

An Investigation of Item Response Time Distributions as Indicators of Compromised NCLEX Item Pools¹

Steven L. Wise, James Madison University

G. Gage Kingsbury, Northwest Evaluation Association

An important problem faced by high-stakes CAT licensure programs, such as the National Council Licensure Examination (NCLEX), is test item pool security. It is challenging to maintain security, however, because items are repeatedly being exposed to examinees who might remember items they received. Such memories can range in detail from a vague idea of the point being assessed by a particular item to a relatively complete recollection of the item's stem and options. To the extent that examinees can accurately recall information after testing about items they were administered, they possess important information that might then be passed on to other examinees. An item is said to be *compromised* to the extent that advance knowledge regarding the item's content is available to examinees prior to testing.

For examinees with advance knowledge of an item pool, test score validity will be degraded. Compromised items will tend to be less difficult for those with advance knowledge, which can lead to positively biased proficiency estimates, with some unqualified examinees attaining passing scores. Moreover, the realization that some examinees have advance knowledge of items represents a serious threat to the integrity of the testing program, as those with advance knowledge have an unfair advantage over those that do not. This inequitable situation can have profound public relations consequences. Hence, test givers are highly motivated to maintain test score validity by controlling item exposure and potential compromise.

In contrast, many—if not most—examinees taking a licensure exam are less interested in attaining a valid test score than they are in receiving a license to working in their chosen profession. That is, they view the exam as an obstacle to overcome in their efforts to become licensed. From this perspective, examinees are motivated to acquire as much advance knowledge as they can about the test items. This has led to organized efforts by test preparation organizations to acquire and provide (often for a price) advance knowledge of items in a CAT pool.

High-stakes CAT programs typically deal with this problem by frequently changing item pools, which minimizes the exposure of individual items. This strategy, however, greatly

¹ Paper presented at the 2006 annual meeting of the American Educational Research Association, Chicago. This research was supported through funding by the National Council of State Boards of Nursing. Correspondence regarding this paper should be addressed to Steven L. Wise, Institute for Computer-Based Assessment, Center for Assessment and Research Studies, James Madison University, MSC 6806, Harrisonburg, VA 22807. E-mail: wisesl@jmu.edu.

increases the resources required to maintain score validity, as many new items need to be developed. Moreover, it is not easy for the test givers to assess the degree to which a given item pool has been compromised and thereby judge whether or not changing an item pool is warranted. If the pool is changed too frequently, an excessive amount of resources will be devoted to item development. In contrast, if the pool is not changed frequently enough, test score validity will be threatened.

Identifying item compromise can be challenging. If a test giver is lucky, there is tangible evidence (e.g., written materials, web sites, etc.) that the content of operational items has been disseminated. Typically, however, test givers have to rely on the behaviors of examinees during the test to detect item compromise. For example, one might expect the proportion of examinees passing a compromised item to exceed what would be expected by the item's behavior during pilot testing, and item fit statistics can be used to identify items whose difficulties have drifted due to item compromise.

An additional examinee behavior that may be related to item compromise, but does not appear to have been previously studied, is item response time. The purpose of the present study was to investigate the usefulness of response time as an indicator of item compromise for the NCLEX.

A challenge of using item response time is that it can have multiple influences—some of which are relevant to the construct being measured and some that are not. When an examinee encounters a multiple-choice test item on a high-stakes exam, for example, the time it takes the examinee to select an answer is influenced by a number of factors (Wise, Bhola, & Yang, 2006). These factors can be generally classified as characteristics of the item, characteristics of the context in which the item is given, and characteristics of the examinee. Items can vary in difficulty, but they also vary in the amount of reading required, and how mentally taxing the item is (e.g., how many steps it takes to complete a mathematics item). In addition, contextual factors such as item position, how free the test administration setting is from distractions or the consequences associated with test performance can influence item response time. Finally, while examinees vary in their levels of proficiency, they also vary on such variables as reading ability, cognitive processing speed, reaction time, fatigue, and test anxiety—all of which can influence item response time. Taken together, these multiple influences make item response time an inherently complex variable for measurement practitioners to use, because of the difficulty of unequivocally interpreting the meaning of a given item response time value.

What is the impact of advance knowledge on response time? Our working assumption in this study is that, other factors being equal, an examinee with advance knowledge of an item will tend to respond more quickly than one without advance knowledge. If an examinee has detailed knowledge of an item's content and correct answer, then the examinee would merely need to recognize the item and recall the correct answer—which should require substantially less time than that needed to read the item, understand its challenge, and do the mental processing needed to identify the correct answer. Even if the examinee had advance knowledge that was less than complete regarding an item, he or she may have enough information to direct test preparation toward the challenge posed

by the item. This directed test preparation should also tend to reduce response time to the item.

