THE PROBLEM OF THE SALTATORY CUT-SCORE: SOME ISSUES AND RECOMMENDATIONS FOR APPLYING THE ANGOFF TO TEST ITEM BANKS

William Coscarelli, Andrew Barrett, John Kleeman and Sharon Shrock

The Problem of the Saltatory Cut-Score: Some Issues and Recommendations for Applying the Angoff to Test Item Banks

William Coscarelli, Curriculum and Instruction, Southern Illinois University, Carbondale, IL 62901-4610, coscarel@siu.edu

Andrew Barrett, Curriculum and Instruction, Southern Illinois University, Carbondale, IL 62901-4610, abarrett@siu.edu

John Kleeman, Questionmark Computing Limited, 5th Floor, Hill House, Highgate Hill, London N19 5NA, john.kleeman@qmark.co.uk

Sharon Shrock, Curriculum and Instruction, Southern Illinois University, Carbondale, IL 62901-4610, sashrock@siu.edu.

Abstract

A fundamental issue in criterion-referenced test (CRT) development is: What should the cut score be to determine mastery? The literature has suggested three types of strategies for answering this question: Informed Judgment, Contrasting Groups, and Conjectural Techniques. For a number of reasons, the Conjectural approaches are probably the most common solution to this problem; and within this class, the Angoff is probably the most commonly used technique for setting cut scores.

The Angoff uses subject matter experts (SMEs) to review each item and assign a weight to the item based on the SME's conjecture that a minimally competent performer would answer the item correctly. These weights—which are fundamentally different from a traditional difficulty index—are then summed to provide the initial recommendation for the cut score. As CRT test development has become more widespread the use of multiple forms of the same test has also become more common. The use of computerized test development tools allows for random selection of questions that would make the number of forms combinatorially large. And thus, a new problem is created. Theoretically each form of the same test could have a different cut-score. This bouncing score would be defensible from a statistical perspective, but might give organizations implementation challenges for political and perhaps legal reasons.

In this paper we look at how the concept of using Angoff weights to determine a cut score for an assessment where questions are selected at random might work, and give an illustrative example to allow people to consider it in action. We are

using a true data set from a certification test and will look at the differences in cut scores using three assumptions: 1)) random sampling from the data set, 2) a "random" sample that draws on extremes of the data set, and, 3) stratified random sampling of the set. We then conclude with suggestions for sampling from item banks based on the size of the bank and the criticality of the test.

Introduction

A fundamental issue in criterion-referenced test (CRT) development is: What should the cut score be to determine mastery? The literature has suggested three types of strategies for answering this question: Informed Judgment, Contrasting Groups, and Conjectural Techniques. Since the Conjectural approaches rely on the use of SMEs, they have become a preferred choice in that they are cost-effective and efficient methods for determining a mastery score. Of the Conjectural approaches probably the most common approaches is the, Angoff technique which asks SMEs to assess the probability a minimally competent master will pass a given item. Thus, each item has a value assigned from 0.0 to 1.0 based on the SME's estimate of a master's performance. (It should be noted that the Angoff is sensitive to difficulty and importance. Thus, for a welder, putting on safety glasses is easy, but all would be expected to be able to perform this task—while landing a plane in a wind shear is difficult, but all pilots would also be expected to pass this task.)

Now, as CRT test development has become more widespread so too has the use of computerized test development tools allows which allow for random selection of questions from a data bank and thus can present many different tests from the same bank that would, in fact, make the number of forms combinatorially large—and thus reduce errors in testing due to familiarization with items or attempts at cheating.

And thus, a new problem is created. Theoretically each form of the same test could have a different cut-score because each item has a unique Angoff weight. The idea that one person passed an exam with a score of 82% while another failed a test on the same content with a score of 83% would be hard to explain to anyone but a psychometrician. This bouncing score would be defensible from a statistical perspective, but might give organizations implementation challenges for political and perhaps legal reasons.

This presentation will look at how the concept of using Angoff weights to determine a cut score for an assessment *where questions are selected at random from within a single objective* might work. We are using a true data set from a certification test and will look at the differences in cut scores using three assumptions: 1) random sampling from the data set, 2) a "random" sample that draws on extremes of the data set, and, 3) stratified random sampling of the set.

The Data Set

We began with a data set of Angoff weights for a single topic (objective).

The data has the following attributes:

- A population of 50,
- A mean = 0.8420,
- A standard deviation = 0.12712,
- Skew=-0.279, and
- Kurtosis=-.550.

The following table shows the frequency distribution of the Angoff weights in the data set.

