An Exploration of Answer-Changing on a Computer-based High-Stakes Achievement Test

Gail C. Tiemann and Neal M. Kingston

University of Kansas

Center for Educational Testing and Evaluation

****DRAFT - Please do not reference without permission from the authors.****

An Exploration of Answer-Changing on a Computer-based High-Stakes Achievement Test

Introduction

In high-stakes educational achievement testing, data forensics has risen as a key approach for supporting the integrity and interpretability of student test scores. Item answer-changing analysis has also emerged as a method of flagging potential cheating behavior. Indeed the approach, also called erasure analysis, has been used recently by investigators to uncover several high-profile incidences of cheating in public schools. While answer-changing itself has been studied in the context of paper-pencil tests for decades, little is known about the construct related to computer-based tests.

As mentioned, erasure analysis via scanned paper and pencil tests has been used to explore and uncover incidences of cheating. Erasure analysis involves examining the instances of student answer-changing behavior on multiple choice tests, looking for aberrant or larger than expected numbers of answer-changes, or wrong-to-right answer changes. Aberrant patterns of answer-changes may be indicative of changes made to items by someone other than the student themselves (Qualls, 2001).

Most notably, erasure analysis was one of several methods used to uncover cheating in the Atlanta Public Schools. Statistical detection of cheating via erasure analysis was implemented and then followed up by a thorough investigative process, where ultimately 178 teachers and principals confessed to changing student answers at 44 schools (Severson, 2011).

Erasure analysis has been used in other states as well. The New York State education department conducted erasure analysis on the high school Regents exams. In this example, one of 64 flagged incidents resulted in the termination of a Bronx assistant principal (Otterman, 2011). This particular analysis showed that of 1,013 items erased on the exam, 94 percent had been changed from wrong-to-right, where typically, in that testing program, about 50 percent of answers are changed from wrong-to-right. Related to the application of erasure analysis in testing programs nationally, a recent USA Today survey (Bello & Toppo, 2011), reported that 20 states and Washington, D.C. routinely conduct erasure analysis on all student tests.

Applications of erasure analysis have typically used a "flagging rule," based on an evaluation of expected student erasure behaviors and then flagging patterns of erasures that exceed a pre-defined threshold. In one analysis conducted in a district in Morgan County Georgia, students were flagged if the count of student wrong-to-right answer changes exceeded the state average plus three standard deviations (Schiliro, 2010). Other flagging rules are more conservative. For example, the Louisiana Test Security policy (Louisiana Register Education Bulletin, 2012) offers a flagging rule of the state average of wrong-to-right answer changes plus four standard deviations. Recently, researchers at the Data Recognition Corporation applied a flagging rule based on the state average of wrong-to-right changes plus eight standard deviations (Primoli, Liassou, Bishop, & Nhouyvanisvong, 2011). In all cases though, the flagging rules are non-definitive and offer one step in a process of investigation. Often analysis is followed up with evaluation of other indicators, for example, the proportion of wrong-to-right changes to total item changes (Louisiana) or erasures per student (Otterman, 2011). Additionally, review of physical answer-sheets and extensive interviews are conducted.

Outside of the applied arena, researchers have been examining answer-changing as a construct for decades (Al-Hamly & Coombe, 2005) with the first study published in 1929 (Mathews, 1929). Many of these studies have explored whether students benefit from changing answers and if the conventional wisdom to stay with a first response is a myth (Benjamin, Cavell, & Shallenberger, 1984). Additionally, researchers have explored what factors possibly explain variability in answer-changing, as well as the frequencies and patterns of wrong-to-right, right-to-wrong, and wrong-to-wrong responses. Two particular studies synthesized the research in the area. Benjamin, Cavell, and Shallenberger (1984) reviewed 33 studies conducted between 1929 and 1984. Benjamin et al. (1984) reported a range of 2.5 to 9.0 percent of multiple choice items changed per test. Of these, the percent of wrong-to-right changes ranged from 44.5 to 71.3 percent. Al-Hamly and Coombe's (2005) study of answer-changing also summarized and reviewed additional answer-changing studies published after 1983. Results across these studies, in percent of items changed and in percent of wrong-to-right items, were consistent with the results in Benjamin et al. (1984). It should be noted that the synthesized studies varied widely in test content and test purpose, with most studies examining higher education populations. Additionally, all of the studies examined were administered via paper and pencil.

Studies have also examined potential sources of variance in answer-changing rates. Of note, numerous studies have examined the relationship between "student proficiency" and answer-changing, with studies defining "student proficiency" in a variety of ways, including test total score, course grades, outside standardized measures of proficiency, and self-report of academic performance. Results have been mixed. Benjamin et al.'s (1984) review cited six studies with statistically significant negative relationships between test scores and the number of answers changed on a test. Five other studies in the review reported "nonsignificant results in the same direction" (Benjamin et. al., 1984, p. 137). Other studies have also reported mixed results. Best (1979) found that students with higher grades tended to change answers less frequently than students with lower grades, and the former made fewer right-to-wrong changes than other students. Al-Hamly and Coombe (2005) also found that higher scorers changed answers less frequently than other students; however, McMorris et al. (1991) did not. Again, all of the tests administered in these studies were paper-pencil based and administered to a variety of audiences, mostly in higher-education contexts.

