

Running Head: DETECTING ITEM DISCLOSURE

Finding a Needle in a Haystack: Detecting Item Disclosure in Large-Scale Testing Programs

Jennifer P. Hatfield

Pearson VUE

Paper presented at the annual

American Educational Research Association Meeting

Chicago, Illinois 2007

Finding a Needle in a Haystack: Detecting Item Disclosure in Large-Scale Testing Programs

With the proliferation of and easy access to information via the internet, item disclosure is becoming more of a threat to large-scale high-stakes testing programs. To date some researchers have used item parameter drift detection strategies to identify potentially disclosed items. A variety of methods has been proposed to do this, including time-dependent IRT models (DeMars, 2004; Bock, Muraki, & Pfeiffenberger, 1998), procedures used to detect differential item functioning (Isham & Donoghue, 1994; Smith, 2004), Analysis of Covariance models (Sykes & Ito, 1993), moving averages (Han, 2003), and analysis of item fit statistics (Lu & Hambleton, 2003). These methods can detect items whose characteristics are changing over time, but only if items have been disclosed in a reasonably large proportion of examinees taking the test. However, in a large-scale testing program it is likely that items will be disclosed to only some relatively small subgroup of examinees, in which case it would be very difficult to detect disclosure since the overall drift effect would be quite small. It would thus be beneficial to have methods that could aid in isolating subpopulations in which item disclosure might be occurring.

What could large-scale testing programs do to monitor test security? Various reports are generated during a testing window to monitor test functionality. Test functionality should be expanded beyond traditional item statistics, reliability, validity, and fairness to include exam security. This paper will present an example of the type of analyses that could be conducted to monitor a test's security. Large-scale testing programs have a wealth of data that could be mined for factors indicative of breaches in test security. Just as data can be gathered that indicate the degree to which a test is valid, reliable, and fair, data also can be gathered which might be able to alert test developers to possible breaches in test security.

If there is more cheating in one group of examinees than in another, the following might be expected for the group in which cheating is occurring: item fit statistics might be more aberrant, decreases in item difficulty might be more prominent, item response time distributions might be more negatively skewed, and the performance of repeat test takers might be better. The key to identifying possible item disclosure lies in the simultaneous examination of multiple data indicators. It is not sufficient to suspect a breach in test security if there is a large amount of misfit for one group of examinees relative to another since such misfit could be related to bias issues rather than cheating. However, if items misfit due to bias, item response times would not necessarily be expected to be aberrant, and significant trends in item difficulty across time should not necessarily occur.

Some expectations were set forth a priori regarding characteristics that might be expected of groups of examinees who have access to disclosed items. Specifically, it was hypothesized that, compared with groups of examinees in which items had not been disclosed, groups in which cheating is occurring should exhibit the following characteristics:

- more marked increase in pass rates across time
- larger score gains between repeat exams
- more item and person mis-fit
- more variability in item response time
- less score stability across content area subsections of exam

Furthermore, it was hypothesized that items seen near the beginning of a testing period by examinees in groups where items have been disclosed should show more drift than other items.

Method

Five cohorts of examinees were established based on the quarter in which they first took the NCLEX-RN[®] examination: Cohorts from the following testing periods were used: (a) January - March 2005, (b) April - June 2005, (c) July - September 2005, (d) October - December 2005, and (e) January - March 2006. All subsequent exam records were found for each member of the cohort. The data of interest to this study were examinees' final Rasch ability estimates, a person-fit statistic, test center in which the exam was taken, and item response data, including item response times. examinees were placed into one of 12 geographical regions based on the location of the test center in which the exam was taken.

The objective of this project was to demonstrate the types of data and analyses that could be used to monitor test security. Even with current desktop computing technology, because of the amount of data involved, such analyses take a considerable amount of time to set up and run. Therefore, in some instances analyses were run for only a subset of the test center groups. The following groups of analyses were conducted:

Analysis of repeat test takers

For examinees in the cohort who failed their first exam and went on to retake the exam, performance differences between their first and second exams were graphically analyzed. The following variables were used as performance indicators for each examinee: (a) Rasch ability estimates, (b) person-fit statistics, (c) mean item response time, (d) item response time variability, and (e) variability of content-area ability estimates.

Screening of items

Moving average p -value analyses were conducted in order to screen items for trends indicative of possible item disclosure. These analyses were conducted for a subset of the test

center regions. Regression lines were fit to the moving average data, and the distributions of regression slope values were compared between test center groups. It might be expected that the distribution of slopes would be more positively skewed for test center regions in which there has been a breach in test security.