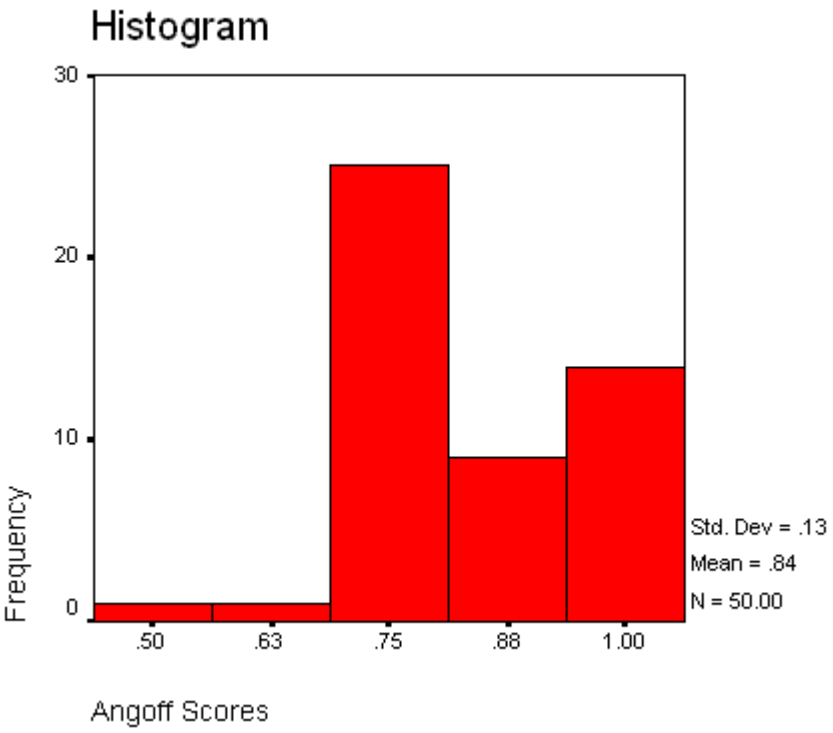

Table 1. Frequency distribution of Angoff weight

Choosing the Test Items and Potential Outcomes

Now, from this bank of items, we could describe three common approaches to sampling the items and their outcomes: We could

- randomly sample and find the sample matches the distribution,
- randomly sample and find the sample does not match the distribution,
- stratify the sample to approximate the distribution.

We began with the assumption in the data set that if the skew and kurtosis are less than plus or minus 1 then the data from the population can be considered normally distributed—which they are in this example even though it is a CRT based test.

We then drew three samples to simulate a test taker having three chances to pass a test of 10 questions (20% of the item bank) each. (Note: Remember that the tests are drawn from an item bank that represents a single topic or objective. Most tests would be longer as they would include items from multiple topics.)

We began the exploration of the issue by looking at means and associated z-scores and found the following results:

Randomly sample and find the sample matches the distribution

We determined the cut-score for three tests based on a simple randomization. The mastery levels were: 8.35, 8.55, and 8.75, which proves to be statistically similar.

Randomly sample and find the sample does not match the distribution

Next we considered the instance when the randomization created three means due to the fact it sampled the highest weighted items, the average weights, and the lowest weighted items. This would be an uncommon, though statistically possible outcome. We found the following for a test of 10 items:

- Highest cut-score weight was 1.0. (standard deviation=0)
- Random sample cut-score was .84. (standard deviation=0.141)
- Lowest cut-score weight was .67. (standard deviation=0.068)

Here we found that both the highest and lowest groups were significantly different.

Stratify the Sample to Approximate the Distribution

Finally we used a stratification algorithm to generate a sample. The algorithm was designed to force a selection process that might accurately reflect the distribution of the Angoff weights in the question bank.

1. The algorithm begins by first determining the median Angoff score of the item bank. The median Angoff score of the item bank is used to divide the item bank into two groups. The first group contains all the items whose Angoff scores are less than or equal to the median Angoff score for the entire item bank. The second group contains all the items whose Angoff scores are greater than the median Angoff score for the entire item bank. The heart of the algorithm involves randomly sampling only from one of these groups at a time.

2. The first item that will be selected for the test is still randomly drawn but only from the items in the item bank that have an Angoff score that matches the item bank median.
3. The second item selected to be part of the test is randomly drawn only from those items with Angoff scores that are less than or equal to the median Angoff score.
4. Similarly, the third item selected to be part of the test is randomly drawn only from those items with Angoff scores larger than or equal to the median Angoff score.
5. The remaining items are alternately drawn in the same way the second and third items were selected, from the group with the smaller scores then from the group with the larger scores.

(For example, if we wanted to create a test with 5 items from an item bank with 50 items and the median Angoff score is 0.8 then:

1. The first item selected is randomly drawn from those items in the bank with an Angoff score of 0.8.
2. The second item selected would be randomly drawn from the remaining items that are less than or equal to the median Angoff score of 0.8.
3. The third item selected would be randomly drawn from the remaining items that are greater than the median Angoff score of 0.8.
4. The fourth item would be drawn in the same way as the second item (from the group with the smaller scores).
5. The fifth item would be drawn in the same way as the third item (from the group with the larger scores).

In the end our 5 item test would have one item with an Angoff score equal to the median of the item bank (0.8), two items with Angoff scores smaller than or equal to the median Angoff score, and two items with Angoff scores greater than the median Angoff score.)

