Comparing Methods to Recalibrate Drifting Items in Computerized Adaptive Testing

James Masters

Pearson VUE

James.Masters@pearson.com

Timothy Muckle

National Board on Certification and Recertification of Nurse Anesthetists

tmuckle@nbcrna.com

Brian Bontempo

Mountain Measurement, Inc.

brian@mountainmeasurement.com

**Introduction**

Computerized Adaptive Testing (CAT) involves the construction of a test dynamically so that the difficulty of each item administered is targeted to the estimated ability of the candidate. Before an item can be administered operationally, item parameters (e.g., item difficulty) must be estimated empirically. Typically, newly developed items, called experimental, pretest, or pilot items are embedded into the operational CAT so that they are randomly administered to examinees amongst the real or operational CAT items. The experimental items are not scored and the response data collected are used to estimate the item parameters through IRT calibration. Once estimated, the item parameters are assigned to their corresponding items, a process referred to as “anchoring.” Operational CAT item pools are composed of these calibrated items, and examinee ability is estimated during the test using the anchored item parameters.

In an operational computer adaptive test, there are situations (such as item overexposure or changes in the curriculum or training, sampling error) when an item’s difficulty (or set of items’ difficulty) has changed significantly from the difficulty calculated when the item was originally pilot tested. This phenomenon, known as item parameter drift, can be the result of both test and non test design factors such as item overexposure or a change in the curriculum or training. A test that suffers from item parameter drift may inaccurately estimate examinees’ ability (Wells, Subkoviak, Serlin, 2002) and must be remedied. Therefore, test makers try to detect items that exhibit drift in item difficulty and reconcile the discrepant estimates.

Monitoring the stability of IRT parameters as a psychometric practice is well-established in the literature (Goldstein, 1983, Bock, Muraki, & Pfeiffenberger, 2005; DeMars, 2004; Donoghue & Isham, 1998; Ito & Sykes, 1994; Sykes & Ito, 1993, Do, Chuah, & Drasgow, 2005). One metric that is used to evaluate the stability of Rasch item difficulty calibrations is the Winsteps (Linacre, 2009) displacement statistic. In any analysis featuring anchored items, Winsteps simultaneously estimates an unanchored item difficulty for each of the anchored items. The displacement statistic is computed simply as the difference between the anchored difficulty value and the estimated difficulty that would be obtained if the item were left unanchored. Readers are referred to Linacre (2009) and Wright & Stone (1999) for technical details on parameter estimation in Winsteps. Typically, large-scale certification programs flag items that display large amounts of drift and begin resolving the problem. Regardless of how drifting items have been identified, rectification of parameter drift (in a 1-parameter CAT environment) can be accomplished in one of two ways:

* Option # 1 (Fresh Pilot): Deactivate (remove from operational use) drifted items and re-administer the items as pilot items. Compute a new item difficulty estimate from the new pilot data.
* Option #2 (Adjusted Operational): Adjust the item difficulty using the Winsteps displacement statistic calculated from the operational data. Essentially, this is the same as using the value that would be obtained by recalibrating the item using the current operational data.

The examinee ability distribution of the sample for these two options varies greatly. The ability distribution of examinees who are administered pilot items (Option #1) is normal. The ability distribution of candidates who are administered operational CAT items (Option #2) is extremely leptokurtic. For items that are drifting (displacing), the mean of this ability distribution may be quite different than the true difficulty of the item. This mis-targeting of the candidates to the item may be such that it is impossible to determine the true difficulty of the item without additional data (Ito & Sykes, 1994). Since the ability distribution of the Option #1 sample is similar to the ability distribution of the sample obtained when the item was initially pilot tested, the more conservative approach is to pilot test the item again and calculate a new item difficulty estimate.

While re-piloting may be the safer route, this process is costly and time consuming. In many cases, re-administration of a drifted item in the pilot stage may not be possible if the pilot test is already filled with newly constructed items. Conversely, it is possible that drifted items placed back in the pilot pool will prevent other newly written items from being piloted in a timely manner. In these situations, it would be ideal if the drifted item could remain in operational use, with the item calibration “adjusted” or updated using the free-estimation value from the operational data (Option #2). However, the advisability of this alternative has not yet been investigated.

The purpose of this study is to determine the extent to which an adjustment to the item difficulty, using the displacement statistic calculated from the operational data, is comparable to a new item difficulty estimate calculated from fresh pilot test data.

**Methods**

The data for this study came from a large-scale, variable-length CAT that is administered to about 100,000 examinees annually.

