Statistical Detection of Cheating

Thomas O'Neill
Jerry Gorham
Michelle Reynolds
Weiwei Liu

Introduction & Rationale

Test publishers and testing services are always concerned with irregular behavior, especially with regard to how it will impact the reliability and validity of the test scores. Irregular behavior can range from minor deviations from the standardized conditions to more significant types of behavior, such as, copying from another test-taker, accessing unauthorized resources (books, notes, cheat sheets, etc.), or attempting to steal the item content. Although test publishers are quite vigilant, it is difficult to catch people doing something that they wish to keep secret. Therefore test publishers and test services typically employ a battery of procedures to discourage the unauthorized behavior and facilitate the detection of violations.

Using the cheating detection classification categories, observational and statistical, as proposed by Cizek (1999), it seems that the greatest successes in proving cheating have come from the observational camp (Dwyer & Hecht, 1994). Observational procedures include training test center staff to be on the look out for violations of the standardized procedures, such as proxy testers, accessing unauthorized materials during the test, and removing/recording item content. These processes can often be enhanced through the use of recording technology. Examinees can be video taped while they test; they can be fingerprinted and photographed upon sign-in. When the observational detection and documentation is done well and assuming there has been the appropriate legal language regarding the conditions under which the examinee agrees to test, it provides very powerful evidence that permits test publishers to take action against cheaters. Nevertheless, even the best training, the best check-in procedures and the best recording of the testing session can be defeated by a proctor that is working against the security procedures. Less nefariously, it can also be that some of the irregular behavior is going on outside of the test center, which is not under the test center staffs' control. Accessing "stolen items" before the test or "brain-dumping" items after the test would fall into this category.

Most of the literature regarding statistical approaches to detecting cheating has focused on comparing the performance of pairs of examinees. Bird (1927, 1929) advocated the comparison of the number of common errors for a pair of examinees to an expected number of common errors based upon the rest of the sample. Since then, extensions, refinements, and variations of this approach to detecting cheating have been proposed (Crawford, 1930; Angoff, 1974; Frary, 1992; Sotaridona & Meijer, 2001a; Sotaridona & Meijer, 2001b; van der Linden & Sotaridona, 2002). Similar procedures have even been mentioned in recent popular literature such as Freakenomics (Levitt & Dubner, 2005).

Although examples in the literature on procedures comparing specific pairs of examinees or searching data sets for pairs that are suspiciously similar are not uncommon in the literature, it does seem strange that there has been very little discussion regarding the pass rates for test centers and comparing them to some expectation. The reason that this is so strange is because it is a very common practice in the industry and it has been for a very long time.

Establishing Thresholds:

Deviation among test centers. When a test is being given for the first time, there is no previous experience to guide the test publisher's expectations with regard to expected pass rates. Nevertheless, one could assert that there are no regional differences and the expected pass rate is merely the mean pass rate of all the test centers. Therefore, one could set some threshold; say 1.96 standard deviations above the mean pass rate as a flagging criterion for further scrutiny. However, it seems reasonable that there could be regional differences in the quality of candidates. If these regional differences are large and the standard deviation of the pass rates is large, then small disturbances in the pass rate for a particular center would be difficult to detect.

Deviation over time. When a test has been administered for several years at the same test centers, it becomes possible to use the previous pass rate(s) of that test center as the expected pass rate. This approach removes the issues related to region differences, but it would not be able to detect if there was cheating that was constant over several years.

Changes in pass rate for a test center could be caused by changes in the demographics of the population that tests at that site, improvements in the educational programs in the area surrounding the test center, or even the time of year. Alternatively, a more sinister mechanism could also be at work. There could be problems related to a specific proctor or examinees in the region have been harvesting items and sharing that repository of items with new examinees.

Methods

Aberrant Test Center Pass Rates

Just as detection of cheating is important at the candidate level, the same is true at the test center level. The effects of a group of individuals conspiring to cheat could possibly begin to be seen at the test center level. The criterion used here to detect significant changes at test centers is the aggregate pass rate at each test center. By compiling pass rates by test centers, we can flag any centers with changes that differ significantly from the population.

Center Deviation from Annual Mean. In this procedure, the mean and standard deviation are calculated for the entire population of test centers over a specified period of time. Two thresholds, 95%, (1.96 * SD) and 80% (1.282 * SD) are determined. Those centers with pass rates greater than the population mean plus the threshold are flagged for further review. Test centers with changes in the negative direction are not flagged at this time, due in part to the fact that if a group's cheating efforts result in fewer people passing, that is not of concern. The responsibility of protecting the public refers only to those who pass as those who fail are deemed not minimally competent, therefore are not allowed to serve the public.