When a test item is pilot tested, because it has not been used in previous test administrations, advance knowledge should be minimal. That is, the distribution of response times across examinees for a secure pilot-tested item should be free of advance knowledge effects. If the item makes it into the operational item pool, and is exposed to an increasing number of examinees, then the numbers of examinees who have advance knowledge of that item should also increase. Examinees with advance knowledge will require less time to respond to the item, which suggests that over time the distribution of response times for an item should shift downward relative to its response time distribution during pilot testing.

Thus, increasing amounts of advance knowledge regarding an item should be reflected by systematic changes in the response time distributions. The current study was intended to explore for evidence of this effect in NCLEX data.

Method

To address the relationship between advance knowledge and response time, the current study used operational data from the NCLEX-RN, a computerized adaptive test that is used as a portion of the process for licensing registered nurses. The study used data collected from the test over the course of several years to investigate response time for items that were flagged as having drifting or poorly fitting measurement characteristics in a term of use after being used operationally for several years with no problems. To the extent that these items also had unusual response times, it may indicate that the item has been compromised.

The NCLEX-RN

The NCLEX-RN is a test that a candidate must pass as part of the process of becoming a licensed registered nurse in the United States. The test is administered on a computer in a standardized, proctored testing center. The test adapts to the performance of each candidate by selecting items from a large item pool by matching the difficulty of the item to the candidate's performance. All items in the item pool are calibrated using the Rasch (one-parameter logistic) IRT model (Wright, 1977). The test uses an adaptive mastery testing procedure (Kingsbury & Weiss, 1983) and ends when a pass-fail decision can be made with a high degree of confidence, or when a maximum test length is reached.

Individuals who don't pass the test face real-world consequences ranging from retaking the test to losing a career in the nursing field. The high-stakes nature of the test means that item compromise is a real and constant threat. All reasonable precautions are taken by the testing agencies, but item exposure that leads to cheating remains a concern. While this makes life difficult for the practitioners, it makes a very nice data set for us researchers. During each term, data from the items is analyzed to determine which items are performing oddly, given their item difficulty. Items with high misfit are identified

and flagged for further analysis or flagged for removal from the item pools. Items flagged for removal due to misfit form the basis of this study, as detailed below.

For security purposes, multiple item pools are rotated into use periodically. A particular item may be used for a time, be taken out of use for a time and then be returned to use at a later date. This means that over time, several opportunities to observe the characteristics of an item are available. This study capitalizes on this feature of the test to observe item response times across several years.

Data Used

The data in this study came from 10 successive administrations of the NCLEX-RN from October, 2000 to January, 2005. For each time period information available included the following:

- The overall score for each candidate (including passing status)
- The items taken by each candidate
- The correct/incorrect status of each item response
- The response time for each item response (the seconds elapsed between item presentation and the confirmed item response)

No information about item content was available. No information about the order in which items were taken was available.

Table 1. Item Pool Size and Number of Administrations

Administration Time Period	Number of Items in Pool	Total Number of NCLEX-RN Administrations
October, 2000	1,653	16,216
January, 2001	1,653	23,025
October, 2001	1,653	16,664
January, 2002	1,653	22,651
October, 2002	1,641	14,895
January, 2003	1,641	26,086
October, 2003	2,000	19,329
January, 2004	2,000	31,891
October, 2004	2,000	22,546
January, 2005	2,000	32,244

Table 1 shows the number of items in each operational item pool and the number of test takers included in the data set for each test term. The number of items in the item pools ranged from 1641 to 2000, while the number of candidates tested ranged from 14,895 to 31,891. Each candidate took between 60 and 250 operational items, with the most common test length being 60 items. If items were administered randomly, this would result in 500 to 600 responses per item in the smallest administration. Since the test is adaptive, the actual item usage will vary substantially.

Target items. The items targeted for investigation in this study were 47 items that were identified during January, 2005 as having abnormally high misfit values in two consecutive testing terms using the Outfit statistic (Wright and Stone, 1979). While these items were not specifically identified as being compromised, it is likely that compromise will cause items to display misfit as the compromise becomes more common. A set of comparison items that didn't have the same misfit characteristics was also identified by selecting an arbitrary sample of non-target items from the January, 2005 item pool.

Analysis

The basic premise of the study is that item compromise should be identifiable by the behavior of candidates taking the compromised items. The current approach taken by the testing agencies uses an omnibus fit statistic to identify any type of misfit in the operational items. While this approach is quite useful for maintaining the quality of the item pool, specific statistical tests may be able to identify certain types of misfit (caused by item compromise) more quickly than the omnibus statistic. The analysis in this study tried to identify whether such an approach might be effective. Toward this end cross-sectional and longitudinal analyses were performed using visual analysis and statistical tests where appropriate.

Cross-sectional analysis. The cross-sectional analysis looked at the characteristics of the target items in the January, 2005 data set. The analysis investigated the following questions:

- Were the proportion correct values of the target items similar to those in the entire item pool?
- Were response latencies of the target items similar to those in the entire item pool?
- Were there differences in response latency for correct and incorrect responses?
- Were correct/incorrect differences in response latency similar for the target items and their comparison items?