More recently, answer-changing has been documented in the research literature in context of statistical cheating detection. Qualls' (2001) study documented baseline answer-changing frequencies, including wrong-to-right frequencies for large-scale, K-12 achievement tests, in both low- and moderate-stakes contexts. The study first examined low-stakes, paper-pencil educational achievement test results from 16 districts in Iowa during the 1994-1995 school year. The results showed that "more than 90% of students changed three or fewer responses" (p. 12), with about 50% changing one response. Across the content areas and grades represented, 50% of students erased at least one answer with 38-64% erasing zero answers, 20-29% erasing one answer, and 7-16% erasing two answers. In addition, when students changed only one answer, 50% of these changes were wrong-to-right, and about 20% were right-to-wrong. Across the programs, as the number of answer changes increased, the number of wrong-to-right answer changes decreased. In the second study reported, results from moderate-stakes assessments were compared to the low-stakes assessment. Qualls found that erasure patterns were "remarkably similar" (p. 15) between the two assessment types. Qualls concluded that "it would be rare to see a student change more than 15% of the items" (p. 15), and that "wrong-to-right changes would not typically exceed 55% for one erasure, and for multiple erasures the number of 100% wrong-to-right changes would be dramatically lower" (p. 15). Ultimately, Qualls felt that erasure counts and wrong-to-right changes above these limits could be used to flag aberrant tests.

Primoli, Liassou, Bishop, and Nhouyvanisvong (2011) also examined answer-changing in the context of large-scale, K-12 achievement testing. Their study examined responses from four state testing programs, examining item- and student-level factors related to answer-changing proportions, wrong-to-right proportions, as well as wrong-to-right to total erasure proportions. Using each program's IRT-based ability metrics, this study found strong, cubic relationships between ability and total erasures proportions, as well as between ability and wrong-to-right erasures. Additionally, the research group applied a flagging rule, equivalent to the number of wrong-to-right erasures, plus eight standard deviations (p. 22). Individual schools were also flagged if more than one student per school was flagged using the eight standard deviation flagging rule. As with the Qualls study, all results were based on paper-pencil tests with erasures detected via optical scanning equipment. Primoli et al. (2011) cautioned readers that erasure counts could vary by program, since optical scanning sensitivity settings often vary.

While the studies described have answer-changing in various settings, with various age-groups, and with tests of various types, all documented the construct in terms of paper-pencil testing, using either hand or optical scanner detection of changed item responses. The purpose of this exploratory study is to document baseline answer-changing behaviors in a computer-based, high-stakes educational achievement testing context, in hopes to further understand this evolving construct.

Methods

One Midwestern state administers 99.7% of all summative achievement tests via computer-based assessment software provided by the state. The software allows students to answer items in any order, as well as to review and change answers as frequently as desired. The test is not timed.

The data used for this study was obtained from one form of the state's third-grade summative achievement test in mathematics administered during the spring of 2012. While students were randomly enrolled into several possible test forms, data from one, 70-item form was used for this study. The results reported represent an estimated 98% of the expected spring testing window for this program. Future analyses will document results from the entire testing window.

In order to capture answer-changing, the computer-based assessment system logged each student's path through the test, noting the answer marked on the screen when the student navigated away from the item, as well as the time the student spent viewing the screen. In this way, the system was able to collect how many times the student reviewed an item, as well as any changes the student made to the item. Additionally, the testing system allows examination of precise answer-changing patterns, for example, whether a student changed an item one, two, or even 10 times.

Because the context of this work is to explore possible cheating detection methods by individuals other than students themselves, we focused on the final change made to the item. While we have noted a variety of answer-changing patterns in the data, we defined an overall "wrong-to-right" item as an item where the final change made to the item was wrong-to-right, regardless of how the student had answered the item initially.

Results

The number of students taking the selected form of the third-grade mathematics test was 5328. The students represented 281 public school districts, 646 public school buildings, and 103 private school buildings across the state. Public school students accounted for 92 percent of the test forms administered.

Total raw scores on the 70-item test ranged from 6-70, with a mean of 56.19 and a standard deviation of 8.19. Based on the 2012 actual data, internal consistency of the form was .89. The distribution of total test scores is shown in Figure 1.

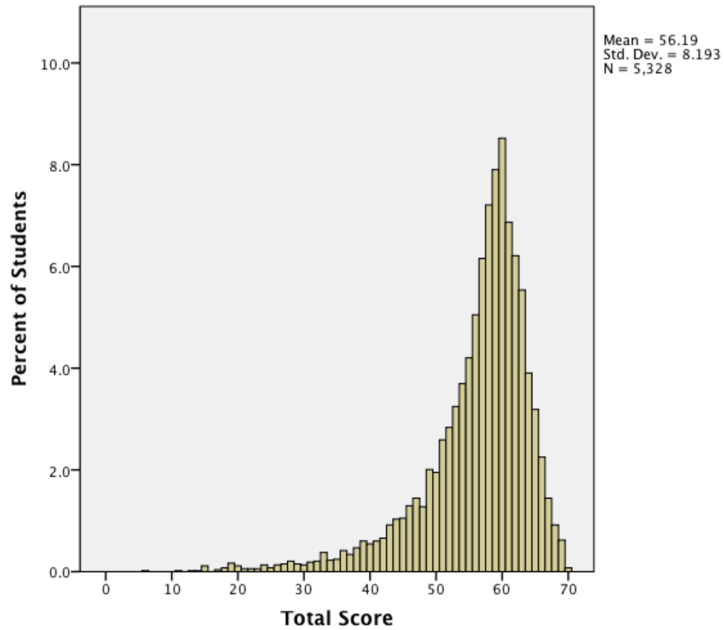

Figure 1. Total test score distribution.