Center Deviation from Mean Change in Test Center Pass Rate. In this procedure, the mean and standard deviation of the difference between the 2004 and 2005 test center pass rates are calculated. Two thresholds, 95%, (1.96 * SD) and 80% (1.282 * SD) are determined. Those centers with pass rates greater than the population mean difference plus the threshold are flagged for further review. Test centers with changes in the negative direction are not flagged.

Center Deviation from Previous Center Pass Rate. In this procedure, the standard error of the pass rate is computed for each test center for both years and then a joint standard error is computed. Next, the mean pass rate (unweighted) is compared with the more recent pass rate. If the difference between the mean pass rate and the recent pass rate is larger than 95% (1.96 joint standard errors) and 80% (1.282 joint standard errors), the center is flagged for further review. This procedure permits each test center to act as its own control and permits the decisions to be based upon the sample size available at each center.

Results

Deviation from 2005 Mean Pass Rate. Currently, there are 206 test centers that administer the NCLEX exam. In January 2005, three international locations were added as well as one new domestic center in late November 2005. The mean pass rate by test center for 2005 was 75.44% (SD 11.08%). When using the 80% threshold 1 test center was flagged and one additional test center was flagged for having a pass rate above the 95% threshold. Its important to note that the test center flagged for the 95% threshold, did not start administering exams until late November 2005, thus having a low number of candidates test at that location could have contributed to the unusually high average pass rate. (See Table 1, Map1)

Deviation from 2004 Mean Pass Rate. Additionally the mean pass rate for test centers in 2004 (N = 202) was calculated to be 73.03% (SD 10.56%). No test centers were flagged for having average pass rates above the 95% threshold, and one was flagged for being greater than the 80% threshold. (See Table 1, Map 2)

Center Deviation from Mean Change in Test Center Pass Rate. While looking at each of the test centers as a population over a specified amount of time, did not indicate results other than what is to be expected, it

seems necessary to consider each individual test center's change in pass rates to see if anything unexpected is occurring. Perhaps a test center has historically had higher pass rates due to such differences as educational programs in the area. Investigating the average change in pass rate from 2004 to 2005 resulted in flagging a total of 18 test centers. The average change in pass rates for all test centers was 2.48% (SD 3.40%). Four test centers were flagged for having increases in pass rates above the 95% threshold (9.15%), and an additional 14 were flagged for changes above the 80% threshold (6.84%). (See Table 1, Map 3)

The two test centers flagged for the 80% threshold in 2005 and 2004 are in close proximity in Ohio. However, these two test centers where not flagged for a pass increase at either the 80% or 95% threshold when comparing the difference between 2004 and 2005 pass rates. This indicates that these test centers may on average have a higher pass rate than the population of test centers, thus have a higher baseline for which to make comparisons in changes to their pass rates.

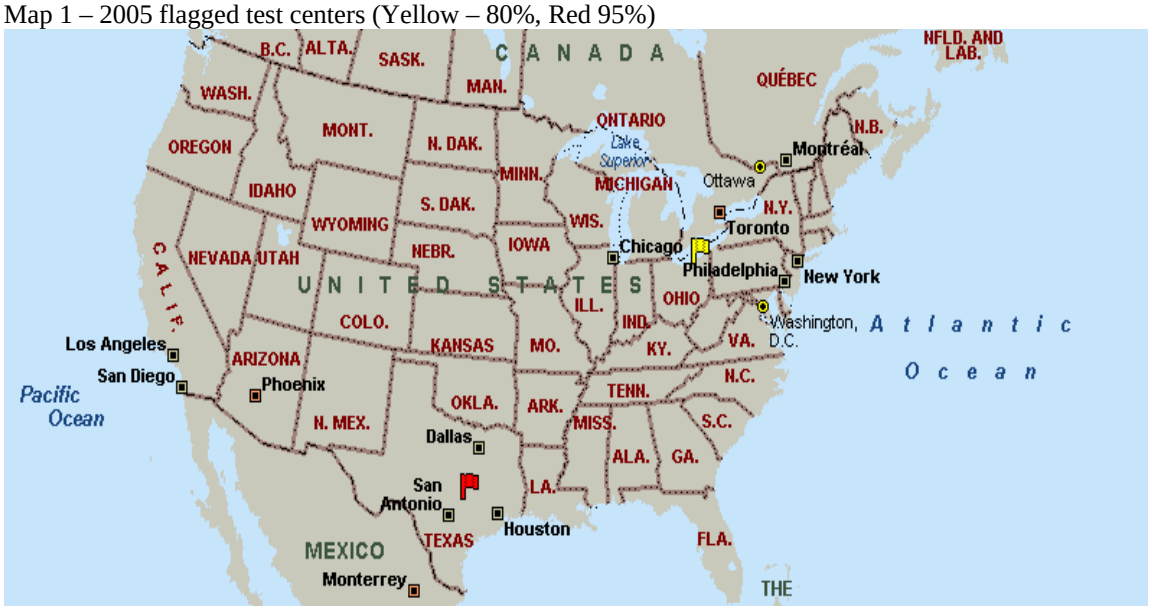
Center Deviation from Previous Center Pass Rate. This procedure identified 18 centers at the 80% confidence level and another 1 at the 95% confidence level. These results looked somewhat similar to the mean difference method. (See Table 1, Map 4)