Longitudinal analysis. The longitudinal analysis concerned changes in the characteristics of target items that may have occurred over time from October, 2000 to January, 2005. This analysis investigated the following questions:

- Did the proportion correct values of the target items remain the same across time or show systematic patterns of drift?

- Did the response latency values of the target items remain the same across time or show systematic patterns of drift?

Results

Cross-Sectional Analysis

Figure 1 shows the distribution of average proportion correct values for each item in the January, 2005 item pool as it related to the number of candidates who saw the item. As would be expected for an adaptive test, most items tended towards 50 percent correct, with the tendency increasing as the number of individuals responding to the items increased.

Figure 2 shows the same information as Figure 1, limited to the target items. It can be seen that the distribution of average proportion correct values for the target items differed substantially from the distribution for the full pool. None of the target items had an average proportion between .4 and .6. In addition, the distribution was bifurcated into a group of items that was relatively too difficult and a larger group that was relatively too easy. It seems possible that different mechanisms explain the misfit observed in these two groups of items.

**Figure 1. Average Proportion Correct for each Item
by Number of Candidates Responding**

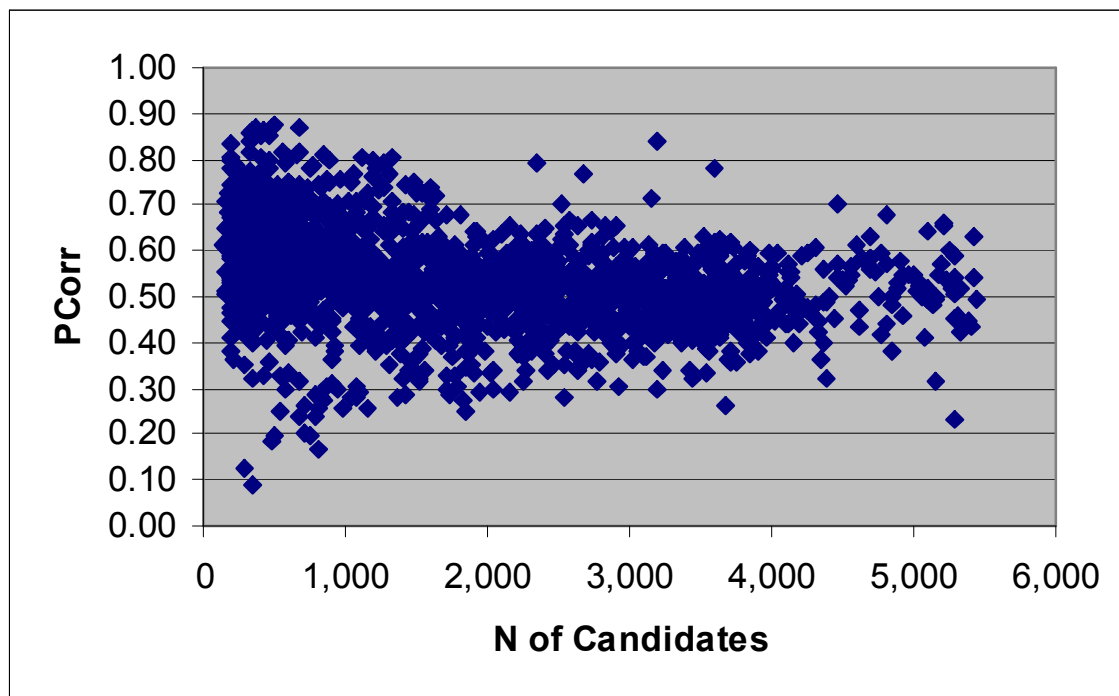


Figure 2. Average Proportion Correct for each Target Item

by Number of Candidates Responding

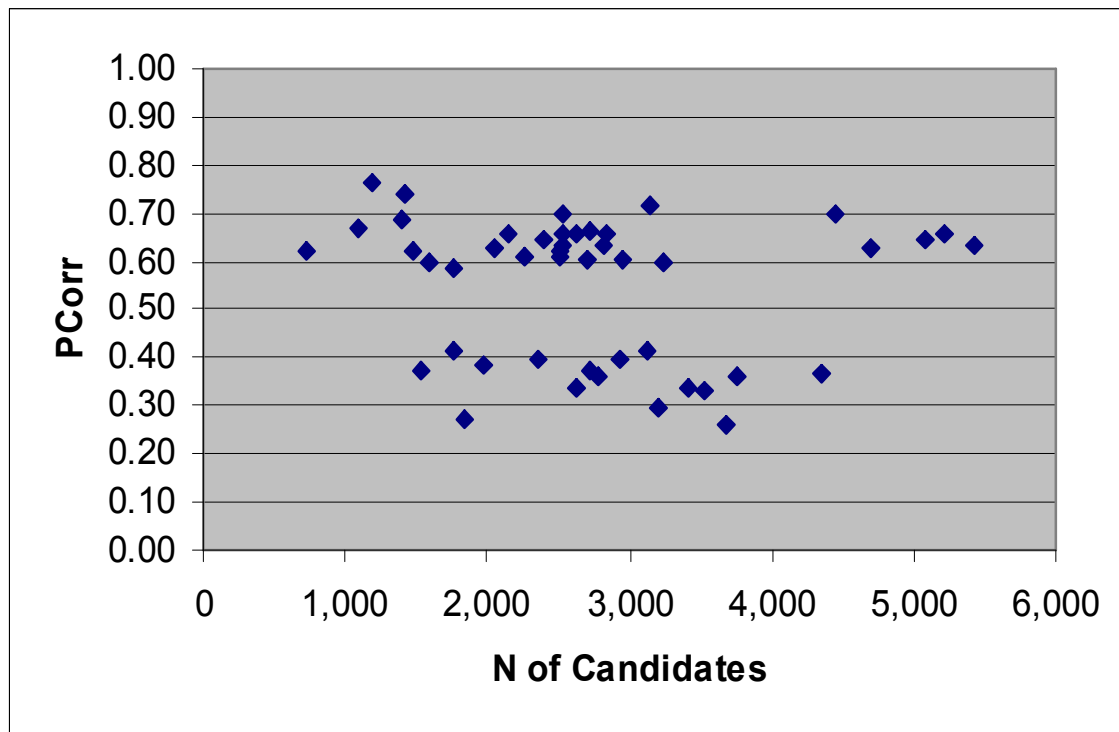


