Patterns of Erasure Behavior for a Large-scale Assessment

Andrew A. Mroch, Yang Lu, Chi-Yu Huang, and Deborah J. Harris

ACT, Inc.

Patterns of Erasure Behavior for a Large-scale Assessment

Erasing behavior on multiple-choice test answer sheets has been increasingly used to screen for test irregularities or unusual patterns of test item responses, particularly to check for test proctors, teachers, or others involved with tests that inappropriately affect student responses or perhaps change students' answers to test items. For example, analysis of erasures was one source of evidence in recent cheating scandals in Atlanta (Severson, 2011) and Pennsylvania (Winerip, 2011).

Erasures refer to item responses that an examinee erased and changed for one or more answer on an answer sheet[1]. As described by Mead, Andersen and Korts (2010), erasures can be explained in three ways: (1) rethinking: where an examinee reconsiders their initial response and makes a change, (2) misalignment: where an examinee inadvertently places answers to a particular item next to a different item (or items) in the answer document, discovers the mistake, and corrects it by realigning responses in the answer document, or (3) irregularity: where cheating occurs. Determining *why* unusual or irregular erasures occur is a more challenging task than determining *that* they occur. Similar to most other psychometric evidence of aberrant responding, aberrant erasure patterns would typically be used to corroborate other evidence or screen examinee responses for further scrutiny.

Erasures are one aspect of examinee responses that can be used to identify *aberrant responding*[2], however examinee responses in the form of correct/incorrect answer or response

[1] Here we are referring to erasures of responses to multiple choice items administered in paper-and-pencil format. However, analogous analysis is possible on a computerized version of a test, for example if examinees are able to go back and change answers, such behavior could be recorded and analyzed. Although not the focus of this study, additional behavior, such as response time, could also be collected in a computerized environment that might be useful for flagging aberrant responding.

[2] Aberrant responding refers to atypical or unusual responding.

option chosen can also be used. A variety of psychometric methods have been proposed and used for identifying aberrant responding (for an overview, see Cohen & Wollack, 2006). Some methods make relatively strong assumptions using statistical models to identify aberrant responding (e.g., Bellezza & Bellezza, 1989; Frary, Tideman, & Watts, 1977; Hanson, Harris, & Brennan, 1987; Holland, 1996; Meijer & Sijtsma, 1995, 2001; Sotaridona & Meijer, 2003; Sotaridona, van der Linden, & Meijer, 2006; van der Linden & Jeon, 2012; Wollack, 1997). Other methods may be more descriptive in nature, making relatively few statistical assumptions (e.g., Qualls, 2001). In this study, we will use a descriptive approach to analyze erasures which will make relatively weak statistical assumptions to identify examinees with aberrant erasure behavior.

Erasure behavior for each item is typically recorded when examinee answer sheets are scanned. The specific types of erasure behavior recorded may vary, but often in addition to recording that an erasure has occurred, the following may be obtained from examinees' responses to each item: changing an answer from wrong-to-right (WR), changing an answer from wrong-to-wrong (WW), or changing an answer from right-to-wrong (RW).

Erasure analysis used with district and state testing programs has typically focused on group-level analyses (e.g., classrooms or schools) and on wrong-to-right (WR) erasures. When a person (e.g., test proctor or teacher) inappropriately affects or otherwise changes incorrect responses to items for a group of examinees (e.g., classroom or school), the idea is that erasures will appear unusual compared to other groups of examinees (e.g., other classrooms or schools) where responses were not changed. For example, a school where systematic cheating occurred may have substantially more WR erasures compared to other schools in a state. This type of pattern is one way that a group of examinee responses could be flagged for further investigation.

While much of the limited literature on erasure, and erasure analysis reported in recent cheating scandals, has focused on group-level erasure patterns (e.g., erasure patterns for classes, schools, or districts), another way of analyzing erasure patterns is to focus on individual examinees (e.g., Primoli, Liassou, & Bishop, 2010). Such analysis compares individual examinee erasure patterns to a larger sample of individual examinees to identify atypical levels of erasure. Few published studies have explored individual erasure patterns, which is one of the reasons we explore them here. In addition, for some tests, examinees testing in the same room may not have a particular relationship among one another and proctors may not know examinees and are unlikely to have particular motivations to change examinees' answers or allow cheating to occur. Examples of such tests include administrations of some college and other post-secondary admissions tests and licensure and certification tests.

The purpose of this study is to examine the potential usefulness of individual-level erasure analysis using data from a large-scale assessment. The analyses conducted in this study provide practical baseline information about likely characteristics of individual erasure patterns that could be used to inform the usefulness of erasure behavior for flagging individual examinees based on aberrant erasure behavior. This study will investigate the following questions:

- Do erasure behaviors differ between high and low ability examinees?
- Does a relationship exist between erasure behaviors and item difficulty/item characteristics?
- Do erasure behaviors differ across test forms (in the same subject)?
- Do erasure behaviors differ across tested subject areas?
- What flagging index and flagging criteria are appropriate when erasure analysis is used on an individual level?

Method

Data

Data for this study came from administrations of four test forms (*form 1, form 2, form 3,* and *form 4*) of a large-scale standardized test that consisted of tests in four subject areas (subject 1, subject 2, subject 3, and subject 4). Each test form consisted of between about 3,000 to about 5,000 examinees in each subject area. Test form lengths ranged from at least 40 items to over 70 items, with subject 3 and subject 4 tests having fewer items than subject 1 and subject 2 tests. The forms were developed over the course of a five year period and may contain slight differences in difficulty. The proportion of items answered correctly by test form and subject are listed in Table 1. One thing to notice is that the proportion of items answered correctly on form 1 was relatively higher than forms 2-4.

Item Difficulty Groupings

To explore possible differences in erasure behavior across item difficulty level, we grouped items by difficulty. After some initial analysis of different possibilities, we chose to define groupings based on dividing the range of observed p-values into four groups, defining the lowest range as the “easy” items, the middle two ranges as the “moderate” items, and the upper range as the “hard” items. The goal of defining item difficulty groupings in this way was to separate “low” and “high” difficulty items from “moderate” difficulty items, as indicated by item p-values (the proportion of examinees answering an item correctly), yet maintain a reasonably large grouping of items in each low, moderate, and high difficulty group. A summary of the cutoff p-values used to define low, moderate, and high difficulty items are contained in each of the item difficulty groupings is listed in Table 2.

