10-23-2008

The Effect of Deleting Anchor on the Classification of Examinees

Tia Sukin
University of Massachusetts - Amherst, tsukin@educ.umass.edu

Lisa Keller
University of Massachusetts Amherst, lkeller@educ.umass.edu

The Effect of Deleting Anchor Items on the Classification of Examinees

Tia Sukin

University of Massachusetts Amherst

PO Box 52; Lake Pleasant, MA 01347

tsukin@educ.umass.edu

(843) 323-6927

Lisa Keller

University of Massachusetts Amherst

School of Education-Hills South; 813 North Pleasant Street; Amherst, MA 01003

lkeller@educ.umass.edu

(413) 545-1528

Abstract

The presence of outlying anchor items is an issue faced by many testing agencies. The decision to retain or remove an item is a difficult one, especially when the content representation of the anchor test is in question. Additionally, the reason for the aberrancy is not always clear, and if the performance of the item has changed due to improvements in instruction, then removing the anchor item may not be appropriate and might produce misleading conclusions about the proficiency of the examinees. This study examines the effect of removing or retaining one aberrant anchor item. The degree of aberrancy was manipulated as well as the ability distribution of examinees, and four IRT scaling methods were investigated (Mean-sigma, mean-mean, Stocking & Lord, and Haebara). The results indicate that the percent of correctly classified students was not affected by either retaining or removing the aberrant item, although the over- and under-classification of examinees was. There was no difference among the methods.

The Effect of Deleting Anchor Items on the Classification of Examinees

In item response theory (IRT), item parameters are assumed to be invariant to the sample of examinees that respond to items, and person parameters are assumed to be invariant to the set of items to which the examinee responds. This is the property of parameter invariance, which allows for the comparison of scores from different test forms through equating. The importance of equating cannot be overstated, especially in the era of No Child Left Behind (NCLB) where it is essential to monitor the performance of students across years.

While item parameters are invariant, they are only invariant up to a linear transformation, which results in the so-called identification problem. The identification problem is usually resolved using one of the popular IRT scaling techniques: Mean-sigma (MS), mean-mean (MM), Stocking and Lord (SL), Haebara (HB), or fixed common item parameter (FCIP). The literature has shown that each of these methods results in slightly different outcomes, and some methods are more robust than others to different testing contexts. Most of these methods are used in the context of the non-equivalent groups anchor test design (NEAT), where a small sample of equating items is presented on both forms of the tests that are to be equated. This practice serves to separate the sources of the differences in the scores of the two groups on the two tests: differences due to the difficulty of the test and differences due to the abilities of the separate groups. Scores can then be adjusted only for the differences in the difficulties of the two tests, thereby preserving the differences due to differing ability.

While the property of parameter invariance is a property of the parameters, it is often applied to the parameter *estimates*. To the extent that the estimates are not invariant, the scaling and equating that result from using this property may not be accurate. As such, it is routine for testing companies to evaluate the functioning of the items used for equating. This is done by

comparing the parameter estimates of the equating items on the two test forms; if the relationship between the item parameters is not linear, then the invariance of the estimate is suspect. In some cases, there are problems and the anchor items do not function the same way from year to year. This could be due to changes in instruction and curricular emphasis or the exposure of an item, to name a few reasons. In these instances, a decision must be made r to include the item in the equating or not.

Previous research involved using real data to determine the effect of removing items from the equating. Michaelides (2006) used the delta plot method to identify aberrantly performing items and considered, among other things, the effect on the classification of examinees into two categories. The study examined four operational tests. Across the four assessments, between one and three items were flagged as being aberrant. The effect of including or excluding the aberrant item was as predicted; if the examinees in Year 2 were performing higher than those in Year 1, including the item in the anchor led to a higher percentage of students being classified as proficient. The study considered the mean/sigma and Stocking & Lord equating methods. Both methods were similarly affected by the inclusion of the aberrant item. Because the real data were used it is impossible to determine which of the classifications was more accurate.

One simulation study investigated the effect of outlying anchor items when using MS and SL scaling methods for a number of conditions where a different method of flagging aberrant items was used (Hu, Rogers, & Vukmirovic, 2008). In this study, the authors found that including the aberrant items led to more systematic error in the equated scores, as would be expected. However, the effect of including/excluding the aberrant item on the classification of examinees was not explored.