Figure 3 shows the average response time for each item in the full January, 2005 item pool as it related to the number of candidates seeing the item. The average response times varied substantially across items, which may be due to a number of factors including item length and complexity. The items that were administered to the most students tended to have slightly lower average response times. The average response time across all items was 72.29 seconds, with a standard deviation of 26.22 seconds.

Figure 4 shows the same information as Figure 3, limited to the target items. In general, the target items had a distribution of average response times somewhat similar to that of the full item pool, but with a lower average response time (65.82 seconds) and a lower standard deviation (17.08).

Since the particular kind of misfit of interest in this study was item compromise, it was reasonable to assume that the items in Figure 2 that form the “too easy” group might be items that had been compromised. (If compromised items become more difficult than expected, the cheaters really need to consider a different line of work.) It also was reasonable to assume that the individuals who had inappropriate access to an item probably didn’t have to work it out when they saw it on a test. This may result in shorter response times for correct responses to exposed items. To identify whether these assumptions have credibility, the next phase of the analysis concentrated on the 31 target items that form the “too easy” group.

Figure 3. Average Response Time for each Item by Number of Candidates Responding

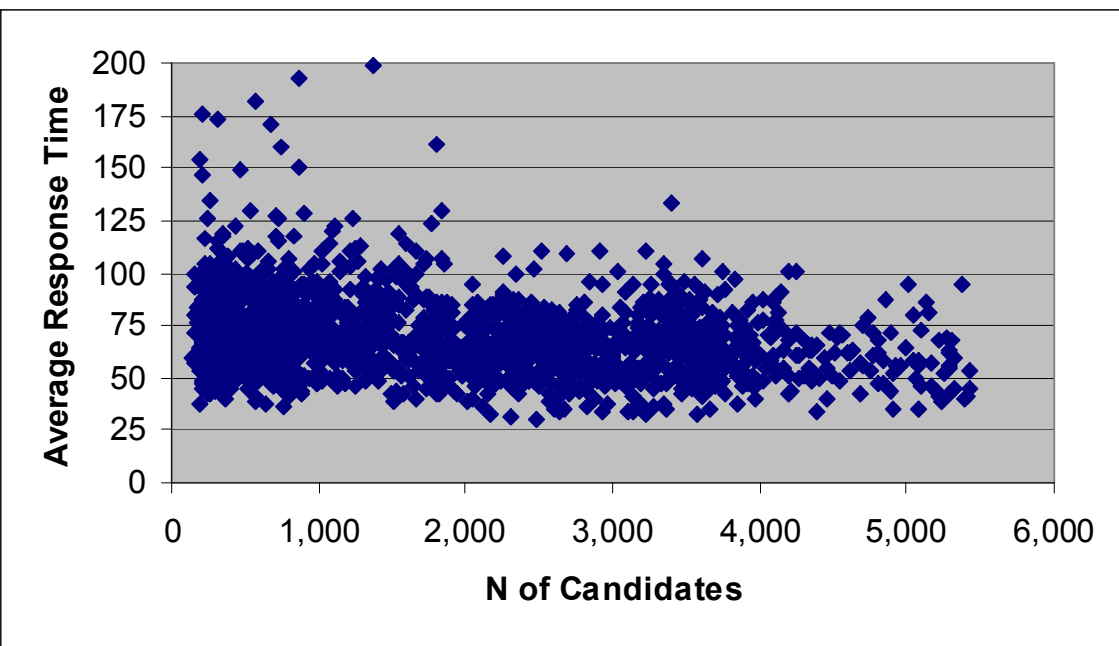


Figure 4. Average Response Time for each Target Item by Number of Candidates Responding