Examinee Ability Groupings

To explore possible differences in erasure behavior across examinees, we grouped examinees by scores as part of the data analysis. Examinee ability groupings were defined as “low”, “medium”, and “high”. The groupings were defined by the lowest 25% of examinees, the middle 50% of examinees, and the upper 25% of examinees.

Item Position Groupings

To explore possible difference in erasure patterns across items at the beginning and end of the test, we identified the “first” and “last” items on the test. “First” items were defined by the first 10% of the items on the test. “Last” items were defined by the last 10% of the items on the test. Note that test lengths varied, so that the number of items included in these groupings differed across test subjects. In addition, first and last items were to a certain extent confounded with item difficulty because items at the beginning of the tend to be easier than those at the end of the test.

Indices

As mentioned above, wrong-to-right (WR) erasures are commonly used to study patterns of erasure. Three erasure indices based on WR erasures were calculated as part of this study. The first is the ratio of wrong-to-right erasures to total number of items (WRT), which is the number of wrong-to-right erasures divided by the total number of items. The second index is the ratio of wrong-to-right erasures to the number of correct responses (WRC), which is the number of wrong-to-right erasures divided by the total number of correct responses. The third index is the ratio of wrong-to-right erasures to the total number of erasures (WRE). Each of these indices can be interpreted as a proportion of wrong-to-right erasures: proportion of total (WRT), proportion of correct (WRC), or proportion of erasures (WRE).

Flagging

If the erasure indices are to be used to flag examinees, a criterion is needed to identify a cut-point separating examinees that will be flagged from those that will not be flagged. The goal is to identify a cut point that corresponds to unusual or aberrant erasures. Such a criterion is to a certain extent arbitrary, requiring judgment that would weigh the advantages and disadvantages of particular cut-offs. Such cutoffs are consequential because different cutoffs will lead to different examinees being flagged, particularly those examinees close to the cutoff.

We based the flagging criterion cutoff on the percentage of examinees falling outside of a particular value of the index. This cutoff is based on identifying the index value that would flag *x* percent of examinees (where we define *x* to be some value, such as 5). There are a number of advantages and disadvantages of this kind of flagging criterion. Several advantages include that (a) it is relatively simple to calculate, (b) it is relatively simple to explain, and (c) it doesn't require making particularly strong statistical assumptions. Disadvantages include that (a) the cutoff is ultimately dependent on the particular sample of examinees being considered; we will always flag a certain percentage of examinees and (b) we may not have precise control over the percentage of examinees flagged because the index values are not continuous (we will elaborate on this in the discussion).

While the simple flagging criterion we use has some limitations, we are suggesting that in practice such a criterion be used as an initial screening for unusual or aberrant erasures. In this case, we would screen some percentage of examinees with the most extreme index values to consider for further analysis or perhaps for further scrutiny using more sophisticated analysis.

We would need to define the percentage of examinees that would be screened; a decision which could vary, depending on the purposes for and implications of screening examinees.

We considered *at least* 5% or *at least* 1% levels for the flagging criterion. We are not suggesting that these are the best values, but they represent a starting point. Given our sample sizes of between about 3,000 to over 5,000 examinees, if we are able to flag roughly 5% and 1% of examinees, such criteria would screen roughly 30 to 250 examinees, depending on the criterion and the sample size.

Erasure analysis

We can break the erasure analysis implemented in this study into the following steps:

1. Obtain erasure results for each examinee, including erasure counts and WR erasure counts.
2. For each examinee, compute WRT, WRC, and WRE for each test subject.
3. Identify index values falling outside of 95% or 99% of the distribution for each index (i.e., identifying the 5% or 1% of examinees with the largest index values).
4. For each subject area (test form, etc.), summarize student flagging information.

Analysis

Due to the exploratory nature of this study, data analysis included primarily summary statistics (e.g., frequencies of values, means, and standard deviations) and plots of erasure counts and erasure indices.

Results

Do erasure behaviors differ across test forms (in the same subject)?

Table 3 contains a summary of the percentage of examinees for number of wrong-to-right (WR) erasures by test form and subject. Form 1 tended to have a larger percentage of examinees

with no WR erasures across test subjects, but all test forms had about 73% to about 85% of examinees with no WR erasures. Of those examinees with WR erasures, most had a single erasure and the percentages fall off dramatically for two and three erasures. The largest number of erasures for an examinee was observed on form 3, where one examinee had 16 erasures. Generally, however, few examinees had more than six erasures. These patterns are illustrated in Figures 1a to 1d. In addition, the ordering of the mean number of WR erasures across forms in Table 3 varied by subject area for forms 2 to 4. *Erasing did not generally differ much by form.* The only discernable pattern was that form 1 tended to have the relatively smaller mean percentage of WR erasures across subject areas compared to forms 2 to 4. Similar patterns were observed for the average total number of erasures by test form in Table 4. It is worth noting that the proportion of items answered correctly on form 1 was higher than the other forms (see Table 1). If we can assume that the examinee population was similar across forms, form 1 was slightly easier than the other forms, which may explain the slightly fewer erasures observed for form 1.

Do erasure behaviors differ across tested subject areas?

Table 3 and Figures 2a to 2d contain the percentages of examinees by number of WR erasures across test subjects for each of the four forms. Across all four test forms, the subject 1 test had larger percentages of examinees with WR erasures compared to the other subject areas. For three of the forms (2, 3, and 4) the subject 2 test had the largest percentage of examinees with no erasures. These differences in percentages of examinees were small but rather consistent across the four forms. The mean percentages of examinees with WR erasures across test subjects displayed at the bottom of Table 3 did not show a strong pattern across forms, although similar to the pattern observed in Figure 2, subject 1 tended to have relatively higher mean percentages of

WR erasures. Similar patterns were observed for the average total number of erasures by test subject in Table 4.

Do erasure behaviors differ between high versus low ability examinees?

Table 5a contains the average percentage of wrong-to-right erasures by test form and subject (Table 5b contains the standard deviations). On the subject 1 and subject 2 tests, the high ability examinees tended to have the largest average percentages of WR erasures, followed by the medium and then low ability examinees. On the subject 3 and subject 4 tests, the low ability examinees had the smallest percentage of WR erasures compared to medium and high ability examinees, and the medium and high ability examinees did not show a consistent ordering across forms.