While there have been studies that have examined the effect of removing anchor items using real test data, there are no studies that could be found that have examined the effect of removing items from the equating on the accuracy of the classification of students into performance categories using simulated data. The benefit to using simulated data is that the true classification of the examinees is known, and the effect of retaining or deleting an anchor item on the accuracy of the classification of students can be ascertained. Therefore, this study seeks to investigate the effect of removing aberrantly performing anchor items on the classification of students into performance categories. The importance of this study is clear in the wake of NCLB, where the accurate classification of students into performance categories is essential. Additionally, with assessments influencing the content of instruction, the likelihood of finding aberrant anchor items is high and so deciding how to deal with these items is of the utmost importance.

Method

The purpose of this study is to investigate the effect of removing aberrant anchor items on the classification of examinees into performance categories. This is accomplished by conducting a simulation study so that the true classification of the examinees is known. Fifty replications are simulated for each condition and the general outline of the study follows:

1. Simulate item response data for two administrations of an exam, including an aberrant anchor item;
2. Calibrate the items using the three parameter logistic model [PARSCALE: SSI, 2003];
3. Determine whether there are aberrant anchor items;

4. Equate the two test forms both with and without the aberrant anchor item(s) [STUIRT: Kim & Kolen, 2004];
5. Classify the examinees into performance categories based on the equated ability distributions obtained in step 4, both with and without the aberrant anchor item(s);
6. Compare the classification of the examinees in step 5, with the true classification of the examinee; and
7. Decide which classification in step 6 leads to the most accurate classification.

Details for each of these steps are provided in the next section.

Test Design

Two administrations of the test were simulated using the NEAT design. A twenty-item internal anchor was chosen such that the anchor test was as similar as possible to the total test in terms of average difficulty and discrimination (see Table 1).

Table 1: Average Item Statistics for Year 1, Anchor, and Year 2 Forms

	Year 1 Form		Anchor Test		Year 2 Form	
	Mean	SD	Mean	SD	Mean	SD
a	1.05	0.26	1.07	0.19	1.07	0.24
b	-0.03	1.00	-0.03	1.02	-0.03	1.00

Note: a = discrimination, b = difficulty, SD = standard deviation.

Parameters

The item parameters used for generating the item responses were obtained from an operational statewide testing program. Sixty dichotomously scored items were chosen for the simulation. The three-parameter logistic IRT model (3PLM) was used to simulate the item responses using WINGEN 2 (Han, 2008). Five thousand examinees were simulated for each of

the two administrations. In the first case, examinees for both administrations were drawn from a N(0,1) distribution. In the second case, growth was simulated between the two years. For the first administration, examinee ability parameters were drawn from a N(0,1) distribution while the examinee ability parameters for the second distribution were drawn from a N(0.2, 1) distribution.

Scaling Methods

Four different scaling methods were examined in this study: Mean-sigma (MS), Mean-mean (MM), Stocking and Lord (SL), and Haebara (HB). These four methods could be classified into one of two categories: moment methods or characteristic curve methods. MM and MS are moment methods, as they use only the first moment (MM) or the first two moments (MS) of the distribution of the item parameters. SL and HB, on the other hand, seek to minimize differences between the characteristic curves of the anchor tests. SL seeks to minimize the differences of the test characteristic curves, while HB seeks to minimize the difference between each of the item characteristic curves. Details of all methods can be found in Kolen and Brennan (2004).

Cut Scores

To simulate a common practice in many testing programs, examinees were classified into one of four performance categories, based on three cut scores. The cut scores were chosen arbitrarily and do not reflect the operational cut scores of the test that was used as the basis of the simulation. Cut scores of -0.75, 0, and 0.75 on the theta metric were chosen to classify examinees.

Aberrant Item

One item was chosen and simulated to be aberrant. This item was made aberrant by shifting the b-parameter between administrations. The b-parameter was shifted by two different values, 0.5 and 0.8, to simulate two degrees of aberrancy. To detect the aberrant item in the data, the "0.3 Rule" (see Huff & Hambleton, 2001), whereby an anchor item is considered aberrant if the difference in the b-values between administrations is greater than 0.3.