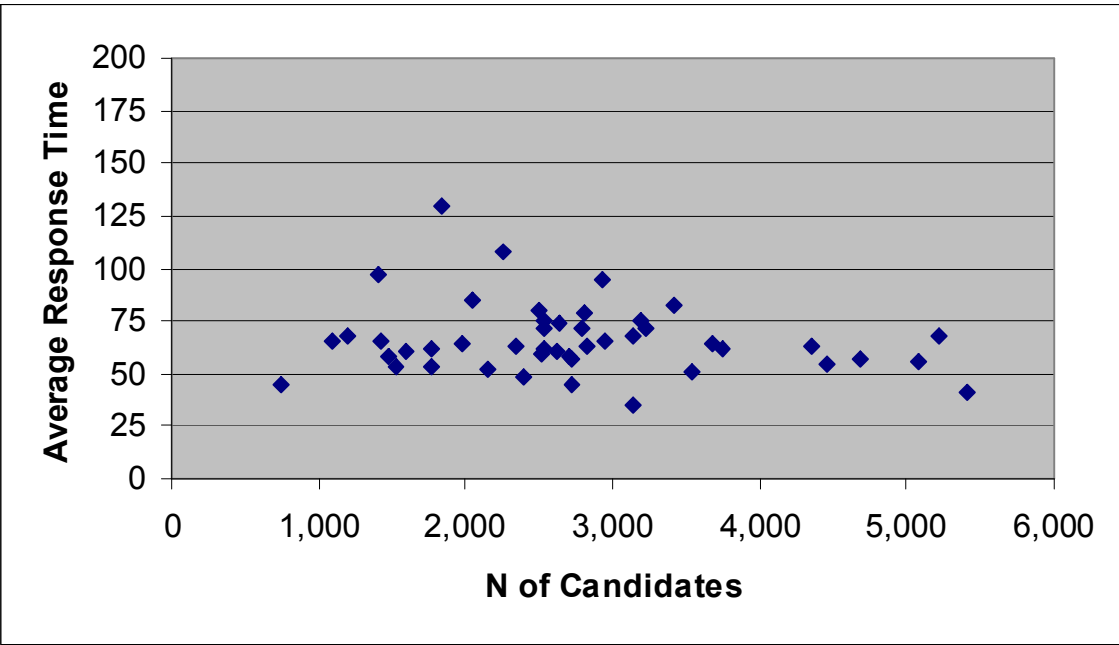


Figure 5. Average response time for correct and incorrect responses to each target item