Does a relationship exist between erasure behaviors and item difficulty/item characteristics?

Table 6 contains the average percentage of items with WR erasures by item difficulty level. For 12 of the 16 form/test subject combinations, the easy items had the smallest average percentages of WR erasures. These 12 form/test combinations included all subject 2 forms, three subject 1, three subject 3 forms, and two subject 4 forms. Table 2 lists the item difficulty cut-points based on classical test theory p-values that were used to define item difficulty groupings. Difficulty groupings were relatively consistent based on p-value, perhaps with the exception of form 1 in subject 2, where the easy and hard cut points were more than .1 higher than the other forms. This reflects the previous observations that form 1 appeared to be relatively easier than the other forms.

Table 7 contains the average percentage of items with WR erasures by item position. For forms 1 and 2, the last items had larger average percentages of examinees with WR erasures. However, forms 3 and 4 did not display this pattern.

What flagging index and flagging criteria are appropriate when erasure analysis is used on an individual level?

Tables 8 and 9 contain the observed percentages of examinees flagged under wrong-right/total (WRT) wrong-right/correct (WRC) and wrong-right/erasure (WRE) indices under 5% and 1% cutoffs, respectively. As we mentioned earlier, we may have difficulty controlling the percentages of examinees flagged because they depend on the distributions of indices that are not continuous. In Tables 8 and 9 we see that the actual observed percentages of examinees flagged may differ quite a bit from the specified cutoffs of 5% or 1%. These differences were particularly prevalent for WRT and WRE under the 5% cutoff. For WRT, the percentages of flagged examinees were from 5.3% and 15.3%. For WRE, the percentages of flagged examinees were from between 10.1% and 17.1%.

WRC tended to best reproduce the intended percentage of examinees defined by the cutoffs. This should not be interpreted as indicating that the WRC is a better or more accurate index; this result illustrates that WRC is behaving more like a continuous variable than WRT and WRE, and was better able to rank order each examinee on the index.

The 5% and 1% cutoffs for WRE were the same; both were defined by a cutoff of 1 or the case where all erasures led to wrong answers being changed to right answers. A relatively large percentage of examinees had index values of 1.

Tables 10a through 11b contain the averages and standard deviations of WR erasures for those examinees flagged by erasure indices (WRT, WRC, and WRE) under 5% and 1% cutoff rules. The average number of WR erasures is relatively small, from about 0.5 to about 2.8. In addition, the WRE average WR erasures for flagged examinees is consistently lower than the average WR erasures for WRT and WRC. Underlying this relatively low average is the fact that examinees tend to erase few items, and when they do erase they tended to erase a wrong response and change to a correct response. As mentioned above, the WRE cutoff was 1, which flagged examinees whose erasures all switched from wrong to right. We would expect that an examinee would have a better chance of correcting a single incorrect response after an erasure than correcting all of multiple erased incorrect responses.

Each of the erasure indices considered (WRT, WRC, WRE) treats wrong-to-right erasures in slightly different ways and therefore is likely to flag different examinees. Tables 12 and 13 display the percentage agreement of flagged examinees for pair-wise combinations of WRT, WRC, and WRE for 5% and 1% flagging cutoffs, respectively. For most forms/test subjects, WRT and WRC had the highest agreement (90% or more) and WRE tended to have somewhat lower agreement with WRT and WRC (roughly 80%-90% in most cases). Exceptions included form 1 in subject 2, subject 3, and subject 4, where WRE tended to have similar or higher agreement with WRT and to a lesser extent, WRC. This pattern is likely due in part to the pattern of fewer WR erasures observed for form 1. While the percentage agreement between erasure indices was generally relatively high, between roughly 80% and 90%, it implies that if we were to use a an index by itself, we would flag a slightly different set of examinees, depending on which index we used.

An alternative to a single index would be to consider multiple indices when screening examinees in an attempt to accommodate slight differences in flagging across indices. One option that we use here, is to track how many of the three erasure indices (WRT, WRC, and WRE) flag each examinee and screen examinees based on the number of indices flagged. Tables 14 and 15 contain the percentages of examinees flagged by 0, 1, 2, or all 3 erasure indices for 5% and 1% flagging cutoffs, respectively. As we have seen before, the vast majority of examinees (typically 80% or more) are never flagged. Across test subjects and forms from 1.8% (subject 2) to 17% (subject 1) were flagged by one index, from 0.8% (subject 1) to 11.5% (subject 2) were flagged by two indices, and from 0.1% (subject 2) to 3.3% (subject 4) were flagged by all three indices. Similar to determining the percentages for flagging cutoffs, determining the number of flagged indices to count as a screening cutoff would depend on how stringent or lenient we would want to be. Because erasing (WR erasures in particular) is relatively rare, the index values flagged in this study tend to flag examinees with relatively few erasures. This has implications for what we might consider as reasonable levels of erasure using the WRT, WRC, and WRE indices for purposes of flagging and screening examinees. We will elaborate on these implications in the discussion below.

Discussion

In this study, we addressed a number of questions regarding patterns of erasing behavior for individual examinees using data from four test forms of a large scale assessment program with tests in four subjects. Our analysis showed us that, at least for the testing program we considered, wrong-to-right (WR) erasures (and any erasures) were relatively rare events. Few examinees had more than six WR erasures on a particular test form. This is useful information for developing baseline expectations for patterns of erasures in a testing program. If we observe

examinees with substantial erasures compared to our baseline, we may want to investigate further. Likewise, a change in erasure behavior on a particular administration may warrant further investigation.

A second consistent finding was that harder items tended to show more WR erasures than easy items for the majority of forms/test subjects studied. It is not entirely clear whether this result is unique to the forms or particular tests we studied, but this pattern could be examined in future studies with additional test forms or with different testing programs. If this finding is consistent, we can expect difficult tests to display more erasures than easier tests.

We observed some variations in WR erasure behavior across the four studied test forms, with one form (that appeared easier) showing less erasing than the other three. Across test subjects, one of the test subjects (subject 1) tended to show more WR erasures than the other subject areas, and on three of the four forms the subject was the easiest test. Low ability examinees generally had relatively fewer erasures than medium or high ability examinees, although the differences were not large. Item position (i.e., first or last items) did not have a clear association with WR erasures, even though item position was confounded to some extent with item difficulty because items at the beginning of the test were easier than items at the end.