Answer-changing frequencies

Of the 5328 students who completed the form, 52% (n = 2791) changed their answer on at least one item. Ninety-six percent of the students changed four or fewer items. The mean number of items changed per student was 1.12 (1.16% of items) with a standard deviation of 1.74. The distribution of items changed is shown in Table 1 and Figure 2.

Table 1

Frequencies of Items Changed

Items Changed	Count of Students	Percent	Cumulative Percent
0	2537	47.6	47.6
1	1419	26.6	74.2
2	652	12.2	86.5
3	337	6.3	92.8
4	155	2.9	95.7

5	78	1.5	97.2
6	58	1.1	98.3
7	33	0.6	98.9
8	18	0.3	99.2
9	12	0.2	99.5
10	10	0.2	99.6
11	4	0.1	99.7
12	1	0.02	99.7
13	2	0.02	99.8
14	3	0.1	99.8
15	3	0.1	99.9
16	2	2	99.9
18	1	1	99.9
19	1	1	100
22	2	2	100

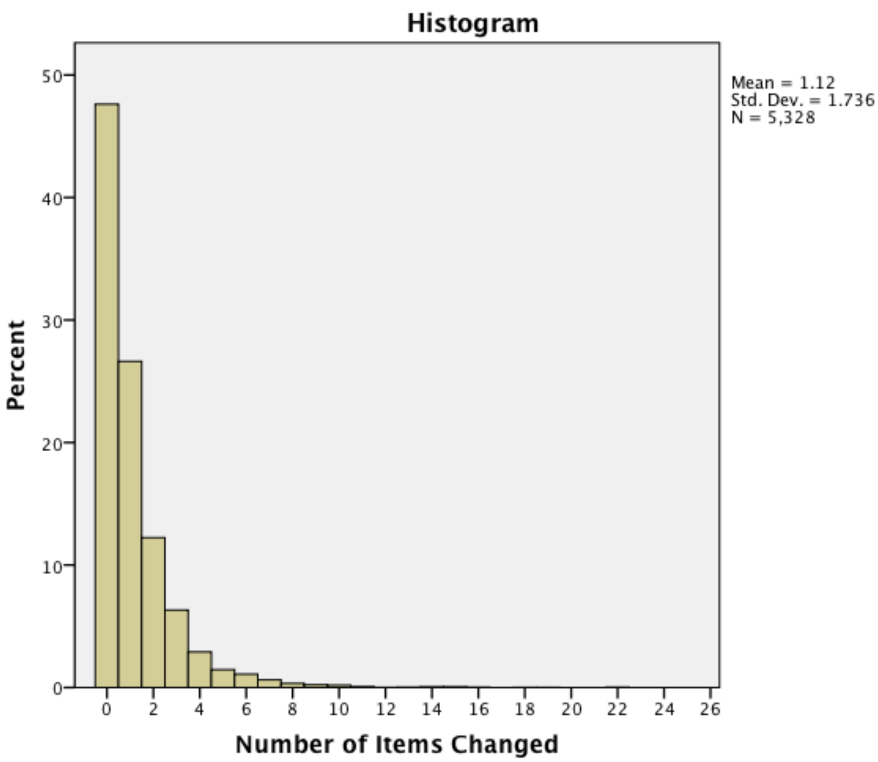

Figure 2. Distribution of total number of items changed.

As mentioned earlier, wrong-to-right answer changes were defined as an item changed from a wrong answer to a right answer based on the last change made to an item. Right-to-wrong answer changes were defined in the same manner. The mean number of wrong-to-right changes per student across this test form was 0.74 with a standard deviation of 1.31. Forty percent of students had at least one wrong-to-right change. Right-to-wrong answer changes were less frequent; the mean right-to-wrong changes per student was 0.38 with a standard deviation of 0.82. Only 26% of students had at least one wrong-to-right answer change. The frequencies for wrong-to-right and right-to-wrong and answer changes are shown in Table 2 and 3.

Table 2.

Wrong-to-right Frequencies

Wrong-to-Right Changes	Count of Students	Percent	Cumulative Percent
0	3158	59.3	59.3
1	1271	23.9	83.1
2	510	9.6	92.7
3	198	3.7	96.4
4	81	1.5	97.9
5	45	45	98.8
6	25	0.8	99.2
7	11	0.5	99.5
8	15	0.2	99.7
9	3	0.3	99.8
10	2	0.1	99.8
11	3	0.06	99.9
14	2	0.04	99.9
15	2	0.04	100
16	1	0.02	100
17	1	0.02	100

Table 3.