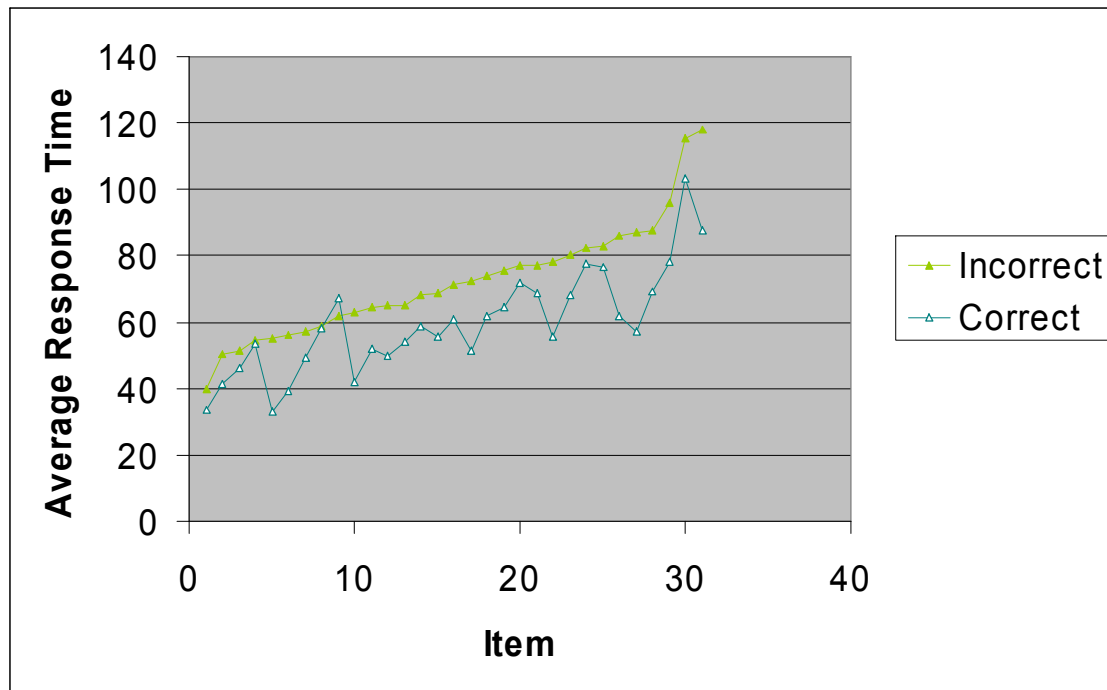


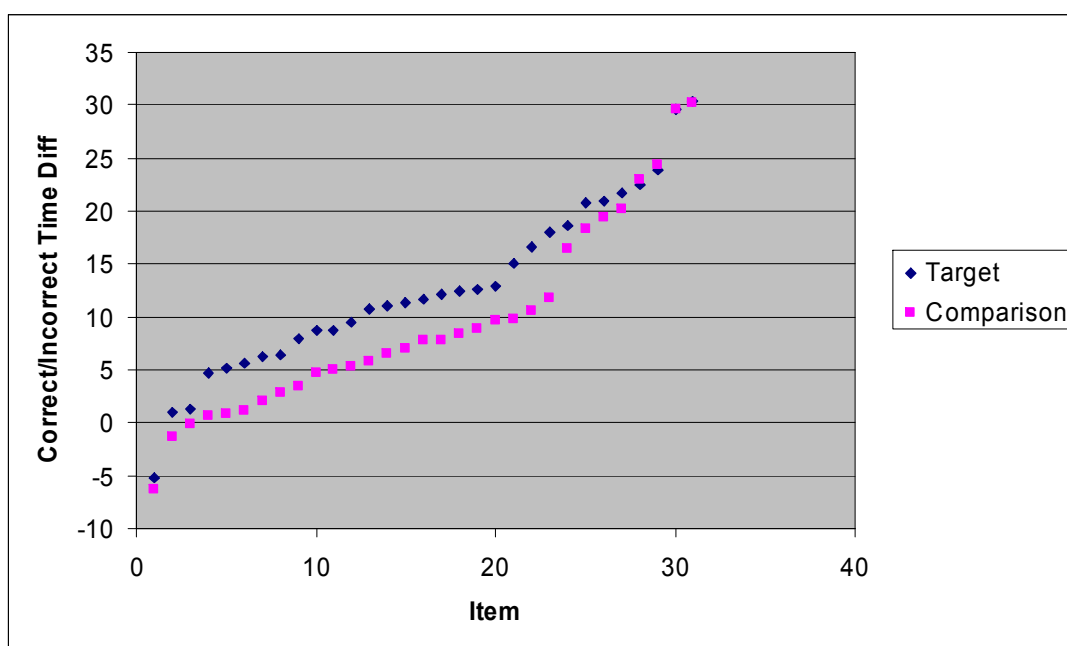
Figure 5 shows the average response time for each item in the “too easy” group, for correct and incorrect responses. This figure is ordered by the average response time for an incorrect response to the item. One clear pattern that can be seen is that correct responses in general took less time (average response time was 59.60 seconds) than incorrect responses (average response time was 72.26 seconds). This pattern held for 30 of 31 of the “too easy” items.

While the pattern of lower response times for correct responses is interesting, it doesn’t directly suggest item compromise, since it may just have taken longer for a candidate to give an incorrect answer than a correct one. To investigate this question further, the correct/incorrect response time differential was calculated for the target items in the “too easy” group by subtracting the average response time to answer correctly from the average response time to answer incorrectly for each item of interest. This analysis was also conducted for the comparison items associated with each target item.

Figure 6 shows the correct/incorrect response time differential for the “too easy” items and their comparison items, ordered smallest to largest. From this figure, it can be seen that there was a consistent pattern in which the target item had a greater difference between the time for a correct response and the time for an incorrect response. The average response time discrepancy for the “too easy” target items was 12.66 seconds, with a standard deviation of 8.23. The average response time discrepancy for the comparison items was 9.47 seconds, with a standard deviation of 9.10. The t-test and

Mann-Whitney U test fell just short of significance ($p < .10$), but the target items had a larger correct/incorrect discrepancy in 27 of 31 comparisons. This pattern suggests that individuals had to spend less time than expected answering the “too easy” items correctly.

Figure 6. Average response latency difference between incorrect and correct responses for each target and comparison item



Longitudinal Analysis

In the longitudinal analysis, changes in proportion passing and various response time percentile points were studied across the 10 administration periods. The 47 target items varied in the number of administration periods in which they appeared, ranging from 10 periods (1 item) down to 4 periods (33 items). Twenty seven of the items were removed from the operational item pool for two or more administration periods and put back into the item pool during a later administration period (which we termed a *hiatus*).

Table 2. Changes in Proportion Correct and Response Time

Variable	Minimum	Maximum	Median
Change in Proportion Correct	-.06	0.20	0.05
Change in P ₁₀	-8.0	1.0	-1.0
Change in P ₅₀	-10.0	6.0	-2.0

One of the items was found to have a very large (roughly 50%) increase in response times after it experienced a hiatus of two administration periods, and it was suspected that this item's content had somehow changed. Because we could not access the item content, however, and thereby verify whether or not the content had changed, we decided to omit the item from the longitudinal analyses. Table 2 summarizes the changes in proportion correct and two response time percentiles (P_{10} and P_{50}) across the remaining 46 target items. Consistent with the cross-sectional analyses, proportion correct tended to increase across administration periods, while response times tended to decrease.

