Alpine Testing Solutions

Detecting Suspect Examinees: An Application of Differential Person Functioning Analysis

Russell W. Smith
Susan L. Davis-Becker

Alpine Testing Solutions

Paper presented at the annual conference of the National Council on Measurement in Education, New Orleans, LA

April, 2011

We would like to acknowledge and thank the test sponsor, who shall remain anonymous, for allowing us to use their data for the analyses in this paper.

Russell W. Smith
Alpine Testing Solutions
Russell.Smith@alpinetesting.com

Abstract

Typical cheating statistical analyses focus on answer copying or impersonation (Cizek, 1999, p. 136). Within some credentialing fields, a different cheating problem has emerged. Examinees are gaining access to the test content prior to the examination being administered via illicit means. In this paper we propose a new approach for detecting this type of cheating through a two-step process. In the first step, examinees are administered a standard test form along with an additional set of security items that are presumed to be uncompromised. In the second step, examinees' performance on the two sets of items (exam form, security items) are compared through differential person functioning analysis and those examinees with suspect results (high performance on exam form and low performance on security items) are flagged. This process is detailed along with an example analysis from a certification testing program. Further analyses explore the accuracy with various numbers and difficulties of security items.

Detecting Suspect Examinees: An Application of Differential Person Functioning Analysis

Cheating is a serious threat to the validity of a testing program. Impara and Foster (2006) state that "cheating introduces what Messick (1989) characterized as construct-irrelevant variance" (pg. 91-92) as it becomes unclear if successful exam performance is due to mastery of the exam content or cheating. Cheating is often a problem in credentialing examines for several reasons. First and foremost, there are substantial stakes associated with an examinee's performance on the exam. A passing score on such a credentialing examination may result in the examinee becoming eligible to work in a field (e.g., licensure), or qualified for a new or higher paid position (e.g., certification). Second, many credentialing examinations are delivered on-demand and computer-based and are therefore more vulnerable to security problems as compared to exams administered in other formats or less frequently (Cohen and Wollack, 2006). Third, given the motivation of examinees to cheat on these examinations, there is a market for exam content. Those seeking to profit from examinees' desire to pass an exam will use extreme measures to gain access to the content. Such stolen content may then made available for purchase over the Internet (e.g., Smith, 2005).

Given the serious implications of cheating for a testing program, it is important for testing programs to enact monitoring programs by which they can identify suspect behavior. Cizek (1999, pg. 145) points out that "even statisticians and others who appreciate the weightiness of probabilistic statements--perhaps especially those people--recognize the limitations of statistical methods." However, he does go on to argue in favor of the expanded use statistical methods for detecting cheating. He suggests that statistical methods are preventative, that they show "concern about the problem of cheating and… commitment to address it" (p. 147). Typical approaches for such monitoring may include evaluating examinee-level performance to identify suspect examinees or item-level statistics to identify potentially compromised test content.

Cizek (1999) summarizes a set of statistical approaches used to detect answer copying. The focus of these methods is to compare the response patterns of pairs of examinees. These approaches attempt to identify individuals that copy other individuals during a test administration. This is a different problem than identifying groups of examinees who have gained access to the examination content prior to taking the examination. The focus of this study is an attempt to develop an approach to identify examinees who have gained access to the exam content.

Currently, some examination programs conduct various types of analyses to detect security problems at the test- or examinee-level. An example of such analyses would be evaluating total exam performance by total exam time. For example, Figure 1 shows a scatter plot of each examinee's total exam time by their examination score from an international certification program that consists of 40 multiple choice items with a median item response time of 43 seconds per item. The cluster of examinees in the upper left corner of the Figure scored at or near 100% correct in a very short amount of time (e.g. under 9 minutes total). This is an all too typical result in certification testing when examinees have prior knowledge of the exam content, recognize exam questions quickly, and respond almost immediately.