Item fit within each of the test center regions was also examined. The statistic for assessing model-data fit was a standardized residual for item i as follows:

$$Z_{ij} = \frac{N_j^{1/2} [P_{+ij} - E(P_{+ij})]}{[E(P_{+ij})(1 - E(P_{+ij}))]^{1/2}}$$

where,

$$P_{+ij} = \frac{1}{N_j} \sum_{g \in j}^{N_j} u_{ig} = \text{observed proportion correct for the } g \text{ examinees in group } j, \text{ and}$$

$$E(P_{+ij}) = \frac{1}{N_j} \sum_{g \in j}^{N_j} P_i(\hat{\theta}_g) = \text{the expected proportion of } g \text{ examinees in group } j \text{ correctly answering}$$

item i as predicted by the Rasch model.

Although the standardized residual can be interpreted according to standard normal distributional theory, as with traditional model-data fit statistics, the power to reject the null hypothesis increases with increased sample sizes. This can become especially problematic with CAT data, as the sample sizes for different items will vary widely. One way this effect might be adjusted for is to rescale the standardized residual by the ratio of the sample size used to calculate the statistic to some desired “reference” sample size:

$$Zadj_{ij} = Z_{ij} \left(\frac{N_{REF}}{N_j} \right)^{1/2}$$

These adjusted Z statistics were used in the analyses. These statistics are also used regularly to screen NCLEX items.

Person fit was assessed with a statistic proposed by Wright and Panchapakesan (1969), which is given by:

$$W2 = \sum_{i=1}^n [u_i - P_i(\theta)]^2 / \sum_{i=1}^n P_i(\theta)[1 - P_i(\theta)]$$

where,

i indexes items, $i = 1, 2, \dots, n$,

u is the item response (0 or 1),

P is the model predicted probability of correct response.

Results

Pass rates broken out by test center region, across the last eight testing windows can be found in Figure 1. All regions, aside from 8, 9, and 11 exhibit a cyclical pattern where pass rates are similar for January, April, and July testing periods but drop in October periods. None of the test center regions exhibits a steady increase in pass rates that might be expected if breaches in test security were occurring.

Analysis of Repeat Test Takers

Figures 2 – 6 present results of the cohort analysis. Figure 2 shows that, across all cohorts, region 9 is a clear outlier in terms of gains in ability estimates between the first and second exam. All cohorts had minor average changes in person fit between the first and second exam (cf. Figure 3). Region 9 stands out as exhibiting consistently higher person fit statistics in the second exam and region 10 exhibits the most fluctuation across cohorts in person fit differences.

If examinees have access to disclosed items it might be expected that their content area ability estimates would be less consistent. Items on which examinees cheated might affect content area ability estimates to a greater degree than the overall ability estimate since content area estimates are based upon fewer items and thus the effect of any one particular item on ability estimation is greater. The difference between each of eight content area subtest ability estimates was computed and the standard deviation of these differences was used as an index of content area ability estimate consistency. It can be seen in Figure 4 that region 9 exhibited consistently higher variability in difference scores than the other regions.

Figure 5 shows that, in terms of item response times, most cohorts took longer, on average, to respond to items in the second sitting of the exam than in the first. Region 9 had the least difference in item response times between the first and second exam, with several cohorts responding somewhat faster to items on the second exam than the first. In general, there was more variability in item response time on the second exam (cf. Figure 6). Response time variability was more similar for April and July cohorts than for the others. It might be expected that there would be more item response time variability for groups in which examinees have access to disclosed items since item response times to disclosed items should be aberrantly faster than an examinee's typical item response time. Though there is a general bias toward more response time variability on the second than the first exam, this bias is likely an artifact of less range restriction in response time for the second exam rather than any breaches in test security. None of the regions exhibited markedly larger positive differences between exam 1 and 2 response time variability that might be suggestive of item disclosure.

Screening of Items

If an item has been disclosed it should exhibit poorer model fit since examinees who answer the item correctly because they had prior access to the item are not responding in accordance to the model. Item fit statistics were analyzed for a six-month testing period for four test center regions with comparable testing volumes. Table 1 presents frequencies of item fit statistics by region. From these tables it can be seen that there clearly is a larger degree of item misfit for region 9 than the others.

However, since there is some degree of symmetry to misfit, these statistics alone likely cannot indicate possible disclosure. For groups in which items have been disclosed there should be more instance of positive fit statistics since these indicate that examinees are more likely to answer the item correctly than the model predicts. However, this likely will be counterbalanced by instances where examinees are less likely to get the item correct than the model predicts because examinee ability estimates in the group will be less precise due to the introduction of a cheating factor. Thus, characteristics of items with extreme positive and negative fit statistics are in order.