Figure 7 shows the histogram of proportion correct for the target items. Although most items became easier over time (averaging a .05 increase), the changes in proportion correct were quite variable across items. Figures 8 and 9 show the corresponding response time distribution histograms for P_{10} and P_{50} , respectively. Most items exhibited modestly shorter response times (averaging a 2 second decrease) across administration periods. As with proportion correct, however, the changes in P_{10} and P_{50} varied substantially across items.

Figure 7. Average change in proportion correct for the target items

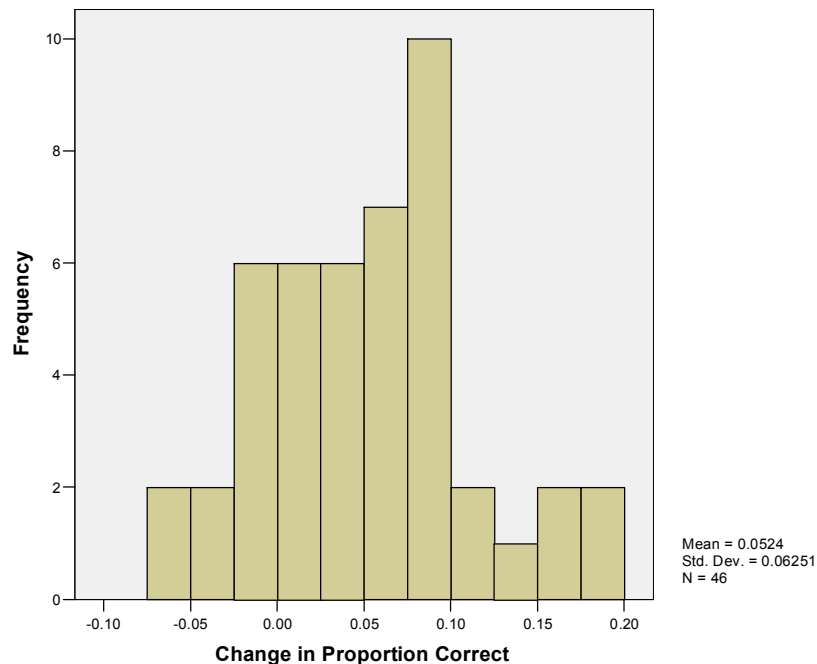
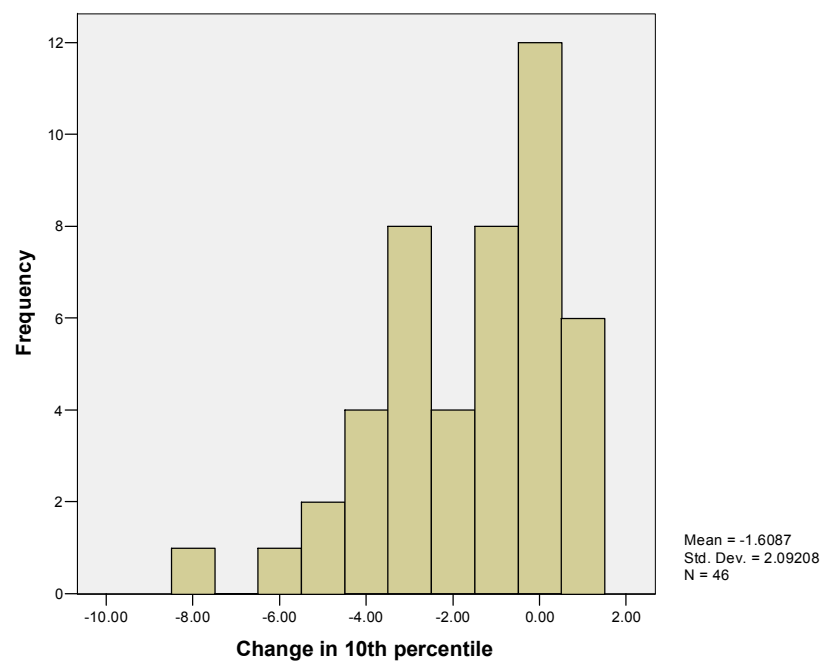
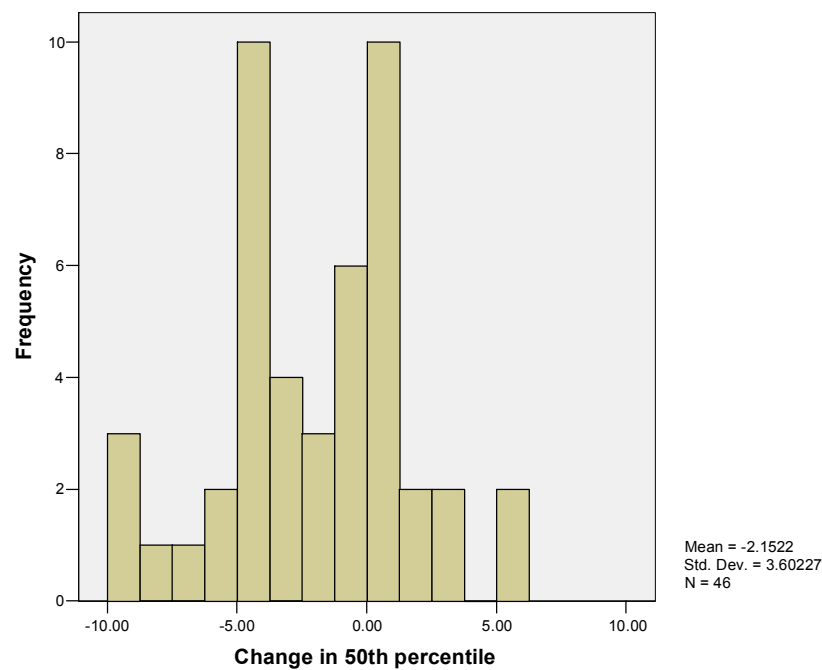