Operational pools for this examination are rotated periodically. Somewhere near the end of an operational pool’s administration cycle, an item analysis is conducted in order to evaluate the performance of the items. Operational items are evaluated for parameter drift, while pilot items are evaluated to determine whether they meet pre-existing statistical criteria (difficulty, goodness-of-fit, point-to-measure correlation). In order to be flagged for significant displacement, an operational item must exhibit the following conditions on two consecutive operational pools:

* be administered to at least 200 examinees
* have an absolute displacement greater than or equal to 0.5 logits
* have a standardized displacement (absolute displacement divided by the Standard Error of Measurement [SEM] associated with the free item difficulty estimate) greater than or equal to 2.0 logits.

The following analyses focus on 152 operational items that exhibited significant displacement in at least two of three operational pools that were available. As the procedures of this testing program dictate, all of these items were removed from operational use and re-administered as pilot items. These items comprised a portion of a pilot pool administered from June to December, 2008 where they were administered to an average of 450 examinees who were taking the test for the first time. Upon completion of the pilot pool’s administration, an item analysis was conducted, resulting in a new calibration for each of the items.

The analyses which follow compare the “adjusted” operational item difficulty (adding original anchor to the displacement statistic) to the “fresh” item calibration from the re-pilot stage. Two types of analyses were conducted. The first was a correlation and the second was a *t*-test with pooled standard error:



where

 is the adjusted difficulty estimate from the operational data,

 is the difficulty estimate from the pilot data,

 is the standard error associated with the adjusted difficulty estimate from the operational data,

 is the standard error associated with the difficulty estimate from the pilot data,

*n0*is the sample size from the operational data, and

*np* is the sample size from the pilot data.

Using the results of the *t*-tests, each difficulty estimate was placed into three categories:

* No significant difference found between fresh difficulty and adjusted item difficulty
* Fresh difficulty was significantly greater than the adjusted item difficulty
* Adjusted item difficulty was significantly greater than the fresh item difficulty

The results of these categorizations were tabulated and interpreted. Where an item was found to drift on two consecutive pools, two *t*-tests were conducted, comparing fresh pilot calibration to both the Pool 1 and Pool 2 adjusted operational calibrations.

**Results**

Figure 1 is a scatter plot of the various adjusted operational and fresh pilot calibrations plotted against the initial item calibration. This figure illustrates the narrow range of item difficulty for the items in the sample. The fact that there is no linear shape to the plot also reveals the degree to which the items had drifted since their initial calibration.

-------------------

Figure 1 about here

-------------------

Figure 2 shows another scatter plot. This is an identity plot of the adjusted operational calibrations plotted against the fresh pilot calibration. This plot shows how similar the results were for both methods. A correlation matrix for the item calibrations is displayed in Table 1. The correlation between the initial calibration and the various rectification methods was between .4 and .7, as expected, somewhat low. On the other hand, the correlation between the calibrations derived from the various methods exceeded .98 in all cases. Given the large sample, all of the correlations were significantly different than 0. By comparing these correlations, it is clear that the difference between the methods is minimal.

-------------------

Figure 2 about here

-------------------

-------------------

Table 1 about here

-------------------

Despite this high degree of similarity, there were still some items that did not produce similar results across the two methods. Figure 3 is a drop-line chart which displays all of the adjusted operational or fresh pilot calibrations. The calibrations for a single item are displayed along a vertical line with each point representing a comparison calibration. The length of the line relates to the disparity in the calibrations for the item. This plot shows that, in general, the items that had extreme difficulty estimates yielded larger differences between the results from the two methods than the non-extreme items.

-------------------

Figure 3 about here

-------------------

The next series of analyses determined how many of the 152 displaced items showed significant differences between the fresh calibrations and the adjusted operational calibration. In addition, if the difference was statistically significant, the direction of change was calculated (positive [more difficult] or negative [easier]).

When items had shown significant displacement on only one previous pool, the fresh calibration was compared to the original value plus displacement. When items had shown significant displacement on two or more consecutive pools, the weighted average of displacement across the two pools and a pooled standard error were used in order to make the comparison.

Table 2 displays the number of items that showed significant differences and the direction of the differences for two different levels of α.The final column of the table shows the results after a Bonferoni correction, which accounts for the family-wise error associated with 152 *t*-tests.

-------------------

Table 2 about here

-------------------

Using strictly the results of the significance tests, one may conclude that there is not much evidence to support the method of adjusting the item calibration using the displacement statistic. In the most favorable case, 41% of the adjusted operational calibrations were either significantly higher or lower than the pilot calibrations. One may conclude that it is advisable to re-pilot the items that show drift and use the new calibration.

However, as in many studies, tests of statistical significance may not tell the complete story. There is a practical component to the comparisons that must be addressed as well. The majority of the comparisons summarized in Table 1 involved confidence intervals constructed using standard errors which were relatively small. (The mean of the pooled variance was M = 0.09, SD = 0.06.) Accordingly, the differences between the adjusted operational and fresh pilot calibrations that were interpreted as statistically significant may have amounted to little practical significance.