If items are being disclosed by examinees who took the test early in a testing period and memorize and disseminate items, item fit should be better near the beginning of a testing period than at later points after more examinees have had access to disclosed items. The items with positive fit statistics should be more likely to have been exposed to examinees early in the testing period than the items with the negative fit statistics, since these statistics are merely an artifact of examinees later in the testing period utilizing knowledge of disclosed items. Figure 7 presents the percentage of administrations within the first month of the testing period for each Z statistic

category. Items with extreme positive fit statistics were not more likely than items with extreme negative fit statistics to be seen by examinees early in the testing period.

Items that become easier over the course of a testing period could be suggestive of disclosure. In order to examine this possibility, moving average p -values were computed for each item that had over 95 responses. In order to summarize results of the moving average analyses, a trend line slope was calculated from the moving average data. It might be expected that there would be more positive slope values for groups of examinees in which items have been disclosed. Figure 8 presents histograms of the slope values for each of the four regions examined.

There was no evidence from the item fit statistic analysis that items administered near the beginning of the testing cycle were more likely to exhibit higher than expected p -values. However, these fit statistics do not take time into account. An item that becomes easier over time due to having been disclosed might not necessarily have a large fit statistic. Therefore, analyses were conducted to determine if there is any relationship between moving average p -value trends and item exposure rates in the first month of the testing period. Scatterplots of trend line slope values versus percentage of administrations within the first month can be found in Figure 9. From these plots, it is evident that items with more of their exposure near the beginning of the testing cycle are not more likely to exhibit increase in p -value than other items.

Discussion

The analyses presented here are only a handful of possible analyses that could be conducted to monitor test security. With time and further exploration, better potential indicators of test security could be found.

The indicators used in this paper suffer a number of shortcomings. First of all, the cohort analyses rely heavily on difference statistics but such statistics have been criticized as being too unreliable. It likely is oversimplistic to restrict the moving average trend to be linear. However, monotonic increasing trends are what is of interest in such an analysis. A monotonic trend will still be fit with a line of positive slope though it might be more precisely defined by a quadratic function. Item and person fit statistics have been shown to be far from perfect and subject to inflated type I error rates or poor power in certain circumstances.

Though the measures used in this study are far from perfect, their utility could perhaps lie in the form of multiple indicators. If a number of statistics are used to monitor test security, evidence for possible breaches to test security could be found among patterns of results, for example certain changes in statistics across repeated assessments or certain groups of examinees that exhibit anomalies in a number of the indicators examined. Large breaches in test security likely would be captured by the methods used in this paper.

A drawback of the approach taken in this paper is that groups must be specified a priori to examine for possible breaches in test security. If there are very large breaches in test security, analyses like those above conducted on the entire group of examinees might indicate the breach. However, if breaches in test security are moderate to small and not localized in a particular group that is analyzed, detecting disclosure would be difficult, if not impossible, using the methods outlined above. Because of the internet, and easy dissemination of information globally, non-localized breaches in test security are more likely than ever. Thus, the development of search strategies to detect subgroups of examinees in which items have been disclosed is in order.

Even if breaches in test security are localized to known groups, the difficulty still lies in how to identify small breaches; instances in which only a handful of items become disclosed to

examinees. Finding statistical methods that would have enough power to identify such instances seems an insurmountable task. Perhaps another direction that a testing program first needs to embark upon is an analysis of the practical implications of item disclosure. Small amounts of item disclosure, though undesirable in principle, might have a negligible effect upon test scores and leave the validity of the cut score criterion intact.

References

- Bock, R.D., Muraki, E., & Pfeifferberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275-285.
- DeMars, C. (2004, April). *Item parameter drift: The impact of curricular area*. Paper presented at the Annual Meeting of the American Educational Research Association, Sam Diego, CA.
- Han, N. (2003). *Using moving averages to assess test and item security in computer based testing* (Research Report No. 468). Amherst, MA: University of Massachusetts, School of Education, Center for Educational Assessment.
- Isham, S.P. & Donoghue, J.R. (1994, April). *A comparison of procedures to detect item parameter drift*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.
- Lu, Y & Hambleton, R.K. (2003). *Statistics for detecting disclosed items in a CAT environment* (Research Report No. 498). Amherst, MA: University of Massachusetts, School of Education, Center for Educational Assessment.
- Smith, R.W. (2004, April). *The Impact of braindump sites on item exposure and item parameter drift*. Paper presented at the Annual Meeting of the American Educational Research Association, Sam Diego, CA.
- Sykes, R. & Ito, K. (1993, April). *Item parameter drift in IRT-based licensure examinations*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.
- Wright, B. & Panchapakesan, N. (1969). A procedure for sample free item analysis. *Educational and Psychological Measurement, 29*, 23-48.