Fifty replications were simulated for each condition.

Evaluation Criteria

To compare the effect of aberrancy on the different scaling methods, the classification accuracy was determined for each of the methods both with and without the aberrant anchor item. Therefore, for each examinee, three classifications were determined:

1. True Classification-- the classification of the examinee based on the true known theta value.
2. Aberrant Classification-- the classification of the examinee obtained when the aberrant item is left in the anchor for equating.
3. Purified Classification-- the classification of the examinee obtained when the aberrant item is removed from the anchor for equating.

Using these three classifications, two contingency tables were created for each method and each replication. Comparisons were made between the true classification and the aberrant classification as well as the true classification and the purified classification. Examinees could be placed into one of sixteen categories as shown in the example presented in Table 2.

Table 2: Contingency Table for Four Performance Levels: Aberrant Classification vs. True Classification

		Aberrant Classification			
		1	2	3	4
True Classification	1				
	2				
	3				
	4				

The black categories indicate a correct classification of examinees, the gray categories represent an over-classification of examinees, and the white boxes represent an under-classification of examinees.

For each method, the percent of accurately classified, over-classified, and under-classified students was computed for each replication and averaged over replications. Thus, for each method the effect that an aberrant anchor item had on the accuracy of classification of examinees into performance categories could be determined.

Results

The proportion of correctly classified students for each of the scaling methods is presented first, followed by the over- and under-classification rates. The results are summarized for each of the two ability distribution conditions: null--where there was no change in the ability of the examinees between the administrations--and the mean-shift case-- where there was a shift of 0.20 standard deviations between administrations. Likewise, there were two conditions of aberrancy for the anchor item: one case where the b-value was shifted by 0.50, and one where

the b-value was shifted by 0.80. Table 3 provides the classification accuracy for each method and all conditions. As can be seen in Table 3, regardless of the condition, the proportion of students accurately classified was the same whether the aberrant item was retained or removed.

Table 3: Proportion of Students Accurately Classified, by Method, for all Conditions

		Ability Distribution			
		Null		Mean Shift	
		Degree of Aberrancy			
Scaling Method		*0.50*	*0.80*	*0.50*	*0.80*
MM	Aberrant	0.69	0.69	0.69	0.69
	Purified	0.69	0.69	0.69	0.69
MS	Aberrant	0.69	0.69	0.69	0.69
	Purified	0.69	0.69	0.69	0.69
SL	Aberrant	0.69	0.69	0.69	0.69
	Purified	0.69	0.69	0.69	0.69
HB	Aberrant	0.69	0.69	0.69	0.69
	Purified	0.69	0.69	0.69	0.69

While the overall classification accuracy was not affected by the existence of the aberrant anchor item, the proportion of students over- and under-classified was. The proportion of under-classified students is provided in Table 4. For both ability distribution conditions, the pattern of results is the same. In the case where the anchor item was changed by 0.50, leaving in the aberrant item led to 1% fewer examinees being under-classified as compared to when the item was removed from the equating. When the degree of aberrancy increased, the percent of under-classified students was 2% less when the aberrant anchor item was included.

Table 4: Proportion of Students Under-classified, by Method, for all Conditions

		Ability Distribution			
		Null		Mean Shift	
		Degree of Aberrancy			
Scaling Method		*0.50*	*0.80*	*0.50*	*0.80*
MM	Aberrant	0.17	0.16	0.17	0.16
	Purified	0.18	0.18	0.18	0.18
MS	Aberrant	0.17	0.16	0.17	0.16
	Purified	0.18	0.18	0.18	0.18
SL	Aberrant	0.17	0.16	0.16	0.16
	Purified	0.18	0.18	0.17	0.18
HB	Aberrant	0.17	0.16	0.16	0.16
	Purified	0.18	0.18	0.17	0.18

The percent of under-classified examinees was affected by the presence of the aberrant anchor item, hence the percent of over-classified examinees was also affected. Table 5 below provides the percent of over-classified examinees for each condition. As can be seen in the Table 5, the percent of examinees that were over-classified was also affected by the presence of the aberrant anchor item. In the case where the b-parameter was shifted by 0.50, the percent of over-classified examinees was about 1% greater when the aberrant item was included in the equating, as opposed to when it was removed. As the degree of aberrancy increased to 0.8, the percent of over-classified examinees was about 2-3% more when the aberrant item was included in the equating as opposed to when it was removed from equating.