Another issue we attempted to address in this study is the extent to which erasure indices would be useful at the individual examinee level for purposes of flagging unusual erasing behavior. The three indices we studied, wrong-right erasures to total number of items (WRT), wrong-right erasures to total number correct (WRC), and wrong-right erasures to total erasures (WRE) are all based on WR erasures and tended to show relatively high agreement under 5% or 1% flagging criteria. However, WRT versus WRC tended to show higher agreement than WRT versus WRE or WRC versus WRE. Despite relatively high agreement between any two indices, a

certain percentage of examinees flagged under one index may not be flagged under another index, which is one of the reasons we also explored using flags from multiple erasure indices to screen examinee responses.

The WRT, WRC, and WRE indices are relatively simple to calculate and interpret, but the percentage of the most extreme erasure index values that we used as the criterion to flag examinees has disadvantages that should be considered if such a flagging criterion is to be used in practice. The first disadvantage is that the amount of erasure represented by the flagging criterion ultimately depends on the particular sample of examinees being considered and is defined by a percentage of these examinees, a percentage that must be specified. If examinees erase a lot (and erasures change wrong answers to right answers), we will flag a certain percentage of examinees with the most extreme index values; if examinees don't erase much (and erasures change wrong answers to right answers), we will still flag a certain percentage of examinees with the most extreme index values. If everyone cheated on the test in a way that led to erasures, we would only flag the examinees with the most extreme index values using this flagging criterion.

A second disadvantage, as we saw in our results (Tables 8 and 9), is that even if we define a flagging criterion based on a certain percentage of examinees, we will not have precise control of the actual percentage of examinees flagged because the erasure indices are not continuous variables; multiple examinees may obtain identical values on erasure indices, which may require identifying a larger or smaller percentage of examinees than specified by our flagging criterion. In our application of the flagging criteria, we specified that the percentage of examinees should be *at or above* 5% or 1%, and for WRT and WRE this often led to relatively larger than the nominal percentage defined by our flagging criterion. WRC flagged closer to 5%

and 1% of examinees because the index was based in part on the total number of items answered correct, which varied from examinee to examinee more than total number of erasures (included in WRE) or the total number of items (included in WRT). This led to clearer rank ordering of individual examinees. As mentioned above, it is not clear that WRC was a particularly better index, but this explains the variations in observed percentages of examinees flagged across indices.

The WRE index specifically has a disadvantage due to the fact that it is a ratio of the total number of erasures. Extreme values of the WRE index imply that a large proportion of the total number of erasures involved changing an answer from wrong to right. However, an extreme value of WRE can be obtained based on a few (or even one) erasure. Recall that the index value associated with 5% and 1% flagging criteria in our study was always 1, which indicates that to be flagged by WRE, an examinee must have obtained correct responses to *all* erasures that were originally incorrect. There may be a large number of examinees that erase a single response and change their answer to correct, which led to a WRE index value of 1 and a flag. While this may be an “extreme” value under our flagging criterion, it is not clear that this index value is particularly meaningful for purposes of flagging examinees, and it highlights a limitation in the WRE index. The index will always flag examinees erasing a single item with a correct answer. One way to avoid this limitation would be to adapt the WRE index and condition on the total number of erasures, or perhaps define some minimal number of erasures that would be deemed “meaningful” before a WRE would be flagged (a conditioning that could also be applied to WRT and WRC). This would require deciding what number of total erasures that would be deemed meaningful, which would have implications for which examinees are flagged.

Another way to avoid some of the limitations of individual erasure indices, and the fact that indices are likely to flag slightly different groups of examinees, would be to combine indices and screen examinees that have been flagged on multiple indices. We explored this possibility above, and pointed out that we would still need to decide how many flags would be required for purposes of screening an examinee. For example, if we use WRT, WRC, and WRE, should we require one flag? Two flags? Three flags? Part of the answer depends on our desired level of stringency or leniency. Part of the answer would also depend on the examinees and the amount of erasing that we observe compared to what would be judged meaningful for purposes of screening examinees. For example, in our study, due to the relatively rare occurrence of erasures, we might want to set fairly stringent criteria for screening examinees, but given the inherent limitations of the WRE index, we might not want to require that all three indices flag an examinee because such a screening procedure might tend to include more examinees with a single (or small number) of erasures.

As mentioned above, WR erasing behavior (and erasing behavior in general) observed here was relatively rare. Because our indices always target a certain percentage of examinees to flag, relatively rare erasing will lead to examinees being flagged with relatively small numbers of erasures. The reasonableness of the flags, or combination of flags, depends in part on our intended purposes. But on the face of it, although an erasure or two may be flagged as “unusual”, we should proceed cautiously before suggesting that such a result is meaningful for purposes of identifying erasing as evidence of cheating. A risk in applying erasure analysis would be to presume that an “unusual” or flagged examinee automatically indicates that someone cheated; such an interpretation should generally be avoided without additional evidence. As we have

indicated, perhaps the best use of this sort of erasure analysis at the individual examinee level is for screening purposes or for identifying examinees for further scrutiny.

We would recommend additional studies to examine refinements to the erasure indices and erasure analysis procedure. Perhaps one of the most important issues is defining what *meaningful* aberrant erasure behavior looks like and how erasure indices can be used to capture (or refined to capture) such behavior. What constitutes meaningful erasure behavior is likely to be specific to a particular testing program or application of erasure analysis. Another consideration would be to conduct analysis of some of the examinees with the most extreme values of WR erasures, WRT, WRC, or WRE. This might involve pulling out a handful of the most extreme cases and studying examinee response patterns more closely to determine if "rethinking" or "misalignment" could be ruled out as an explanation for the erasures. Additional refinements of erasure indices would be useful to explore, such as conditioning each index on the total number of erasures. Finally, the individual-level indices we explored could be aggregated to study potential group-level (e.g., classroom, school) differences in erasure patterns.

Aggregating the individual-level erasures indices, perhaps by comparing the number of examinees flagged in a particular group (e.g., school) to those in other groups (e.g., other schools in a state), would be an implementation of the individual-level analysis of erasures worth pursuing for situations where group-level irregularities are more likely. For example, such aggregation might involve comparing the proportion of flagged examinees in a particular group (e.g., school) to the proportion of all examinees flagged (e.g., an entire state).