Table 1. Frequency Distributions of Item Fit Statistics by Region

Z Statistic	Region 4		Region 8		Region 9		Region 11	
	N	%	N	%	N	%	N	%
<= -4	94	5.1%	30	2.1%	265	14.6%	36	3.1%
-3 to -4	89	4.8%	70	4.9%	100	5.5%	38	3.3%
-2 to -3	158	8.5%	139	9.8%	138	7.6%	108	9.4%
-1 to -2	232	12.5%	207	14.6%	171	9.4%	166	14.5%
0 to -1	318	17.1%	263	18.5%	195	10.7%	203	17.7%
0 to 1	325	17.5%	270	19.0%	211	11.6%	247	21.5%
1 to 2	285	15.3%	222	15.6%	189	10.4%	179	15.6%
2 to 3	180	9.7%	125	8.8%	173	9.5%	93	8.1%
3 to 4	92	5.0%	59	4.2%	136	7.5%	53	4.6%
> 4	84	4.5%	35	2.5%	242	13.3%	24	2.1%

Figure 1. Pass rates for four-month testing windows from January 2005 to October 2006 broken out by test center region

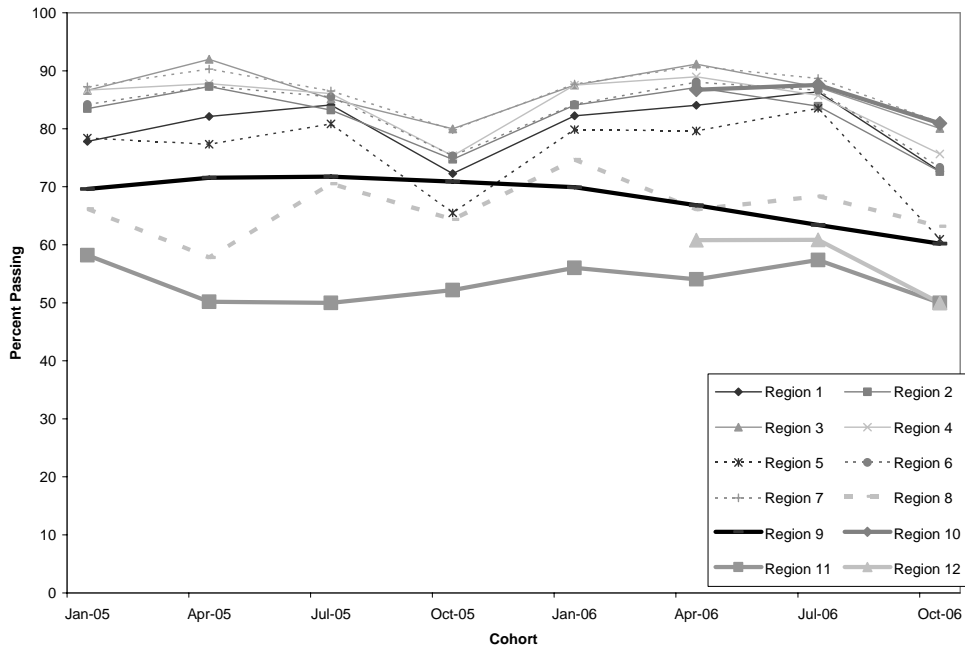


Figure 2. Difference in Rasch ability estimates between first and second exams for each of five cohorts broken out by test center region

