Title: Cheating on Tests: A Threat to Response Validity

Mark Albanese, PhD[1,2] and James Wollack, PhD[2]

Presented at the 2013 Statistical Detection of Potential Test Fraud Conference
Madison, Wisconsin
October 19, 2013

Cheating on standardized tests and what is done about it is one of the dark secrets of standardized testing. Many of the methods for detection and the aftermath tend to be either proprietary and/or confidential. Research in the published literature is generally about different indexes of "unusual similarity" and often employs simulated data to show the effectiveness of the different indexes for detecting various patterns of similarity. The *Standards for Educational and Psychological Testing* (1999) primarily deal with cheating in standards 8.7-8.13. Standard 8.7 calls cheating fraud when it involves someone else impersonating the nominal test taker. In later standards in the section, cheating is subsumed under "testing irregularities" and "suspected misconduct."

For psychometricians, one of the most critical questions is how does cheating impact on test scores and how should it be treated. Standard 8.7 states that "The validity of test score interpretations is compromised by inappropriate test disclosure."(p. 88) However, it is not just the interpretations of scores that are impacted by disclosure of confidential testing material, fraudulent test takers or other types of cheating. Cheating compromises the entire standardized testing process. Scores do not accurately reflect the examinee's abilities and their score profile can be disjointed if they copy answers from a portion of the examination from another examinee. A successful cheater will have scores that exceed their ability and will be ranked higher than examinees who are of higher ability, but did not cheat. The cheater who copies on a portion of the examination, will produce a disjointed score profile where they answer some items correctly (where they cheated) that are beyond their ability. This will make items appear to be less homogeneous than they may actually be which will reduce the test internal consistency reliability estimate. Further, the examinee scores will be higher than they should be, which will disrupt correlations with other types of academic performance, should they be used for establishing the validity argument for scores.

In this paper, we propose that cheating should be treated as a threat to response validity. In this conceptualization, the validity argument approach advocated by Kane, Crooks and Cohen (1999) would be used to make four levels of action. The first level action would be to purge the responses from the test data-base. The second level action is whether there is sufficient evidence of cheating to launch a "forensic" search of the evidence. A forensic search would build the validity argument for the higher level actions. The third level of action is whether to withhold reporting scores to the examinee or their designee(s). The fourth, and final, level of action is whether the validity argument is sufficient to pursue charges of fraud, copyright infringement or other serious allegations. In the framework of cheating being a threat to response validity, the

[1] National Conference of Bar Examiners

[2] University of Wisconsin-Madison

actions taken are a function of the strength of the response validity (or invalidity) argument that is compiled.

A level 1 action is to remove the examinee's responses from the data-base used to calibrate item parameters and equate and scale the scores. Invalid responses will add error into the estimates of the various item parameters as well as to the equating links. Cheating will have a similar disruptive effect on classical item and test statistics such as the difficulty and discrimination indices and the raw to scale score conversions. The different classical equating approaches such as mean and equal percentile equating will also be impacted. The lower asymptote in the three parameter model may be the most sensitive parameter to cheating since cheating is likely to be more prevalent among examinees in the lower ability levels. Cheating may be a factor in the difficulty encountered in obtaining stable estimates of the lower asymptote. Thus, one might see more stable estimates for the lower asymptote if examinees with even low levels of likely cheating were removed from the calibration and equating process (smaller standard errors and convergence with fewer iterations). A level 1 action could follow from the screening for evidence of irregularities that most testing agencies engage in as general practice or from some type of report of unusual events. Common screening approaches involve proctor reports of irregularities or from examinees who report the irregular behavior of other examinees. A level 1 action could also come from data-mining test results for unusual similarity among responses or other unusual response patterns.