In order to investigate the practical meaning of the difference in the calibrations, a subsequent analysis was conducted evaluating the absolute magnitude of the difference between the adjusted operational and the fresh pilot calibrations. Figure 4 is a histogram of the absolute difference between the values arising from the two recalibration methods. This distribution is skewed with about half of the values being less than 0.27 logits. If we apply the same rules that are used in this certification program to flag items exhibiting parameter drift (the difference being statistically significant and absolute difference greater than 0.5 logits), only 23 of the 152 items (15%) would have been flagged.

-------------------

Figure 4 about here

-------------------

A substantial portion of the 23 items that yielded differences between the two options were very difficult or very easy items. Figure 5 is an X-Y plot comparing the absolute difference between the adjusted operational and fresh pilot calibrations to the original item difficulty estimate. This figure illustrates that larger discrepancies between the two recalibration methods were observed for items which had original calibrations at the extremes of the difficulty continuum. Eight of the aforementioned 23 items had original calibrations less than 0 logits, and five had calibrations greater than 2 logits. Generally speaking, for this exam, there are very few candidates with abilities less than 0 logits or more than 2 logits. Therefore, the error associated with the adjusted calibrations was greatest for items less than 0 logits or more than 2 logits. As can be seen in Figure 5, less than half of the items would be flagged for drift in the “high traffic” range.

-------------------

Figures 5 and 6 about here

-------------------

**Discussion**

This study demonstrated that there was a very strong correlation between a displaced item’s adjusted operational calibration (difficulty plus displacement) and the fresh pilot calibration. On the other hand, through *t*-tests, this study showed that more than 40 percent of the adjusted operational calibrations were significantly different than the fresh pilot calibration. Despite the sizeable number of points that were significantly different, the actual size of these differences was generally small. Moreover, items within the “high-traffic” zone, tended to have less discrepancy than more extreme items. Test makers interested in implementing an adjusted calibration approach may consider using the adjusted calibration for items within the “high traffic” range and using fresh pilot calibrations for items outside the range. In summary, the results provide support for recalibrating drifting items by adjusting the anchored item difficulty estimate with the displacement value obtained from a calibration with operational CAT data.

# References

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (2005). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, *25*, 275-285.

DeMars, C. E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education*, *17*, 265-300.

Do, B. R., Chuah, S. C., & Drasgow, F. (2005). Effects of Range Restriction on Item Parameter Recovery with Multistage Adaptive Tests. Manuscript submitted for publication.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, *22*, 33-51.

Goldstein, H. (1983). Measuring Change in Educational Attainment Over Time: Problems and Possibilities. *Journal of Educational Measurement*, 20, 369-377.

Ito, K., & Sykes, R. C. (1994). *The effect of restricting ability distributions in the estimation of item difficulties: Implications for a CAT implementation*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Linacre, J. M. (2009). Winsteps 3.68.1 [Computer Software]. Chicago: Mesa Press.

Sykes, R. C., & Ito, K. (1993). *Item parameter drift in IRT-based licensure examinations*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, *26*, 77-87.

Wright, B. & Stone, M. (1999). *Measurement Essentials, 2nd Edition*, Wilmington, DE, Wide Range, Inc.

*Figure 1*. Scatter plot of item recalibrations plotted against the initial anchor calibration for the item.

*Figure 2*. Scatter plot of adjusted operational calibrations plotted against the fresh pilot calibration.

*Figure 3*. Drop-line chart displaying the difference between adjusted operational and fresh pilot calibrations.

*Figure 4*. Histogram of absolute difference between adjusted operational and fresh pilot calibrations.

*Figure 5*. Scatter plot of original item calibration vs. absolute difference between adjusted operational and fresh pilot calibrations.

*Figure 6*. Scatter plot of original item calibration vs. standard error of measurement.