There are two main benefits of using this algorithm. First, it is not possible to end up with a test with an extremely high or an extremely low cut off score. The second benefit of this algorithm is that it is relatively simple. (And, of course, the algorithm is only as good as the quality of the Angoff weights that have been made and assigned to *each* item.)

It is important to note that using this algorithm does not completely eliminate the bouncing cut off score phenomenon; it just makes the bounce a little less bouncy. It is still possible to produce tests that are significantly different but it makes that possibility far less likely. The only way to eliminate the bouncing cut off score completely is to force all the tests to have the same cut off score but this has the negative consequence of drastically reducing the number of different tests that are possible from a given test bank. The algorithm provides a good compromise

between pure random assignment and forced uniformity in cut off scores. The algorithm does start off at the median and this does provide more exposure to items with this statistical characteristic. This was done to anchor the sample, but if one were concerned that doing so provides more exposure for the median level Angoff items, then one might begin with a random item selection and then proceed with the algorithm.

Using the algorithm, the mean for the third sample was 0.85 and the standard deviation was 0.1117. The cut of score for this sample was 8.55. Again we found that both the highest and lowest groups were significantly different from the population mean.

Detailed Statistical Analysis

We decided to re-examine the data to determine if there was a difference among the means chosen from the sample based on seven possible conditions. (We did this realizing that conducting multiple z tests increases the chances of producing a type I error; so we then looked at the data from a more stringent point of view using ANOVA and a Tukey Multiple Range Test.)

We created tests that drew:

- from the lowest Angoff weights
- from the highest Angoff weights
- based on a stratification algorithm
- randomly

Using SPSS, the following one-way ANOVA table was generated.

ANOVA

Angoff Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.558	5	.112	9.101	.000
Within Groups	.662	54	1.226E-02		
Total	1.220	59			

An F of 9.101 was found to be significant at an alpha of less than .001. Subsequent Tukey post hoc tests revealed differences among groups as show by the following homogenous subsets.

Angoff Score

Tukey HSD[a]

Sample	N	Subset for alpha = .05		
		1	2	3
Lowest	10	.6700		
Random 1	10		.8350	
Random 2	10		.8550	.8550
Stratified	10		.8550	.8550
Random 3	10		.8750	.8750
Highest	10			1.0000
Sig.		1.000	.965	.053

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 10.000.

This can also be represented by the following multiple range test:

Lowest Random 1 Random 2 Stratified Random 3 Highest

_____

_____

_____

One can see three outcomes from this range test:

- There was a statistically significant difference between the Lowest and Highest mean samples.
- There was never a difference among the Randomized and Stratified samples.
- There was a statistically significant difference between the Lowest and all the other sample means.
- There was only one random sample that was statistically significantly different from the Highest mean sample.

Recommendations

There are several factors that come into play when considering using random or stratified random sampling. The size of the sample influences the likelihood that a given sample will differ significantly from the item bank. More importantly the consequences of passing or failing the test will drive which sampling method to use.

In very small samples sizes it is very unlikely that even extreme samples will be significantly different from item bank. Since Angoff scores range from 0 to 1, the

range of possible cut off scores with a sample of 3 items is 0 to 3. This small range forces all the possible cut off scores to be fairly close. The cut off scores get even closer when you consider that it is unlikely that an item with a score of 0 would ever be included in an item bank since getting such an item correct or incorrect would not influence an individual's overall score. Based on statistical tests similar to those described above, it was found that there were only differences among extreme groups (the highest Angoff scores and the lowest Angoff scores) when samples were 3 items in a test. When the number of items in a test rose to 5 or higher it was found that there were some differences, not only among the extreme groups but also among the extreme groups and more typical groups.

Looking at the study as a whole it would seem that we might make the following suggestions:

- For small samples, random selection of items seems acceptable. (The only wrinkle is that when there is a small sample there is a greater likelihood that an extreme group will be selected (higher chance of picking 3 items with Angoff score of 1 than picking 10 items with an Angoff score of 1).
- The larger the sample the greater need to stratify.

Finally, if the CRT test is being used to make critical decisions about the individual taking the test then great care needs to be taken to ensure that the likelihood of passing or failing a test due to a bouncing cut off score is extremely low. It is worth repeating that using the algorithm described above does not guarantee that the bouncing cut off score problem will never appear but it does make that problem much less likely. Beyond reducing the likelihood of the problem, using such an algorithm also demonstrates that a reasonable effort was made to avoid the problem.

It should be clear that using pure random item selection for high stakes tests could lead to legal, ethical, or professional problems if the results of the CRT test are ever called into question. Less critical tests are less likely to be put under the same level of scrutiny so random sampling of the item bank may fine but stratification is still safer. In summary

- For low stakes tests randomly sample within the item bank
- For medium stakes tests, one can probably sample within the bank if the distribution is statistically normal, but stratification is safer.
- For high stakes tests, one should consider stratification of the sample for increased precision.