Right-to-wrong Frequencies

Right-to-Wrong Changes	Count of Students	Percent	Cumulative Percent
0	3949	74.1	74.1
1	1002	18.8	92.9
2	248	4.7	97.6
3	76	1.4	99.0
4	24	0.5	99.5
5	15	0.3	99.7
6	5	0.1	99.8
7	4	0.1	99.9
8	3	0.06	100
10	1	0.02	100
16	1	0.02	100

The following contingency tables explore the number of wrong-to-right and right-to-wrong changes by number of item changes made, up to 10 total changes. As noted in the frequency tables above, few students changed more than 5 total answers on the test. Of note on the diagonals are students who changed 100% of the items wrong-to-right. These numbers drop off significantly after 3 total changes.

Table 4.

Changes from Wrong-to-Right by Total Items Changed

	Wrong-to-Right Count																	
Items Changed	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	504	**915**																
2	92	252	**308**															
3	18	75	131	**113**														
4	6	19	42	60	**28**													
5		5	16	8	27	**22**												
6		4	8	10	14	11	**11**											
7	1	1	3	3	6	7	6	**6**										
8				3	1	4	3	2	**5**									
9			1		3		1	1	4	**2**								
10				1	2	1	1	1	3	1	**0**							
11							1				**2**	**1**						
12			1															
13								1				1						
14							1		1			1						
15									1						1	**1**		
16									1							1		
17																		
18																	1	
19																		1
20																		
21																		
22							1								1			

Number of students = 5238 Number of items = 70

The following contingency table complements the Table 4. The number of students changing more than 5 total answers is also small. Additionally, students with a higher number of changes had right-to-wrong items in addition to wrong-to-right items.

Table 5.

Changes from Right-to-Wrong by Total Items Changed

	Right-to-Wrong Count																
Items Changed	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	915	**504**															
2	308	252	**92**														
3	113	131	75	**18**													
4	28	60	42	19	**6**												
5	22	27	8	16	5												
6	11	11	14	10	8	4											
7	6	6	7	6	3	3	1	**1**									
8	5	2	3	4	1	3											
9	2	4	1	1		1		1									
10		1	3	1	1	1	2	1									
11	1	2				1											
12											1						
13			1				1										
14				1			1		1								
15	1	1						1									
16		1							1								
17																	
18			1														
19			1														
20																	
21																	
22									1								1

Number of students = 5238 Number of items = 70

The final contingency table plots wrong-to-right changes against right-to-wrong changes. Note that a few students with higher than average counts of wrong-to-right also had higher than average counts of right-to-wrong.

Table 6.

Right-to-Wrong Frequencies Versus Wrong-to-Right Frequencies

	Right-to-Wrong Count									
Wrong-to-Right Count	1	2	3	4	5	6	7	8	9	10
1	252	75	19	5						
2	131	42	16	3	4	1	1			
3	60	8	10	3	3		1			1
4	27	14	6	1	3	2				
5	11	7	4		3					
6	6	3	1	1	1	1		1		
7	2	1	1		1	1				
8	4	3					1	1		
9	1									
10	2									
11		1	1							
14	1							1		
15	1									
16		1								
17		1								

Number of students = 5238 Number of items = 70

Answer-changing patterns

While the above results focused on the number of *items* changed, the following results focus on the number of *changes* made to each item and the *pattern* of those changes. The ability to log the number of changes to an item and the precise pattern of those changes is a unique benefit of computer-based testing. Table 7 describes the patterns found in the data. Note that when answers were changed, the majority of the time they were changed only one time. Items changed three or more times were rare; no single answers were changed more than four times.

For items changed twice, the pattern of wrong item choices is further denoted. For example, some students changed from a wrong item response to a right item response, and back to the same wrong item response. Others changed from a wrong response to a different wrong response. Those changes are denoted in the table with subscripts. Additionally, patterns that would yield an overall wrong-to-right label in this study are denoted with an asterisk.

Table 7.

Answer-Changing Patterns

Pattern	Number of Students	Number of Items
Items Changed 1x	2731	5641
WR*	2101	3733
RW	672	818
WW	869	1090
Items Changed 2x	240	286
RWR*	126	137
WWR*	61	66
RWW	7	7
W_1RW_1 (same wrong)	18	18
W_1RW_2 (different wrong)	26	26
$W_1W_2W_1$ (same wrong)	23	23
$W_1W_2W_3$ (different wrong)	8	8
Items Changed 3x	25	26
WRWR*	7	7
WWWR*	7	7
RWRW	4	4
WRWW	6	6
WWWW	2	2
Items Changed 4x	4	4
RWWWR*	1	1
WRWWR*	1	1
WWWRW	2	2

Number of students = 5238 Number of items = 70 *denotes a WtR Pattern

Relationships between answer-changes and total score

The relationship between student total score and total answer changes, as well as the relationship between total score and wrong-to-right answer changes, were explored via linear regression. As mentioned earlier, results from prior studies in this area have yielded mixed results. In this study, for students who changed at least one answer, the relationship between total test score and total item changes was statistically significant, likely due to large sample size. However, total score explained less than 2 percent of the variance in total item changes. The relationship between total score and right-to-wrong item changes was not statistically significant, as is likely confounded (Mueller & Wasser, 1977). Scatterplots for both relationships are presented in Figures 3 and 4.