References

Bellezza, F. S., & Bellezza, S. F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology, 16(3),* 151-155.

Cohen, A. S. & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Ed.), *Educational measurement* (4th Ed., pp. 17-64). Westport, CT: American Council on Education and Praeger.

Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics, 2,* 235.256.

Hanson, B. A., Harris, D. J. & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of copying.* ACT Research Report (87-15).

Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the k-index: Statistical theory and empirical support.* ETS Technical Report (96-4).

Mead, R., Andersen, K., & Korts, J. (2010). *Erasures and Rasch residuals.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Meijer, R. R. & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review and new developments. *Applied Measurement in Education, 8,* 261-272.

Meijer, R. R. & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25,* 107-135.

Primoli, V., Liassou, D., & Bishop, N. S. (2010). *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,* 9-16.

Severson, K. (2011, July 6). Systematic cheating is found in Atlanta's school system. *New York Times*, p. A13. Retrieved from http://www.nytimes.com/2011/07/06/education/06atlanta.html

Sotaridona, L. S. & Meijer, R. R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement, 40,* 53-69.

Sotaridona, L., van der Linden, W. J., & Meijer, R. R. (2006). Detecting answer copying using the kappa statistic. *Applied Psychological Measurement, 30,* 412–431.

van der Linden, W. J., & Jeon, M. (2012). Modeling answer changes on test items. *Journal of Educational and Behavioral Statistics, 37,* 180-199.

Winerip, M. (2011, July 31). Pa. joins states facing a school cheating scandal. *New York Times,* p. A11. Retrieved from http://www.nytimes.com/2011/08/01/education/01winerip.html.

Wollack, J. A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement, 21,* 307-320.

Table 1. Proportion of Items Answered Correctly by Test Subject and Test Form

Statistic	Subject 1				Subject 2				Subject 3				Subject 4			
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
mean	0.71	0.62	0.65	0.62	0.74	0.55	0.55	0.55	0.60	0.59	0.58	0.57	0.63	0.54	0.54	0.58
sd	0.19	0.19	0.18	0.18	0.19	0.19	0.18	0.18	0.21	0.20	0.18	0.17	0.17	0.16	0.15	0.16

Table 2. Item Difficulty (p-value) Cut-points Used to Define Item Difficulty Groupings

Item Difficulty Cut-point	Subject 1				Subject 2				Subject 3				Subject 4			
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
Hard	0.56	0.48	0.41	0.40	0.54	0.36	0.36	0.35	0.50	0.44	0.37	0.39	0.45	0.36	0.38	0.37
Easy	0.79	0.74	0.76	0.72	0.83	0.72	0.73	0.70	0.69	0.69	0.67	0.69	0.76	0.69	0.71	0.73

Table 3. Percentage of Examinees by Number of Wrong-to-right Erasures (WR)

Count	Subject 1				Subject 2				Subject 3				Subject 4			
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
0	82.3%	74.7%	75.4%	73.3%	85.6%	82.0%	82.3%	80.2%	85.8%	79.5%	72.6%	76.1%	84.8%	77.7%	76.5%	78.5%
1	12.3%	17.5%	17.5%	18.7%	11.2%	14.2%	13.5%	15.3%	11.0%	15.2%	18.7%	17.9%	11.5%	16.2%	17.4%	16.0%
2	3.4%	5.2%	4.9%	4.8%	2.2%	2.7%	3.3%	3.5%	2.4%	3.9%	5.9%	4.5%	2.6%	4.3%	4.4%	4.2%
3	1.1%	1.8%	1.4%	2.0%	0.7%	0.7%	0.8%	0.9%	0.5%	0.9%	1.9%	0.9%	0.9%	1.3%	1.4%	0.9%
4	0.5%	0.6%	0.6%	0.8%	0.1%	0.2%	0.08%	0.09%	0.2%	0.3%	0.6%	0.5%	0.2%	0.3%	0.3%	0.3%
5	0.3%	0.1%	0.08%	0.2%	0.07%	0.09%	0.03%	0.11%	0.03%	0.1%	0.3%	0.2%	0.00%	0.04%	0.2%	0.06%
6	0.00%	0.04%	0.1%	0.06%	0.00%	0.06%	0.03%	0.03%	0.00%	0.04%	0.1%	0.03%	0.03%	0.06%	0.06%	0.03%
7	0.03%	0.04%	0.00%	0.03%	0.03%	0.00%	0.00%	0.00%	0.03%	0.04%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
8	0.03%	0.00%	0.00%	0.06%	0.03%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.03%	0.00%	0.00%	0.00%
9	0.03%	0.02%	0.00%	0.00%	0.03%	0.00%	0.00%	0.00%	0.00%	0.00%	0.03%	0.00%	0.00%	0.00%	0.00%	0.00%
>=10	0.00%	0.02%	0.03%	0.00%	0.03%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%	0.00%	0.00%
mean	0.265	0.369	0.354	0.395	0.197	0.235	0.230	0.543	0.185	0.280	0.406	0.326	0.207	0.307	0.323	0.287
sd	0.707	0.768	0.773	0.806	0.603	0.583	0.559	0.977	0.522	0.650	0.803	0.685	0.568	0.686	0.684	0.632

Table 4. Total Erasures by Test Subject and Test Form

	Subject 1				Subject 2				Subject 3				Subject 4			
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
mean	0.44	0.68	0.67	0.69	0.35	0.55	0.51	0.54	0.35	0.54	0.70	0.65	0.36	0.56	0.70	0.53
sd	1.02	1.17	1.20	1.17	0.95	1.04	1.00	0.98	0.80	0.98	1.12	1.09	0.82	1.00	1.12	0.95

Table 5a. Mean Percentage of Wrong-to-right Erasures (WR) by Examinee Ability Level

Ability	Subject 1				Subject 2				Subject 3				Subject 4			
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
High	0.27	0.43	0.43	0.49	0.21	0.31	0.28	0.29	0.19	0.34	0.44	0.34	0.20	0.37	0.43	0.35
Medium	0.27	0.39	0.36	0.43	0.19	0.23	0.24	0.28	0.21	0.30	0.44	0.37	0.24	0.34	0.31	0.31
Low	0.22	0.26	0.27	0.23	0.14	0.18	0.18	0.20	0.15	0.17	0.29	0.19	0.10	0.17	0.22	0.17