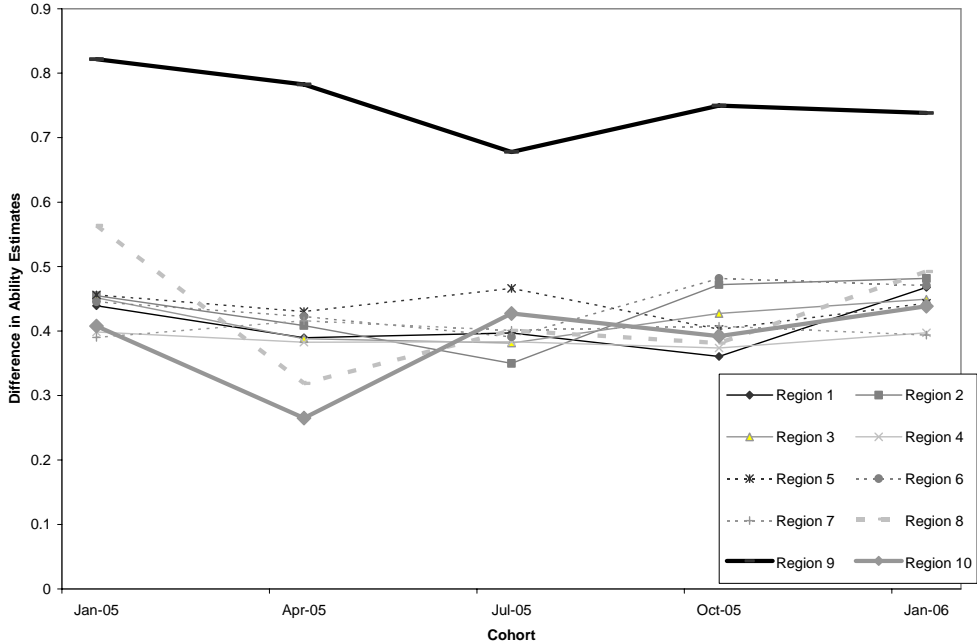


Figure 3. Difference in person fit statistics between first and second exams for each of five cohorts broken out by test center region

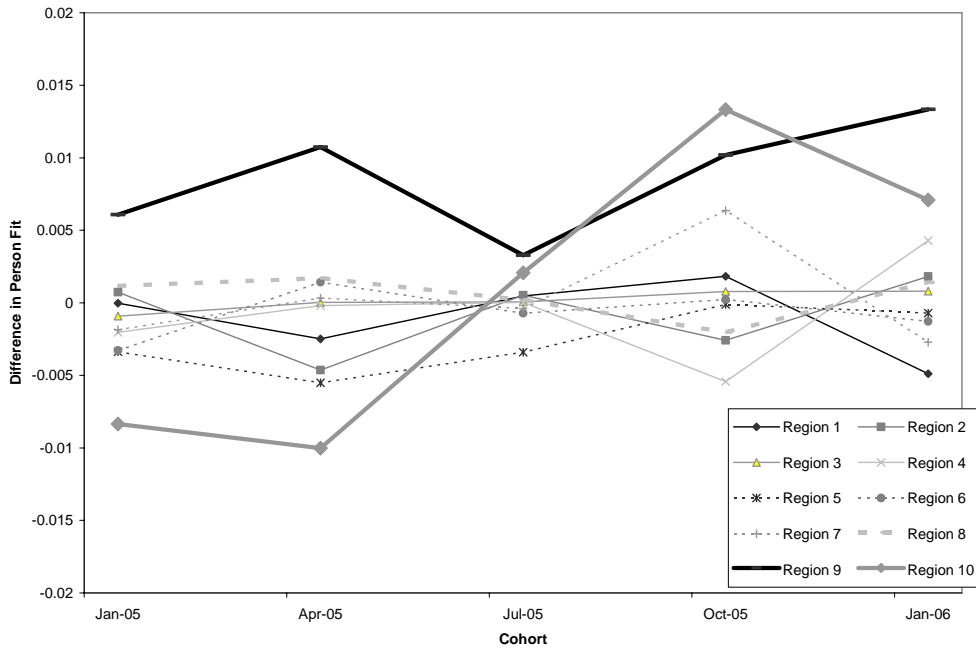


Figure 4. Difference in the standard deviation of subtest ability estimates between first and second exams for each of five cohorts broken out by test center region