Table 5: Proportion of Students Over-classified, by Method, for all Conditions

		Ability Distribution			
		Null		Mean Shift	
		Degree of Aberrancy			
Scaling Method		*0.50*	*0.80*	*0.50*	*0.80*
MM	Aberrant	0.14	0.15	0.14	0.16
	Purified	0.13	0.13	0.14	0.14
MS	Aberrant	0.14	0.16	0.14	0.16
	Purified	0.13	0.13	0.13	0.14
SL	Aberrant	0.14	0.15	0.15	0.16
	Purified	0.13	0.13	0.14	0.14
HB	Aberrant	0.14	0.14	0.15	0.15
	Purified	0.13	0.13	0.14	0.14

When considering the over- and under-classification, Tables 4 and 5 illustrate that in most cases, the percent of under-classified examinees was greater than the percent of over-classified examinees, regardless of equating method or ability distribution shift.

Discussion

This study was designed to investigate the effect of retaining or removing an aberrant anchor item from equating on the classification of examinees into performance categories. The results indicated that removing the aberrant item in the anchor leads to more under-classification than over-classification, but did not lead to any difference in the percentage of examinees that were correctly classified. Therefore, it is up to the test developer to decide whether it would be better to err by having more over-classification or under-classification. However, if there are concerns about the content representativeness of the anchor test, then the aberrant item should be left in for purposes of equating. For certification purposes however, it may be desirable to err on

the side of caution and allow more under-classification than over-classification, thereby taking out the aberrant item will provide a more conservative classification.

The degree of aberrancy was also included as a factor to see how the magnitude of the aberrancy would affect the results. Although there was no effect on the overall classification, greater aberrancy led to more students being over-classified and fewer students being under-classified when the aberrant item was retained in the equating. It should be noted that there were other items that were flagged as being inconsistent, however the particular items were not consistent across replications. As the removed item was the greatest offender this item was chosen for study. Regardless, since there were other questionable items that were not treated, the effect of these items is unknown, and follow-up study regarding these items is warranted.

The ability distribution of the examinees between administrations was also manipulated to investigate what would happen if there were actual growth in the ability of the examinees. The results indicated that there were no differences in the results when the groups of examinees were distributed differently.

The last factor that was manipulated was the scaling method used. All methods were similarly affected and no clear differences were observed between the MM, MS, SL, and HB scaling methods. Hence, all methods appeared to be equally effected by the aberrancy.

While this study is preliminary in nature, it is the first step in determining the effect of aberrant anchor items on the classification of students into performance categories. The results of this study indicate that in this instance, the presence of the item had little impact on the accuracy of classification of examinees, although it did effect the over- or under-classification of examinees. However, the generalizability of these results may be limited and further investigation is warranted.

References

Han, K. (2008). *WINGEN 2* [Computer software]. Amherst, MA: University of Massachusetts.

Hu, H., Rogers, W. T., & Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Applied Psychological Measurement, 32*(4), 311-333.

Huff, K. L., & Hambleton, R. K. (2001). *The detection and exclusion of differentially functioning anchor items* (Research Report 415). Amherst, MA: Laboratory of Psychometric and Evaluation, University of Massachusetts.

Kim, S., & Kolen, M. J. (2004). *STUIRT* [Computer software]. Iowa City, IA: Iowa Testing Programs, The University of Iowa.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking: Methods and practices.* (2nd ed.). New York: Springer.

Michaelides, M. (2006). *Effects of misbehaving common items on aggregate scores and an application of the Mantel-Haenszel statistic in test equating* (CSE Report 688). Los Angeles, CA: Center for the Study of Evaluation, University of California.

PARSCALE [Computer software]. (2003). Lincolnwood, IL: Scientific Software International, Inc.