Table 1. Test Center Pass Rates Statistics

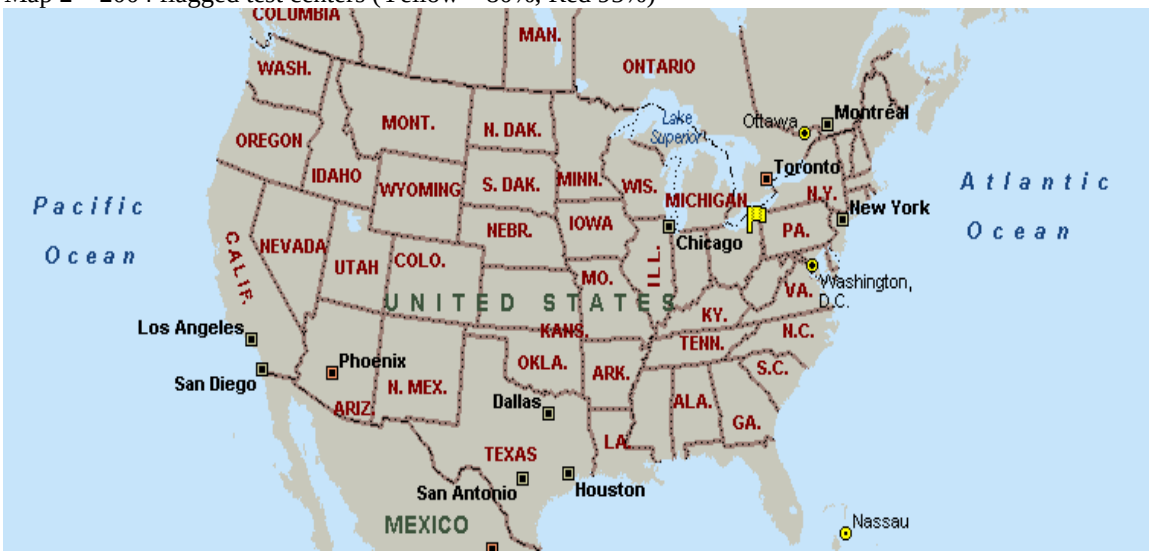
				95% Confidence Threshold (μ + 1.96 SD)		80% Confidence Threshold (μ + 1.96 SD)	
YEAR	# of test centers	Average Test Center Pass Rate	SD	Pass Rate Threshold*	# Flagged	Pass Rate Threshold*	# Flagged
2005	206	75.44%	11.08%	97.16%	1**	89.64%	1
2004	202	73.03%	10.56%	93.73%	0	86.57%	1
Difference of (2005 - 2004)	202	2.48%	3.40%	9.15%	4	6.84%	14
Difference using Joint Standard Error	202	-----	-----	-----	1	-----	18

* Test centers flagged for 80% were not flagged for 95% but meet criteria for 80%.
** This test center did not open until Nov. 2005 and only had 7 candidates test. Its pass rate for 2005 may be inflated due to a small number of candidates testing.