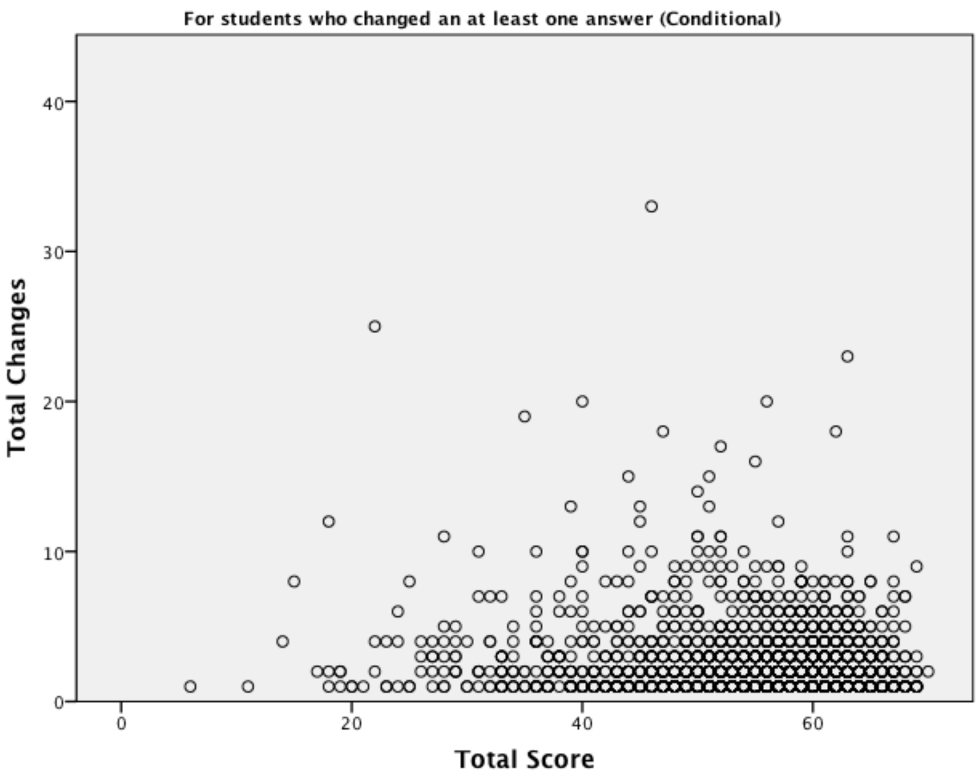

Figure 3. Plot of total score to total item changes for students who changed at least one item.

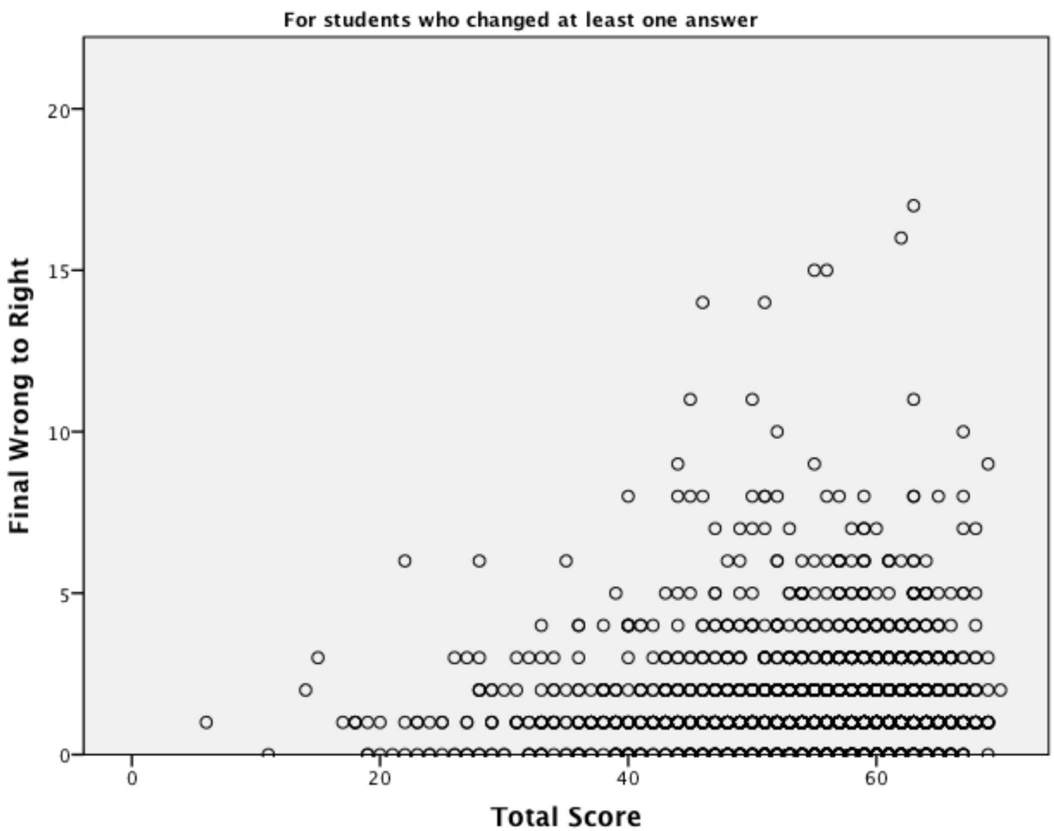

Figure 4. Plot of total score to right-to-wrong answer changes for students who changed at least one item.