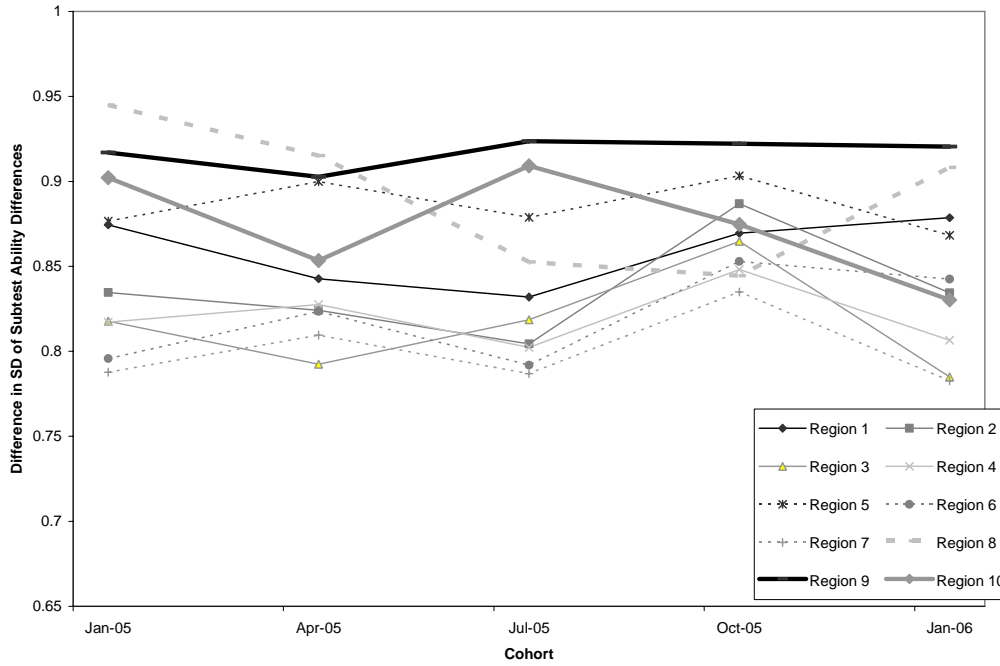


Figure 5. Difference in average item response times between first and second exams for each of five cohorts broken out by test center region

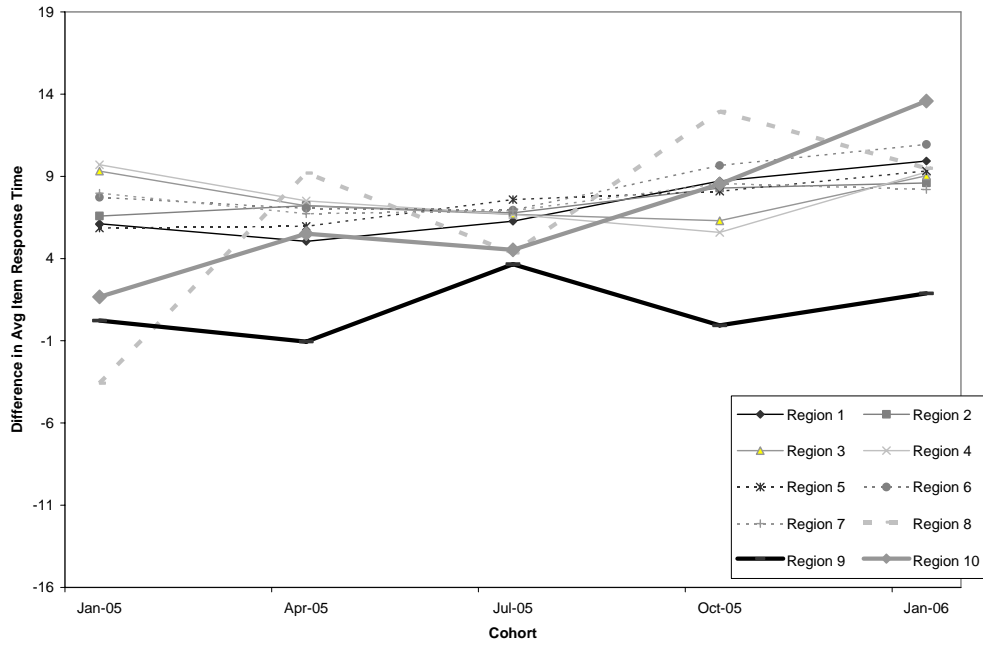


Figure 6. Difference in item response variability (as standard deviation) between first and second exams for each of five cohorts broken out by test center region