Table 5b. Standard Deviation of the Percentage of Wrong-to-right Erasures (WR) by Examinee Ability Level

Ability	Subject 1				Subject 2				Subject 3				Subject 4			
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
High	0.69	0.83	0.85	0.98	0.62	0.69	0.62	0.69	0.53	0.72	0.87	0.72	0.55	0.81	0.83	0.68
Medium	0.75	0.79	0.79	0.82	0.61	0.58	0.55	0.60	0.55	0.68	0.83	0.73	0.63	0.70	0.65	0.67
Low	0.63	0.63	0.63	0.54	0.40	0.48	0.52	0.48	0.46	0.47	0.64	0.47	0.36	0.46	0.53	0.44

Table 6. Average Percentage of Items with Wrong-to-right Erasures (WR) by Item Difficulty

Item Difficulty	Subject 1				Subject 2				Subject 3				Subject 4			
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
Hard	0.29	0.60	0.53	0.62	0.59	0.48	0.37	0.42	0.55	0.82	0.89	0.65	0.51	0.46	0.55	0.66
Moderate	0.38	0.51	0.50	0.59	0.32	0.38	0.40	0.47	0.45	0.75	1.21	0.84	0.56	0.84	0.83	0.72
Easy	0.31	0.38	0.40	0.38	0.22	0.34	0.35	0.39	0.40	0.49	0.66	0.88	0.47	0.81	1.03	0.75

Table 7. Average Percentage of Items with Wrong-to-right Erasures (WR) by Item Position

Item Position	Subject 1				Subject 2				Subject 3				Subject 4			
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
First*	0.42	0.62	0.68	0.48	0.25	0.31	0.30	0.40	0.40	0.88	1.16	1.33	0.30	0.83	1.01	0.69
Last*	0.50	0.64	0.49	0.64	0.52	0.50	0.40	0.33	0.64	1.12	1.66	1.02	0.68	0.89	0.69	0.67

* First = first 10% of items, last = last 10% of items

Table 8. Observed Percentages of Examinees Flagged under 5% Cutoffs for Erasure Indices by Test Subject and Test Form

Flagging Index	Subject 1				Subject 2				Subject 3				Subject 4			
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
WRT	5.3%	7.8%	7.1%	8.0%	14.4%	18.1%	17.7%	19.8%	14.2%	5.3%	8.7%	6.1%	15.3%	6.0%	6.2%	5.5%
WRC	5.2%	5.0%	5.1%	5.1%	5.1%	5.1%	5.3%	5.4%	5.0%	5.1%	5.0%	5.2%	5.3%	5.1%	5.1%	5.4%
WRE	12.2%	15.4%	14.3%	17.0%	10.6%	11.0%	11.3%	12.4%	10.1%	13.4%	17.1%	14.3%	11.1%	14.9%	14.5%	14.4%

Table 9. Observed Percentages of Examinees Flagged under 1% Cutoffs for Erasure Indices by Test Subject and Test Form

Flagging Index	Subject 1				Subject 2				Subject 3				Subject 4			
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
WRT	2.0%	2.6%	2.2%	1.1%	1.0%	1.1%	4.2%	1.1%	3.2%	1.4%	2.8%	1.6%	1.2%	1.7%	1.8%	1.3%
WRC	1.0%	1.0%	1.0%	1.2%	1.1%	1.0%	1.1%	1.0%	1.3%	1.1%	1.1%	1.0%	1.0%	1.1%	1.2%	1.1%
WRE	12.2%	15.4%	14.3%	17.0%	10.6%	11.0%	11.3%	12.4%	10.1%	13.4%	17.1%	14.3%	11.1%	14.9%	14.5%	14.4%

Table 10a. Average Number of WR Erasures for Examinees Flagged under 5% Cutoffs for Erasure Indices by Test Subject and Test Form

Erasure Index	Subject 1				Subject 2				Subject 3				Subject 4			
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
WRT	2.66	2.50	2.51	1.30	1.36	1.30	1.30	1.30	1.31	2.42	2.51	2.42	1.35	2.40	2.42	2.33
WRC	2.54	2.63	2.57	1.88	2.04	1.89	1.89	1.92	1.81	2.17	2.79	2.33	2.01	2.32	2.37	2.15
WRE	1.32	1.33	1.33	1.25	1.24	1.25	1.25	1.23	1.28	1.31	1.40	1.27	1.28	1.32	1.28	1.28

Table 10b. Standard deviations of WR Erasures for Examinees Flagged under 5% Cutoffs for Erasure Indices by Test Subject and Test Form

Erasure Index	Subject 1				Subject 2				Subject 3				Subject 4			
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
WRT	1.14	0.94	1.18	0.61	0.96	0.70	0.61	0.65	0.68	0.86	0.92	0.81	0.75	0.95	0.77	0.68
WRC	1.28	1.20	1.45	0.81	1.40	1.04	0.81	0.92	0.93	1.09	1.14	1.01	0.97	1.14	0.97	0.86
WRE	0.76	0.64	0.66	0.53	0.57	0.62	0.53	0.51	0.66	0.69	0.79	0.60	0.60	0.78	0.62	0.61

Table 11a. Average Observed Number of WR Erasures Examinees Flagged under 1% Cutoffs for Erasure Indices by Test Subject and Test Form

Flagging Index	Subject 1				Subject 2				Subject 3				Subject 4			
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
WRT	3.77	3.50	3.66	4.56	4.00	3.52	2.26	3.37	2.36	3.62	3.56	3.61	3.39	3.43	3.41	3.39
WRC	4.16	3.61	3.89	3.81	3.31	2.91	2.56	2.89	2.50	3.17	3.78	3.47	3.03	3.31	3.29	3.03
WRE	1.32	1.33	1.33	1.39	1.24	1.25	1.25	1.23	1.28	1.31	1.40	1.27	1.28	1.32	1.28	1.28

Table 11b. Standard Deviation of Observed Number of WR Erasures Examinees Flagged under 1% Cutoffs for Erasure Indices by Test Subject and Test Form