Figure 8. Average change in P₁₀ for the target items**Figure 9. Average change in P₅₀ for the target items**

Correlations among several of the study variables are found in Table 3. Change in proportion correct showed significant negative correlations with both change in P_{10} and P_{50} . This was consistent with expectations that item compromise should be indicated by both higher proportion correct and shorter response time. Other correlation values in Table 3, however, were unexpected. Number of administration periods was uncorrelated with change in proportion correct or response time. Somewhat more surprisingly, hiatus length (i.e., how long an item was pulled out of the operational pool) showed a significant positively correlated with change in proportion correct and significantly negative correlations with change in response time. These results suggest that if compromised items exhibit higher proportion correct and shorter response time, then removing the item from the pool for a time did not mitigate the item compromise effects. In fact, the longer the item was held out, the greater the item compromise effects appeared to be.

Table 3. Correlations Among Study Variables in the Longitudinal Analyses

Variable	1	2	3	4	5
1. Change in P_{10}	--	.76*	-.52*	.00	-.32*
2. Change in P_{50}		--	-.52*	.06	-.30*
3. Change in Proportion Correct			--	.06	.44*
4. Number of Administration Periods				--	-.26
5. Hiatus Length					--

* - $p < .05$

Discussion

The results of this study are tantalizing, but hardly definitive. In both the cross-sectional and longitudinal studies, indications were found that indicated a relationship between unusual response times and item misfit. However, neither analysis indicated a clear finding that response times could be used as a measure of item compromise.

The study was limited in several ways, including the following:

- First, the scope of the project prevented a complete analysis of correct and incorrect latency differences in the longitudinal analysis. A follow up study that allows this analysis is an obvious next step.
- Second, the items targeted were items with poor fit statistics, not items that had been identified as compromised. The poor fit could be due to a variety of factors not involving compromise. If a set of items could be identified as compromised, it would strengthen the study. Unfortunately, the folks who administer the test prefer not to have compromised items in the operational pool (no sense of adventure).

- The actual content of the items was not available. Since response time is often associated with the physical characteristics of the items this reduced the ability to do a detailed item-by-item analysis.
- The order in which candidates took items was not available. Since item position within a test is often associated with response time, this reduced the ability to identify whether fatigue was a factor in the response times observed.

Even with its limitations, the study began to reveal interesting aspects of item response times that might be related to anomalous performance of the items. While this is clearly a complex issue, several findings of the study are worth noting.

- First, the tendency for the target items to develop shorter response times across administrations indicates that anomalous response times may be an indicator that an item is beginning to drift as it is used for a period of time.
- Second, items that were “on hiatus” still showed shorter response times when they returned to active use. This may indicate that once an item has been compromised, it continues to lose its performance characteristics as time passes and it becomes more widely known. While this is speculation at this point, it is speculation that matches the facts at hand.
- Third, the element of item response time that might be easiest to study for signs of drift may be the difference in response times for correct and incorrect responses. This index would automatically correct for overall response time differences among items, and seems to show consistently larger values for the target items in the study.

The characteristics of item response times can be indicative of a variety of characteristics of test takers. Among these, motivational differences, differences due to cheating, and differences due to conditions that cause individuals to rush may all contribute to lower score reliability and validity. As we learn more about response times, they may lead us to view test and item performance differently. In time, we may be able to actively filter individual scores that may be inaccurate, and instantly remove items from operational item pools that begin to show drift in response times. While it is clear that this research is far from complete, it does indicate a small sample of the capabilities that will soon be practical.

References

- Kingsbury, G. G. & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.
- Wise, S. L., Bhola, D. S., & Yang, S. (2006, April). *Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Wright, B. D. (1977). Solving measurement problems with the Rasch model, *Journal of Educational Measurement*, 14, 97-116.
- Wright, B. D. & Stone, M. H. (1979). *Best Test Design*. Chicago, IL: MESA Press.