A level 2 action is to launch a forensic search for sufficient evidence of response invalidity that further action should be taken. A forensic search generally follows from the screening for evidence of irregularities described above (proctor reports and data-mining). Additional common screening approaches that could yield a forensic search involve screening web-sites and email or other communications for live test material. Searching web-sites and email communications for live test material or copyright breaches and data-mining for unusual similarity can be time consuming and complex. Some testing agencies employ security companies that specialize in such screening activities. If the argument is sufficiently strong that the responses of a particular examinee are invalid, one or more higher levels of action can be taken.

A level 3 action would be to withhold scores. This would take a stronger validity argument than either a level 1 or 2 decision. It also invokes standards 8.11-8.13 that prescribe timely communication with the examinee about the action being taken and providing them with a means of challenging the action. Whereas a level 1 or 2 action could be supported by data-mining for unusual similarity in responses, a level 3 action would generally need to be bolstered by more than analysis of data, such as visual reporting of copying behavior by proctors or violating the conditions of testing (e.g., bringing a cell phone into the testing room). It is the strength of the response validity (or invalidity) argument that determines whether a level 3 action is taken.

A level 4 action would be to make allegations of fraud, copy right infringement or some other serious breach of ethical behavior by the examinee. This level of action is beyond the scope of further discussion in this paper.

The question arises, then, what type of evidence would yield a validity argument supporting a level 1 decision to remove examinees from the data-base. The consequences of removing examinee data from the data-bank are relatively minor. No examinee is really put at risk by this action as long the identity of examinees whose scores are removed are maintained in confidence. Further, as long as there are over 500 examinees left after removing the suspected cheaters, the item parameters will be adequately estimated. Because the risks are relatively minor, the strength of response validity evidence supporting a level 1 action could be relatively low. In fact, they could be driven by data-mining in which similarity statistics are computed for each examinee and outliers detected and removed. The problem with data-mining is that the methods are not well-developed for this type of action and they can be complex and yield a large number of 'hits' from purely random factors. The accumulative effect of such type 1 errors (probability of concluding unusual similarity between two examinees when there is no difference) would be to remove a lot of examinees whose scores relate for purely random reasons. If one does a comparison of response profiles for all possible pairs of 1,000 examinees, the result is 499,500 comparisons. If one uses a 5% probability as warranting a level 1 action, there would be 24,975 comparisons that would be significant, leading to removal of potentially the large majority of the 1,000 examinees if not all. A more practical strategy would be to use a Bonferroni-type correction to establish a fixed criterion as a threshold for evaluating whether or not to eliminate an examinee from the data-base. For example, one could use the following: $p'=2p/(n-1)$, where p'=the probability of a given pair of examinee responses being as similar by chance, p=overall target proportion of the examinees estimated to be submitted to a forensic analysis and n= the number of examinees. If we use a target of approximately 5% of the examinees, an examination given to 1,000 examinees would give a threshold $p'=0.0001$. Thus, in a data-mining undertaking where all pairs of examinees were evaluated for unusual similarity, one would not eliminate an examinee unless one or more of their pairwise comparisons had responses so similar that it would occur only 1 time out of 10,000 (or more). This would result in approximately 50 examinees out of a 1,000 having their responses removed from the data-base.

If data-mining were the sole source of the evidence for response validity, it would be critical to document the criteria applied and the number of examinees who met the criteria and were omitted from the data-base. In cases of small testing programs such as subspecialty examinations in medicine, omitting examinees could push the limits on the minimum numbers of examinees needed to provide viable estimates of item parameters. However, it would be better to have estimates that are poorly estimated from too few examinees with clean scores than to increase their number by adding estimates from scores contaminated by cheating.