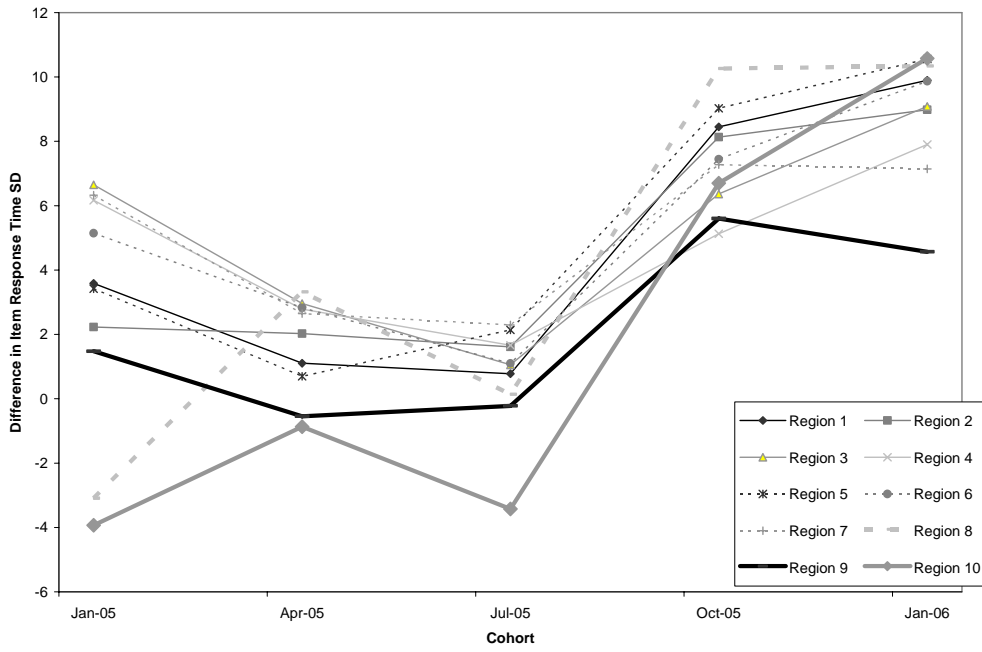


Figure 7. The relationship between item fit Z statistics and early item exposure for four test center regions

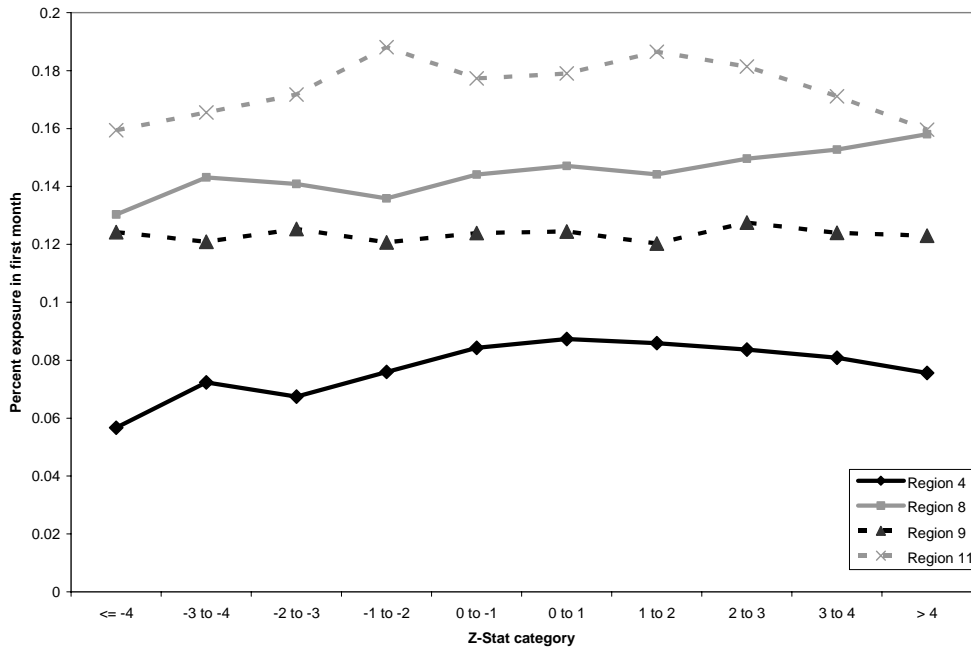
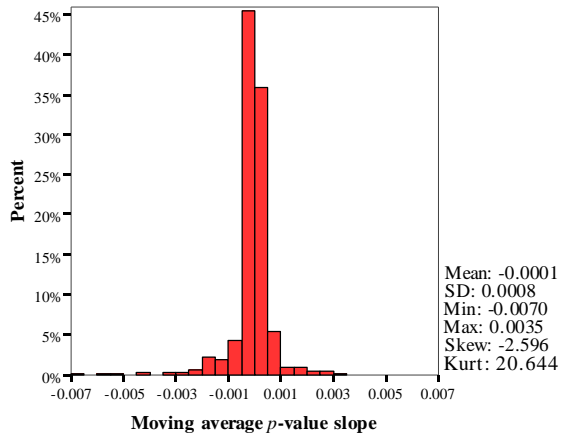
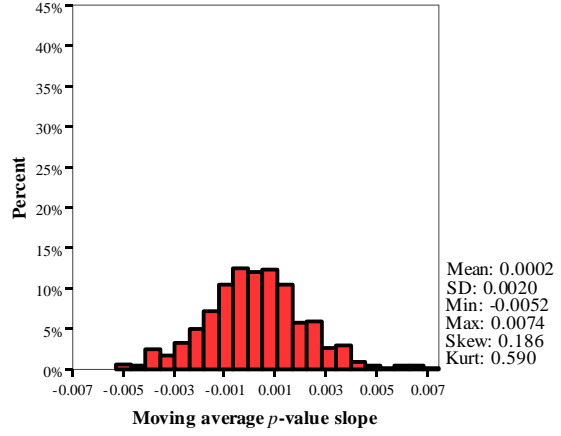