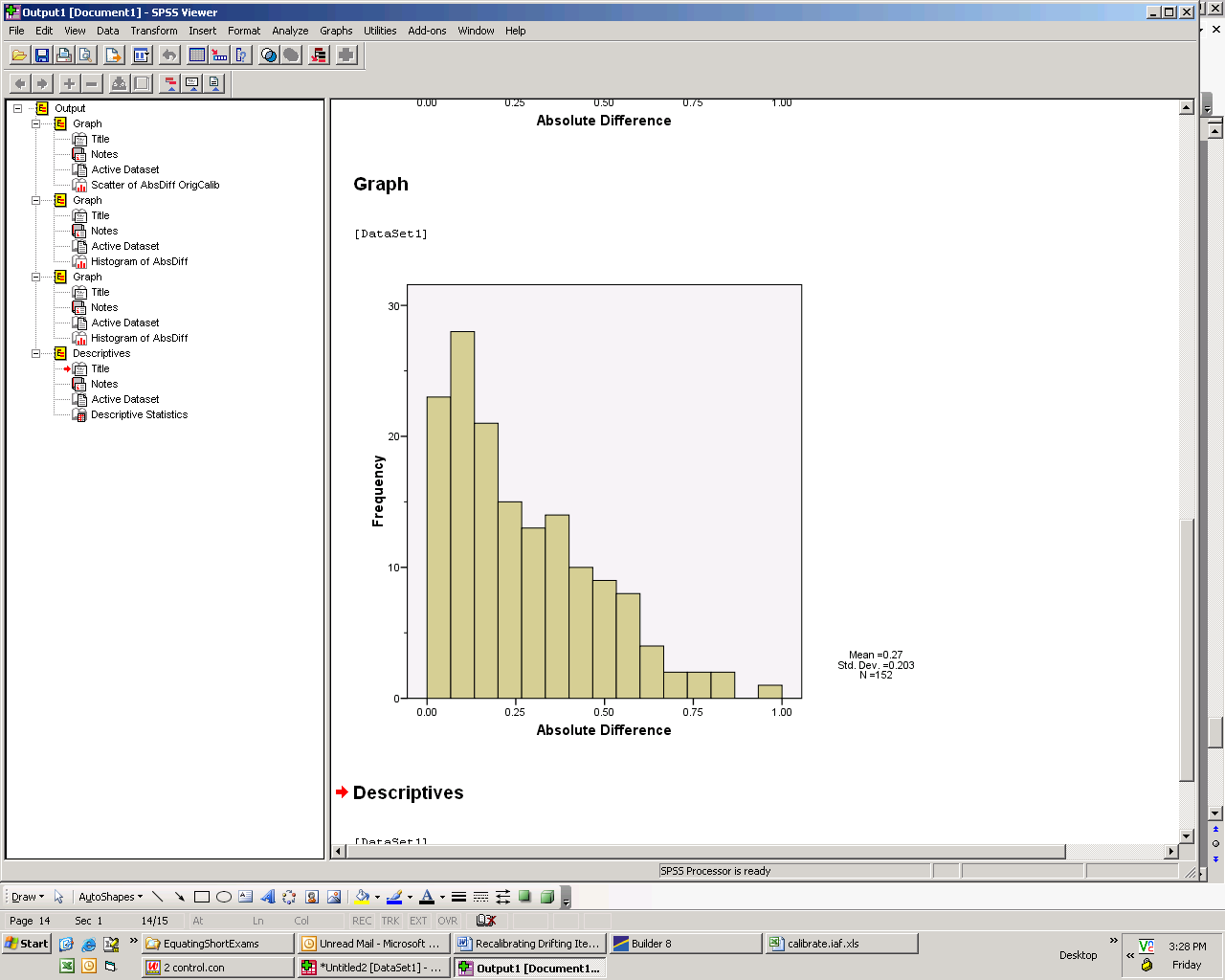
*Figure 1*. Scatter plot of item recalibrations plotted against the initial anchor calibration for the item.

*Figure 2*. Scatter plot of adjusted operational calibrations plotted against the fresh pilot calibration.

*Figure 3*. Drop-line chart displaying the difference between adjusted operational and fresh pilot calibrations.



*Figure 4*. Histogram of absolute difference between adjusted operational and fresh pilot calibrations.



*Figure 5*. Scatter plot of original item calibration vs. absolute difference between adjusted operational and fresh pilot calibrations.



*Figure 6*. Scatter plot of original item calibration vs. standard error of measurement.

Table 1

*Correlations between Item Calibrations at Different Time Points*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Calibration | Original (Pilot) | | Adjusted Op 1 | Adjusted Op 2 | Adjusted Op 3 | ‘Fresh’ Pilot Calibration |
| Original (Pilot) | --- | | .425\*\* | .762\*\* | .404\*\* | .408\*\* |
| Adjusted Op 1 |  | | --- | .993\*\* | .988\*\* | .985\*\* |
| Adjusted Op 2 |  | |  | --- | .995\*\* | .983\*\* |
| Adjusted Op 3 |  | |  |  | --- | .984\*\* |
| ‘Fresh’ Pilot Calibration | |  |  |  |  | --- |

Table 2

*Summary of Comparison of Fresh Pilot and Displacement-Adjusted Calibration, by t-test* α *Level*

|  |  |  |  |
| --- | --- | --- | --- |
| Comparison between  Pilot and Adjusted Calibration |  | *α* Level |  |
| 0.05 | 0.01 | 0.05/152 |
| Positive Difference | 34 | 31 | 22 |
| No Significant Difference | 56 | 68 | 89 |
| Negative Difference | 62 | 53 | 41 |