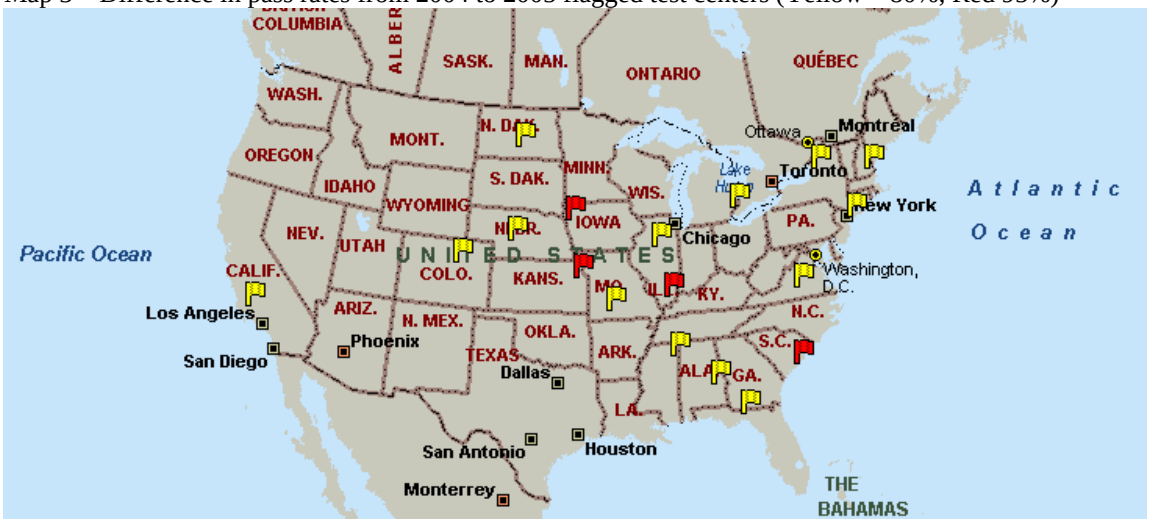
Map 1 – 2005 flagged test centers (Yellow – 80%, Red 95%)



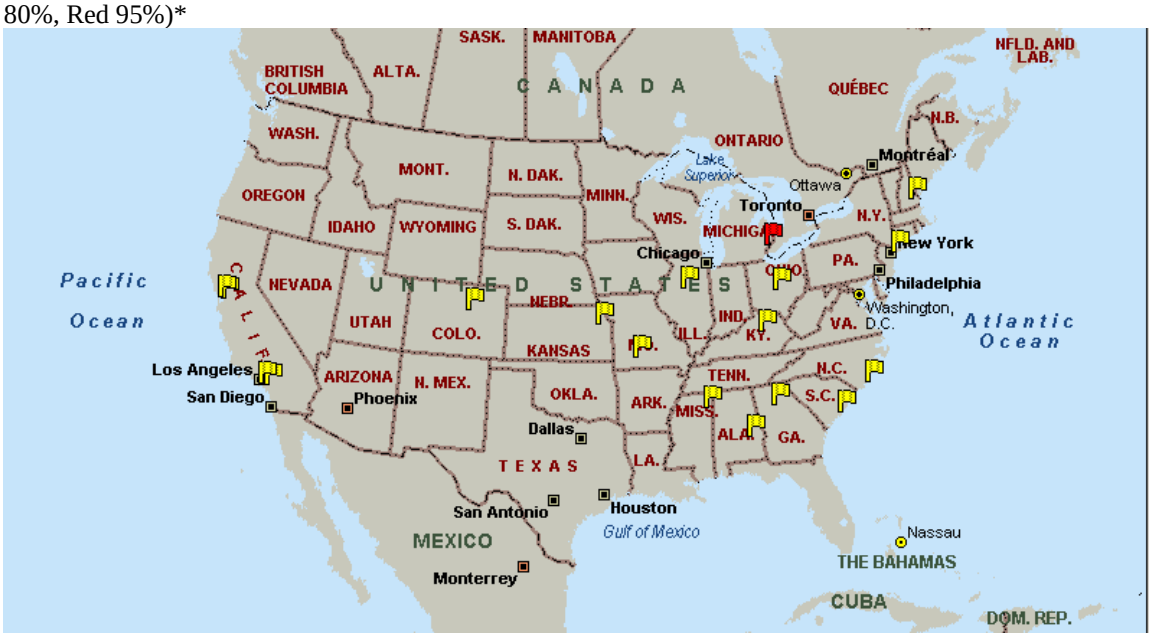
Map 2 – 2004 flagged test centers (Yellow – 80%, Red 95%)



Map 3 – Difference in pass rates from 2004 to 2005 flagged test centers (Yellow – 80%, Red 95%)



Map 4
Difference in pass rates from 2004 to 2005 flagged **(using Joint Standard Error)** test centers (Yellow – 80%, Red 95%)*



* Guam's test center was also flagged at the 80% threshold.