Flagging Index	Subject 1				Subject 2				Subject 3				Subject 4			
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
WRT	1.23	1.07	1.62	1.07	1.02	0.86	0.60	0.79	0.79	0.97	0.98	0.78	0.99	1.32	0.78	0.68
WRC	1.68	1.80	2.42	1.52	2.29	1.35	0.99	1.17	1.24	1.49	1.47	1.23	1.36	1.83	1.22	1.17
WRE	0.76	0.64	0.66	0.77	0.57	0.62	0.53	0.51	0.66	0.69	0.79	0.60	0.60	0.78	0.62	0.61

Table 12. Percent Agreement* between erasure indices under 5% Cutoffs by Test Subject and Test Form

Erasure Indices	Subject 1				Subject 2				Subject 3				Subject 4			
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
WRT vs. WRC	98.4%	96.1%	96.5%	96.2%	90.6%	87.0%	87.6%	85.6%	90.8%	97.2%	95.8%	97.5%	90.0%	97.6%	97.6%	98.0%
WRT vs. WRE	88.0%	84.6%	85.7%	83.8%	96.2%	93.0%	93.6%	92.6%	95.9%	87.3%	83.4%	85.6%	95.9%	86.0%	85.5%	86.5%
WRC vs. WRE	88.0%	83.8%	85.2%	82.7%	90.5%	88.5%	88.6%	87.4%	91.3%	86.9%	82.6%	84.7%	90.2%	85.7%	84.8%	86.0%

* Percentage of cases where indices had identical flags.

Table 13. Percent Agreement* between erasure indices under 1% Cutoffs by Test Subject and Test Form

Erasure Indices	Subject 1				Subject 2				Subject 3				Subject 4			
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
WRT vs. WRC	98.7%	98.0%	98.4%	98.9%	99.1%	99.0%	96.7%	99.0%	97.7%	98.9%	97.9%	98.9%	99.3%	98.8%	98.6%	99.1%
WRT vs. WRE	87.4%	84.1%	85.4%	82.9%	89.3%	89.0%	89.2%	87.4%	90.8%	86.6%	82.8%	85.2%	89.0%	85.2%	85.0%	85.6%
WRC vs. WRE	87.2%	83.9%	85.3%	82.6%	89.3%	88.5%	88.4%	87.2%	90.1%	86.6%	82.6%	85.3%	89.0%	85.0%	85.0%	85.8%

* Percentage of cases where indices had identical flags.

Table 14. Percentages of Examinees Flagged out of Three Erasure Indices* under 5% Cutoffs by Test Subject and Test Form

Number of Flags*	Subject 1				Subject 2				Subject 3				Subject 4			
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
0	85.0%	80.4%	81.7%	79.2%	85.6%	82.0%	82.3%	80.2%	85.8%	83.6%	78.8%	82.1%	84.8%	82.2%	82.1%	82.8%
1	9.5%	12.9%	11.9%	13.6%	1.8%	4.2%	3.7%	4.6%	2.2%	10.9%	13.8%	12.0%	2.1%	11.9%	11.9%	11.4%
2	3.3%	4.9%	4.4%	5.0%	9.5%	11.5%	11.4%	12.6%	7.8%	3.5%	5.3%	4.1%	9.8%	3.5%	4.2%	3.4%
3	2.2%	1.8%	2.0%	2.1%	3.1%	2.3%	2.6%	2.6%	3.2%	2.0%	2.1%	1.8%	3.3%	2.4%	1.8%	2.4%

*WRT, WRC, and WRE.

Table 15. Percentages of Examinees Flagged out of Three Erasure Indices* under 1% Cutoffs by Test Subject and Test Form

Number of Flags*	Subject 1				Subject 2				Subject 3				Subject 4			
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
0	86.5%	82.9%	84.3%	82.0%	88.5%	88.1%	86.8%	86.7%	88.7%	85.7%	81.3%	84.4%	88.2%	84.2%	84.1%	84.8%
1	12.0%	15.2%	14.0%	17.0%	10.4%	10.9%	10.1%	12.1%	8.5%	13.0%	16.6%	14.4%	10.6%	14.3%	14.4%	14.0%
2	1.4%	1.8%	1.5%	0.8%	0.8%	0.8%	2.7%	1.1%	2.2%	0.9%	1.8%	0.9%	0.8%	1.2%	1.3%	0.8%
3	0.1%	0.1%	0.2%	0.2%	0.3%	0.2%	0.4%	0.1%	0.6%	0.3%	0.3%	0.3%	0.4%	0.3%	0.2%	0.4%

*WRT, WRC, and WRE.