What type of validity argument, then, would lead to a level 2 action of a forensic analysis of the available response validity data? The main deterrence to doing a forensic study is the cost. A forensic analysis would involve computing much more complex statistics such as Omega (Wollack, 1997) or some of the more complex Bayesian estimators. In some cases, an expert analyst will be commissioned to undertake the level 2 analyses. Whether to take a level 2 action depends upon the type of evidence and the potential damage that the particular irregularity might inflict. Even the hint of exposure of one or more test forms should yield a forensic study because the stakes are so high and the potential loss so great. The use of the data for determining whether to conduct a level 2 action should generally involve a more targeted consideration than the data-mining operation describe above. For example, seating charts could be used to have each

examinee have their responses scored against the responses of every examinee taking the test within viewing distance. Examinees whose scores using their proximal peers as the key are higher than using their actual key could be targeted for a level 2 action. Other similarity indexes could also be used such as the longest sequence of identical responses or the number of identical incorrect responses. These different similarity indexes could be computed for each examinee and triangulated to determine if a consistent and compelling validity argument arises.

The strength of the validity argument needed to support a level 3 action, where the score is withheld, would require a more demanding level of evidence and preferably multiple types of evidence. Data-mining might initiate an investigation of a particular examinee, but action taken to delay or withhold a score would generally require the data-mining evidence to be far more compelling and there would generally need to be additional supporting evidence. Whereas data-mining evidence showing similarity in data between a suspected cheater and source that would occur only 5% in a base-line data set in combination with multiple indexes could support a level 2 forensic analysis, a level 3 action might demand a 1% occurrence. Examples of the additional evidence beyond response similarity might be a seating chart showing the examinee suspected of cheating and the suspected source(s) were within proximity and visual sight-lines and/or proctor reports in which the examinee suspected of cheating was looking at the exam sheet/computer screen of suspected source(s).

Importance

Cheating incidents on standardized tests are usually treated confidentially and in a relatively adversarial manner. Even the methods used to detect cheaters tend to be proprietary or company secrets. What gets published are studies of different types of indexes of similarity and their sensitivity to different types of response patterns one would expect if an examinee were cheating. The process by which these indexes are used in the screening and detection of cheating in operational tests are rarely described. The stakes are high and it is not unusual for examinees to challenge the allegations of cheating. As a consequence, allegations of cheating are rarely pursued without overwhelming evidence of unusual behavior on the part of the examinee and data from test score analysis that find similarities beyond the realm of reasonable chance probabilities. Further, even though low levels of cheating (copying on 5-6 items on 100 item test) on a wide scale could compromise the validity of scores for many uses, it generally is not evaluated in that manner. The statistical methods are not sensitive to low levels of cheating, particularly if they are wide-spread. It takes an extensive amount of copying behavior for the statistics to indicate unusual similarity. And, because of the adversarial and sensitive nature of cheating allegations, they generally are not pursued unless they are egregious. So, it is likely that a substantial amount of low level cheating goes on without detection because we have neither the means nor the will to do so.

If we consider cheating a matter impacting on the response validity of the examination, it changes the nature of the process. The process moves from being an action taken against individuals and the potential backlash that such action can yield to developing an argument for whether the responses should be pulled from the data-base. This is a low stakes decision that merits the development of tools to detect low levels of cheating and their impact on test analysis

data and other aspects of the test development process. In the process of developing that argument, it may come to light that a higher level action is merited, but that would not be the focal point. Processes for confirming response validity could become standard practices that would be incorporated into the data that would be a by-product of normal test analyses, much like difficulty and discrimination-type data are done currently.

The other part of considering cheating as a threat to response validity is that it enables the validity argument to be crafted to the situation. As technology expands the ways in which examinees can find creative ways to an item answer besides knowing it, test developers need to have creative means of detecting them. The ability to have a validity argument in their arsenal will help to level the playing field.

References

JOINT COMMITTEE ON STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING OF THE AMERICAN EDUCATIONAL RESEARCH ASSOCIATION (AERA), THE AMERICAN PSYCHOLOGICAL ASSOCIATION (APA), AND THE NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION (NCME), STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING (AERA, 1999).

Kane, M., Crooks T., & Cohen, A, Validating measures of performance. *Educational Measurement: Issues and Practice.* 1999(summer), 5-15.

Wollack, J. A. (1997). A nominal response model approach for detecting answer copying.*Applied Psychological Measurement*, 21(4), 307-320.