Elapsed Time

The computer-based assessment system used to administer this test form collects the elapsed time that a student spends on the screen from the point where an item is displayed, to the point where the student navigates away from the item. Elapsed time for answers changed from wrong-to-right (the final change in the pattern) ranged from 2 seconds to 17 minutes, with a mean of 56 seconds and a standard deviation of 81 seconds. In the context of cheating, one might expect answers to be changed from wrong-to-right quickly. Therefore, the frequencies of wrong-to-right changes for elapsed screen times of 60 seconds or less (n = 2866 items) is shown in Table 8. The distribution of these values is shown in Figure 5.

Table 8.

Frequencies of Wrong-to-Right Items Changed 60 Seconds or Less Elapsed Screen Time

Seconds	Count of Wrong-to-Right Items	Percent	Cumulative Percent
1	0	0	0
2	16	.6	.6
3	67	2.3	2.9
4	111	3.9	6.8
5	90	3.1	9.9
6	82	2.9	12.8
7	101	3.5	16.3
8	72	2.5	18.8
9	86	3.0	21.8
10	66	2.3	24.1
11	84	2.9	27.0
12	68	2.4	29.4
13	76	2.7	32.1
14	70	2.4	34.5
15	75	2.6	37.1
16	81	2.8	40.0
17	63	2.2	42.1
18	67	2.3	44.5
19	53	1.8	46.3
20	69	2.4	48.7
21	58	2.0	50.8
22	50	1.7	52.5
23	65	2.3	54.8
24	55	1.9	56.7
25	58	2.0	58.7
26	46	1.6	60.3
27	49	1.7	62.0
28	61	2.1	64.2
29	53	1.8	66.0
30	45	1.6	67.6
31	59	2.1	69.6
32	41	1.4	71.1
33	33	1.2	72.2
34	47	1.6	73.9
35	49	1.7	75.6
36	33	1.2	76.7
37	31	1.1	77.8
38	46	1.6	79.4
39	45	1.6	81.0

40	38	1.3	82.3
41	34	1.2	83.5
42	29	1.0	84.5
43	30	1.0	85.6
44	38	1.3	86.9
45	30	1.0	87.9
46	22	.8	88.7
47	31	1.1	89.8
48	30	1.0	90.8
49	18	.6	91.5
50	34	1.2	92.6
51	29	1.0	93.6
52	23	.8	94.5
53	21	.7	95.2
54	28	1.0	96.2
55	28	1.0	97.1
56	16	.6	97.7
57	18	.6	98.3
58	17	.6	98.9
59	17	.6	99.5
60	14	.5	100.0
Total	2866	100.0	