Figure 8. Histograms of trend line slopes from moving average p -value analysis for (a) region 4, (b) region 8, (c) region 9, and (d) region 11

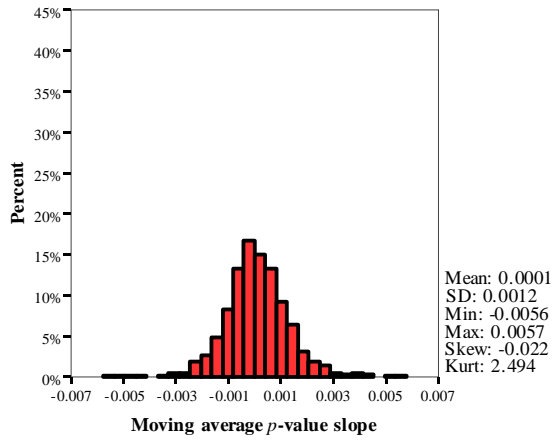
a



b



c



d

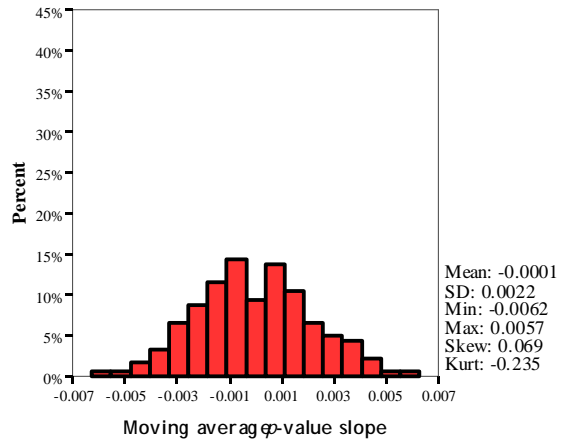
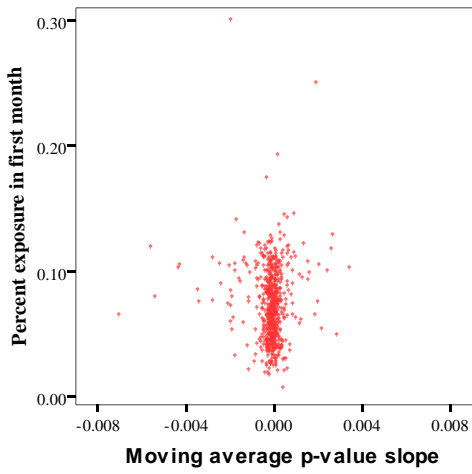
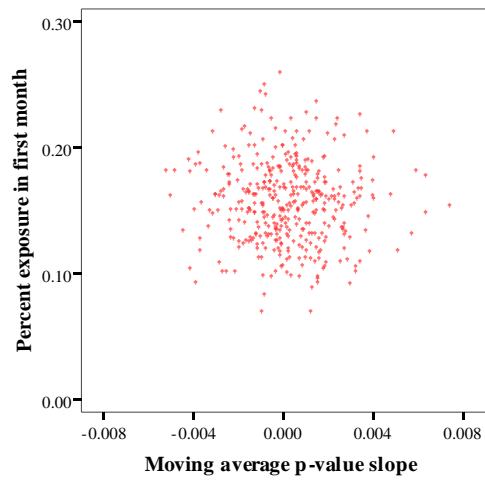


Figure 9. Scatterplots of trend line slope values versus percentage of administrations within the first month for a). region 4, b). region 8, c). region 9, and d). region 11

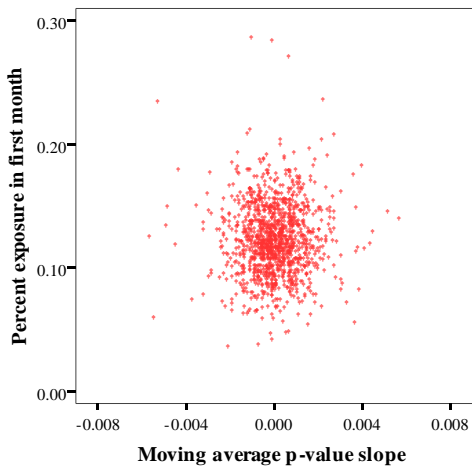
a



b



c



d