Discussion

The first point worth mentioning is that there is a very common practice, examining pass rates by test center, that is not very well documented in the literature. There are probably several reasons for this. First, is that test publishers are probably a little shy about discussing the specific details of their vulnerabilities. This is understandable from both a business and a legal perspective. It also seems that because most of the methods that quickly come to mind are not very sophisticated, many researchers will select other research topics that promise greater prestige and glamour.

Second, just as it is imprudent to rely solely on probabilistic evidence when trying to prove that an examinee cheated (Dwyer & Hecht, 1994; Buss & Novick, 1980), it is equally imprudent to use probabilistic evidence alone to establish that there is cheating going on at a test center. It seems far more sensible to use this type of information as part of an on-going quality control program and treat flagged results as a trigger for a more detailed investigation. However, instead of using these procedures in an exploratory manner, these procedures could be used in a confirmatory manner as well. For example, if it was alleged that a proctor was permitting people to use proxy testers or to bring in notes in exchange for money, one would want to have eyewitnesses establish that this happened. However, there could be attempts to impeach the credibility of the eyewitness. It would help the case if there were also some independent statistical evidence that could confirm the allegation. Although these two points are pretty well accepted in the testing community, they are asserted here for the sake of thoroughness.

Third, it is helpful to use mapping software to interpret the data. If the problem is not specific to one test center, but rather test centers in a particular region, being able to see their locations would aid tremendously in trying to look for alternative causes for the jump in pass rate. There might be a school or a review course in that region that is especially good at preparing people for your examination. On the other hand, there might be a secret compilation of your items being passed around.

Fourth, it is important to select a useful timeframe to examine. This is largely driven by the number of candidates that test. If there are too few candidates only the largest deviations will be detectable. Although this paper has not done so, it would be important to consider the effects of Type I error. It seems that the

inflation of Type I error only causes a few false hits in exploratory situations and perhaps could be considered acceptable; however in confirmatory situations that reasoning is less persuasive. Also, what level of confidence is the most useful for detecting problems is an area that needs to be addressed.

Finally, procedures to detect cheating are difficult to examine empirically when only a few people attempt to cheat. In nursing, the large investment of time, money, and education make risks of cheating quite high. Also, there seems to be a culture of integrity and compliance with rules. In nurse licensing, I do not believe that cheating is rampant. I believe the vast majority follow the procedures as stated and try very hard to conduct themselves in an ethical and professional manner. This does make the administration of the examination much easier, but it makes it very difficult to find examples of cheaters that one can use to validate cheating detection models.

References

Bird, C. (1927). The detection of cheating on objective examinations. *School and Society, 25* (635), 261-262.

Bird, C. (1929). An improved method of detecting cheating in objective examinations. Journal of Educational Research, 19 (5), 341-348.

Buss, W. G., & Novick, M. R. (1980). The detection of cheating on standardized tests: Statistical and legal analysis. *Journal of Law and Education, 9* (1), 1-64.

Cizek, G. J. (1999). *Cheating on tests: how to do it, detect it, and prevent it*. NJ: Lawrence Erlbaum.

Dwyer, D. J., & Hecht, J. B. (1994). *Cheating Detection: Statistical, Legal, and Policy Implications*. Normal, Illinois: Illinois State University. (ERIC Document Reproduction Service No. ED 382 066).

Frary, R. B. (1992). *Statistical detection of Multiple-Choice Test Answer Copying: State of the Art* (Report No. RR-01-07). Enschede, Netherlands: University of Twente. (ERIC Document Reproduction Service No. ED 351 358).

Levitt, S. D., & Dubner, S. J. (2005). *Freakenomics: A rogue economist explores the hidden side of everything.* NY: Harper Collins.

Sotaridona, L. S., & Meijer, R. R. (2001a). *Statistical Properties of the K-Index for Detecting Answer Copying* (Report No. RR-01-06). Enschede, Netherlands: University of Twente. (ERIC Document Reproduction Service No. ED 467 372).

Sotaridona, L. S., & Meijer, R. R. (2001b). *Two New Statistics to Detect Answer Copying* (Report No. RR-01-07). Enschede, Netherlands: University of Twente. (ERIC Document Reproduction Service No. ED 467 373).

Van der Linden, W. J., & Sotaridona, L. S. (2002). *A Statistical Test for Detecting Answer Copying on Multiple-Choice Tests*. Enschede, Netherlands: University of Twente. (ERIC Document Reproduction Service No. ED 473 531).