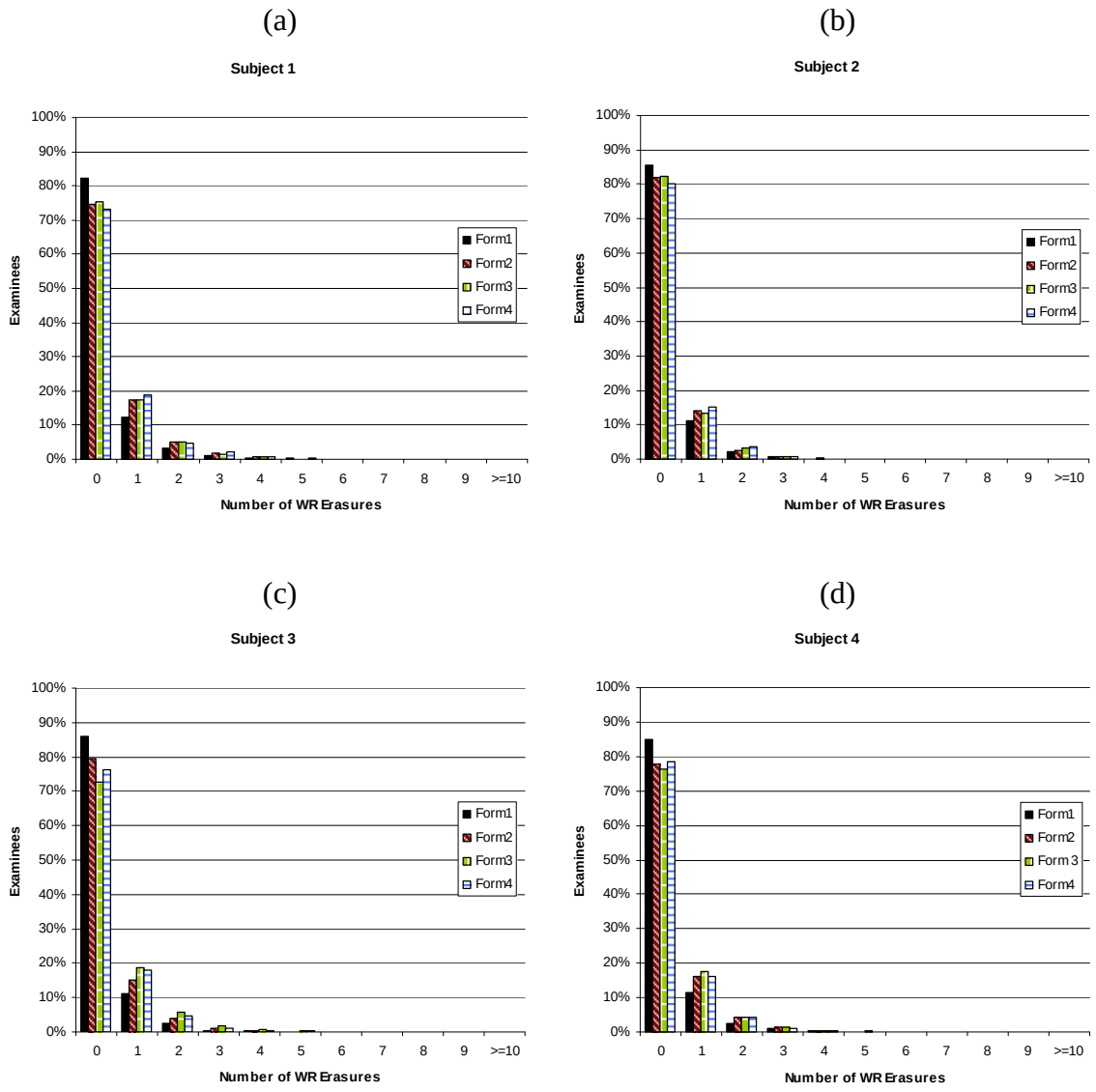

Figure 1. Percentage of examinees as a function of number of wrong-to-right erasures for four test forms across four test subjects.

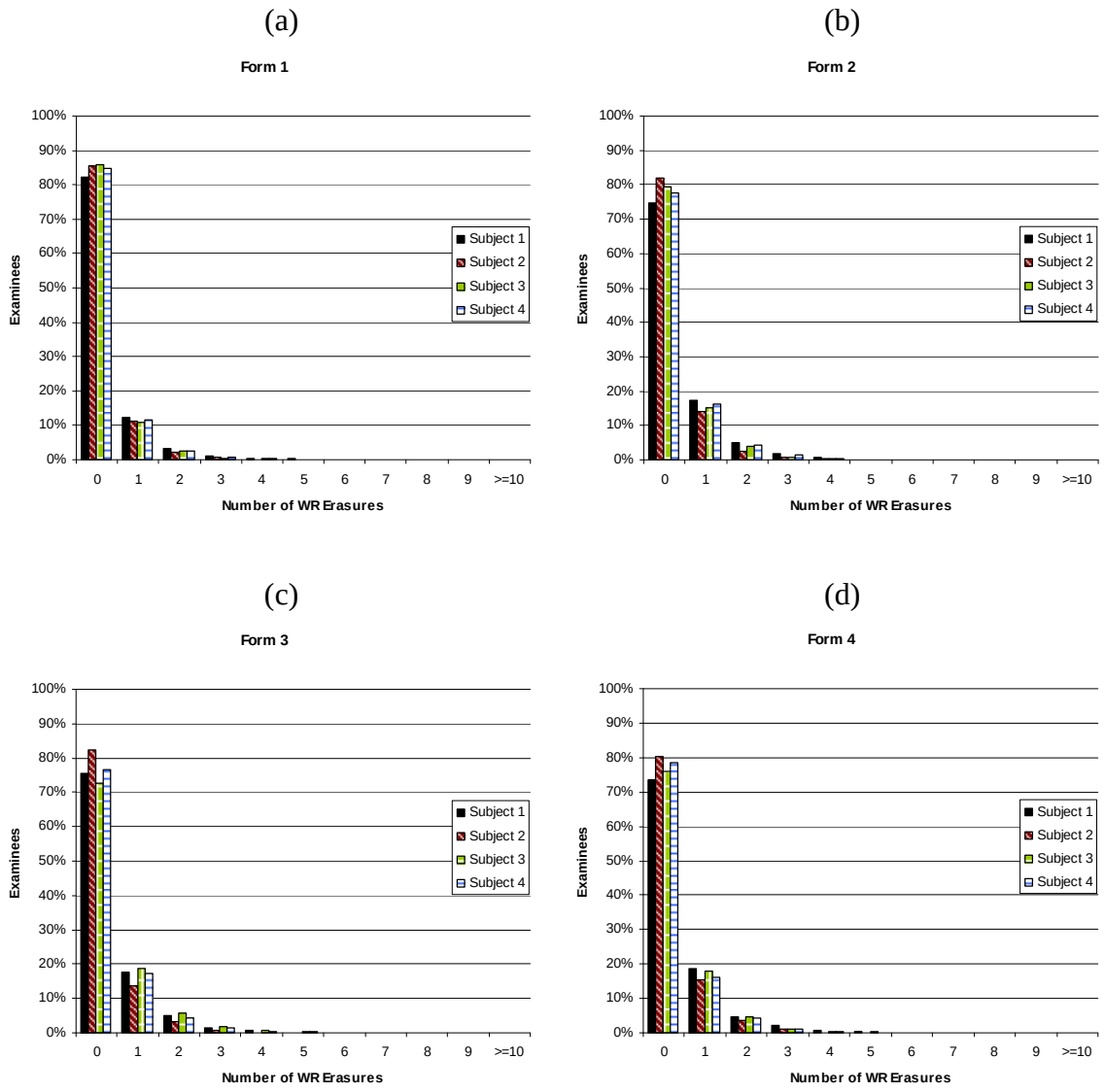

Figure 2. Percentage of examinees as a function of number of wrong-to-right erasures for four test subjects across four test forms.