructors, freshman composition instructors, and higher education administrators from across the geographic regions and institution types in the state.
  - 30 panelists – 26 females, 4 males
  - Teaching experience – Mean 18.2 years, SD 9.34 years

Procedural

- Orientation including the purpose of the standard setting meeting, intended use of the resultant cut score, and a description of the test (including a high level overview of the CAT algorithm) began the meeting.
- Panelists were broken into 5 smaller groups of 6 people to discuss and draft performance level descriptors (PLDs) of the borderline examinee.
- Large group discussion of the drafts created in the small groups followed and was led by the facilitator to produce the final PLDs which were then transcribed, copied, and distributed.

Procedural

- Panelists received training on the methodology where they were instructed to take the test responding to each question either correctly or incorrectly as they would expect the borderline examinee to respond without consideration for which incorrect answer to choose – only that it was incorrect.
- During training panelists took the test multiple times changing their response patterns to help them understand the adaptive nature of the test

Procedural

- Panelists completed two rounds of ratings by taking the test as if they were the borderline examinee.

- Between Round 1 and Round 2 of ratings, a large group discussion was held and feedback provided which included the individual cut score of each panelist, the panel's average and median cut score, and impact data based on a nationally representative norm group of examinees who have taken the test.

Procedural

- Panelists completed an evaluation form following Round 2 concerning:
  - Efficacy of the orientation
  - Understanding of the PLDs and borderline examinee
  - Training on the rating task
  - Helpfulness of discussion and feedback
  - Level of confidence in the resulting standards
- All mean ratings for training and adequacy of time allowed were between 3.8 and 4.0 on a scale from 1= Unsuccessful/Inadequate to 4 = Successful/Adequate.
- Confidence ratings ranged from 1 = Not Confident to 4 = Confident with a mean rating of 2.75.

Procedural

- Panelist comments on the evaluation forms and during Round 2 discussion indicated that there was some confusion about the rating task during Round 1.
- Specifically, panelists spent a large amount of time debating over which incorrect answer to choose even though instructed that it did not matter which as long as it was incorrect when appropriate.
- Additionally, one panelist indicated that they did not understand during Round 1 that they were taking it as the borderline student and had an “Aha!” moment during the discussion between rounds of ratings.

Internal

- Test score scale ranges from 20 -120.
- Round 1 median recommended cut score was 54.0 with standard error of 29.6.
  - Large range of variation most likely due to confusion about providing an incorrect response and focusing on the borderline examinee.
- Round 2 median was 69.0 with a standard error of 25.0.
  - Variation in Round 2 was slightly smaller but there is still concern that panelists are having difficulty connecting their judgments to the score scale due to the adaptive nature of the test.

External

- Collecting and evaluation external evidence is challenging for testing programs.
- In this study no resources were available to collect external evidence but it is important that moving forward, especially for potentially large reaching cut scores as expected for Smarter Balance that the collection of this evidence be planned.
- Using a second methodology to collect additional evidence and determine reasonableness of cut scores and/or consideration of historical data may be used to provide external evidence.

Recommendations/Research Needed

- Consideration of panelists familiarity with technology, adaptive testing, etc. in future research
- Determining how to modify standard training protocols to ensure panelists understood their task and how their judgments connect to the recommended cut score.
- How should judgments be made to consider examinee's experience with an adaptive test?
- What type of feedback would be meaningful to panelists to help them understand, inform and modify their judgments?
- Measurement practitioners and policymakers need to consider impact of technology and measurement strategy as they interact in determining final cut scores.