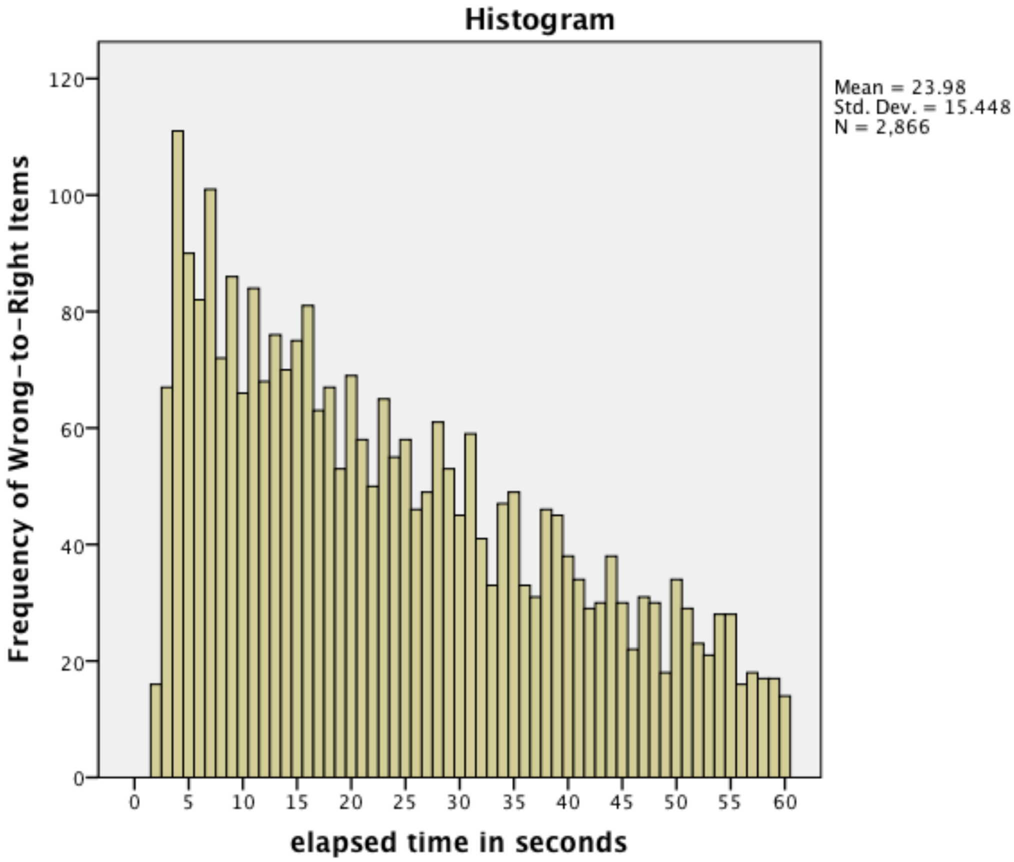

Figure 5. Frequency of wrong-to-right items changed in less than 60 seconds elapsed screen time.

Additionally, Table 9 shows the number of wrong-to-right items per student changed via elapsed screen times of less than 10 seconds. Very few students changed more than 3 items from wrong-to-right in less than 10 seconds. The histogram of these values is displayed in Figure 6.

Table 9.

Wrong-to-Right Items Per Student Changed in 10 seconds or Less Elapsed Screen Time

Wrong-to-Right Items per Student	Count	Percent	Cumulative Percent
1	456	82.8	82.8
2	75	13.6	96.4
3	9	1.6	98
4	4	0.7	98.7
5	4	0.7	99.5
6	1	0.2	99.6
7	1	0.2	99.8
8	1	0.2	100
Total	551	100	

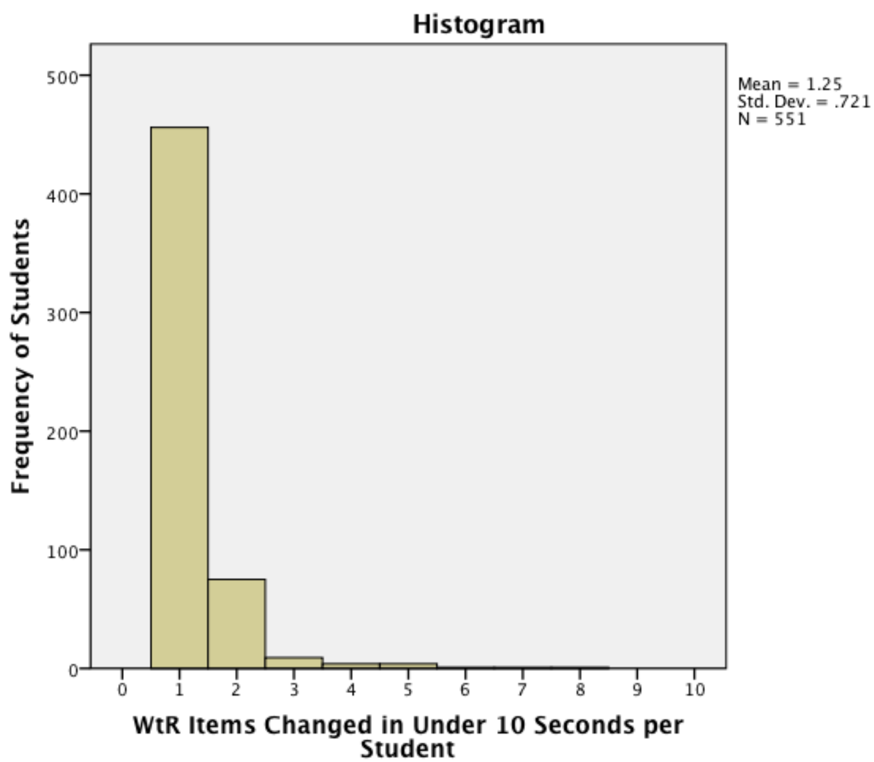

Figure 6. Wrong-to-right items changed per student, 10 seconds or less of elapsed screen time

Flagging Rule Application

As mentioned earlier, flagging rules based on wrong-to-right answer counts have been used to flag potential instances of cheating. Flagging rules consisting of the mean wrong-to-right count for the state plus 4, 5, or even 8 standard deviations have been noted in the literature (Louisiana Board of Elementary and Secondary Education, 2012; Primoli et al., 2011; Schiliro, 2010). The four and eight standard deviation flagging rules were applied, in an exploratory manner only, to the data in this study. Based on the state average for *the target form only*, the cut point for wrong-to-right answer changes is 6 using the four standard deviation rule, and 11 using the eight standard deviation rule. Application of these cut points yielded 65 flagged students and 6 flagged students, respectively. Wrong-to-right changes to total change proportions were then examined for the 6 flagged students; these proportions ranged from 0.64 to 1.00. The elapsed screen time of the wrong-to-right changes for these students ranged from 4 seconds to 236 seconds. In addition, for all students who were flagged, we checked the server logs to see if tests were submitted on weekends. No such tests were found. Additionally, flagged students did not demonstrate multiple changes to single items.

The field is uncertain as to the actual distribution in the population of wrong-to-right answer counts (Bishop, Liassou, Bulut, Dong, & Stearns, 2011). Under a normal distribution, one would expect less than one student to be flagged using the four standard deviation flagging rule. Additionally, under the Poisson distribution, we would expect the only 3 students to have more than five wrong-to-right changes.

Discussion

The purpose of this study was to explore the construct of answer-changing on a computer-based, high-stakes K-12 achievement test. There are significant limitations. First, the study targeted a single test form administered during the spring testing window in 2012. Of this spring window, due to time limitations, only an estimated 98 percent of the student test takers were captured in this data; future studies will be completed on the entire window. Additionally, the study examined only one subject and one grade in one year. Thus, no generalizations should be made to other forms, other grades, other subjects, or tests taken in other years.

Results showed that about half of students (52 percent) changed one or more answers, an average of approximately 1 answer change per student. This value is slightly lower than the range of average items changed per student reported in Benjamin et al. (1984), but very much in line with total item answer-changing percentages reported by Qualls (2001).

Related to wrong-to-right answer changes, 40 percent of students in this study had at least one wrong-to-right answer change. This value is slightly lower than those reported by Benjamin et al. (1984). In addition, of the students in this study who changed only one item, 915 (64%) changed that item from wrong-to-right. This value is higher than the value reported by Qualls (2001). Again, previous literature was based on paper-pencil administered tests in low to moderate stakes assessments.

In this study, no relationships were found between total test score and total answer changes or between total test score and wrong-to-right changes. However, work by Primoli et al. (2011) did find a cubic relationship between "ability," as measured by the continuous IRT-based metric of ability, and answer-changing proportions and wrong-to-right proportions. Future analysis of this data using the IRT-based ability metric could further explore this relationship in this data. Additionally, future analysis should explore other potential sources of variability including item-based factors (difficulty, item position, subject), and additional student-based factors (demographics, districts, buildings).

Results also showed that the majority of the wrong-to-right answer changes on this test were made in less than 60 seconds of screen time and that the distribution of screen times for wrong-to-right changes is not normally distributed. Additionally, a small number of students had more than two wrong-to-right answer changes in less than 10 seconds. Future research could explore the utility of a wrong-to-right by screen time flagging rule.

We found very little evidence of suspicious behavior in this data. A few students had larger than expected wrong-to-right changes in less than 60 seconds of screen time. For these students, more in-depth analysis of the server logs could explore where the changes were made in the student's path through the test. A large number of single wrong-to-right changes made in quick succession at the end of the student's path would lead to more alarm than changes made at various or random positions in the path.

While several flagging rules were applied to this data, no claims are made regarding cheating behavior by the students or their teachers. Results are purely exploratory. It is clear that computer-based tests provide significantly greater information to probe anomalies that might indicate cheating than was previously available from scanned paper-pencil answer sheets. It is also hoped that by further understanding typical answer-changing patterns, more can be known related to potential aberrant patterns. As always, many questions remain unanswered and more research is needed to document this important construct.

References

Al-Hamly, M., & Coombe, C. (2005). To change or not to change: Investigating the value of mcq answer changing for gulf arab students. *Language Testing, 22*(4), 509-531.

Bello, M., & Toppo, G. (2011, September 13). Few states examine test erasures, *USA Today*. Retrieved from http://www.santafenewmexican.com/localnews/State-to-check-for-cheating-teachers

Benjamin, L. T., Jr., Cavell, T. A., & Shallenberger, W. R., III,. (1984). Staying with initial answers on objective tests: Is it a myth? *Teaching of Psychology, 11*(3), 133-141.

Best, J. B. (1979). Item difficulty and answer changing. *Teaching of Psychology, 6*(4), 228-230.

Bishop, N. S., Liassou, D., Bulut, O., Dong, G. S., & Stearns, M. (2011). *Paper four: Application and comparison of alternative procedures*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Louisiana register education bulletin 118 - statewide assessment standards and practices § 309 (2012).

Mathews, C. O. (1929). Erroneous first impressions on objective tests. *Journal of Educational Measurement, 20*(4), 280-286. doi: 10.1037/h0071721

McMorris, R. F., & et al. (1991). Why do young students change answers on tests? (p. 22).

Mueller, D. J., & Wasser, V. (1977). Implications of changing answers on objective test items. *Journal of Educational Measurement, 14*(1), 9-14.

Otterman, S. (2011, September 23). State says it analyzed test erasures for cheating; 62 schools proved suspect, *The New York Times*. Retrieved from http://www.nytimes.com/2011/09/24/nyregion/in-reversal-new-york-state-says-it-used-erasure-analysis-to-detect-cheating.html?pagewanted=all

Primoli, V., Liassou, D., Bishop, N. S., & Nhouyvanisvong, A. (2011). *Paper two: Erasure descriptive statistics and covariates*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Qualls, A. L. (2001). Can knowledge of erasure behavior be used as an indicator of possible cheating? *Educational Measurement: Issues and Practice, 20*(1), 9-16.

Schiliro, K. (2010, February 18). Questionable crct answer sheets at mces?: "Minimal concern" category by two-tenths of a percent, *Morgan County Citizen*. Retrieved from http://www.morgancountycitizen.com/?q=node/12946

Severson, K. (2011, July 5). Systematic cheating is found in Atlanta's school system, *The New York Times*. Retrieved from http://www.nytimes.com/2011/07/06/education/06atlanta.html