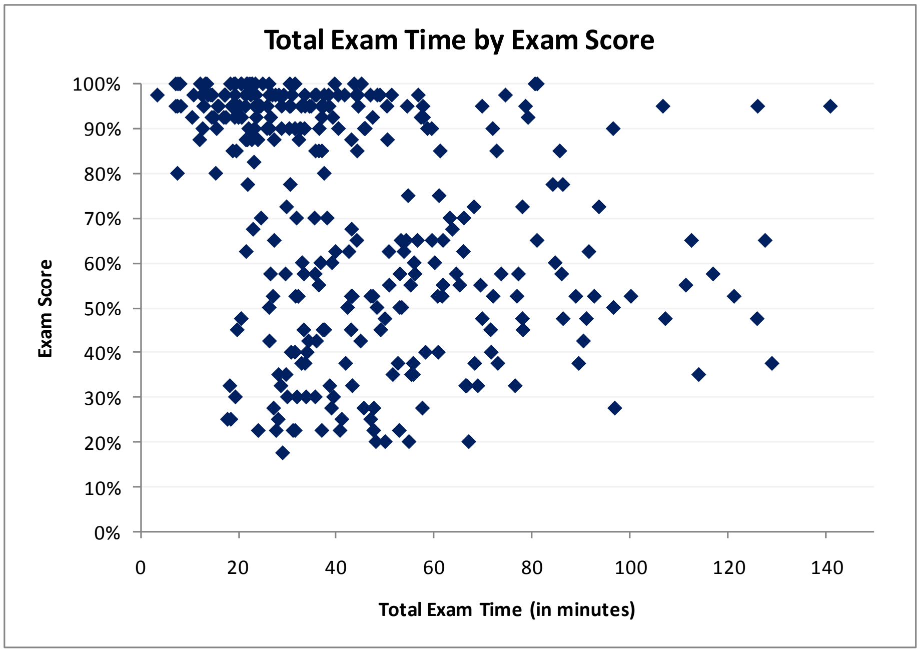

Figure 1. Examinee Response Times by Exam Score

As is evidenced in Figure 1, evaluating total exam time by exam performance can help identify a potential exposure problem at the test level. However, test sponsors cannot rely on such information to identify suspect examinees because it would be easy for a cheating examinee to modify their behavior and slow their response time.

The purpose of this study is to investigate the feasibility of a method for detecting examinees who likely had prior knowledge of item content or answers through differential person functioning (DPF) analysis. This is a two step approach. In the first step, a testing program administers a set of items embedded within the operational examination (security items) to each examinee. These security items would be drawn from a bank of items aligned to the test content and would be believed to be uncompromised. In addition, these items would be continuously refreshed within a test form in a much shorter window (e.g. once a month, or after a given number of exposures) than is possible for the operational test content simply because there would be a small number of security items relative to the number of operational items.

In the second step of this process, DPF analysis would be used to identify examinees who demonstrated suspect behavior based on the comparison of their performance on the scored items to their performance on the security items. For example, if an examinee passes the operational examination and also scores sufficiently high on the security items there would be additional evidence of the examinee’s knowledge and ability within this professional content area. Conversely, an examinee with a high score on the operational examination who performed poorly on the security items would be flagged for suspect behavior. By way of example, if the operational forms of an examination consist of 60 items, a program could reasonably include 10 security items without substantially increasing the burden on the test takers. An examinee who scored 59 out of 60 on the operational part of the test and only 2 out of 10 on the security items might be considered suspect. In other words, a sufficiently high score on the operational items associated with sufficiently low scores on the security items would prompt a concern for the validity of the operational score.

There are several appealing aspects to this approach. First, it would be difficult for cheaters to modify their behavior to get around this approach to flagging. Second, examinees would be aware of the changing content and therefore might be deterred from trying to cheat knowing they would not obtain the full test content a priori. Third, and most importantly, it could help identify examinees that do not have the knowledge and ability to legitimately earn the credential.

The purpose of this study was to (1) demonstrate the process and outcomes of including these constantly refreshed security items and (2) investigate the number of security items that would be necessary to detect egregious offenders with sufficient confidence to make decisions or take appropriate actions.

Data

Figure 1 showed an example of data from an international certification examination program with a serious security problem as indicated by the large number of examinees who scored very well on the exam in an extremely short amount of time. The administrators of this program recognized the security problem and attempted to address it using the solution described above. In this initial attempt, 642 examinees were randomly administered one of two pre-equated forms; each consisting of 40 dichotomously scored items. Additionally, examinees were randomly administered 25 presumably unexposed dichotomously scored security items from a bank of 61 items developed for this examination.

Methods

DPF is statistical analysis approach, in this case using a Rasch measurement model, which holds the item and person parameters constant except for the person for whom DPF is being calculated. The examinee's ability measure is estimated on each subset of items. A log-odds estimate of the difference between the two ability measures is calculated. Given the joint standard error between the measures, a probability is calculated for each examinee that indicates the likelihood of a particular combination of scores. For this analysis, the subsets of items are (1) the 40 operational scored items for each examinee and (2) the security items.

The success of this method in identifying suspect examinees is based on the assumption that even if the scored content has been exposed, the smaller set of continuously refreshed items has not been exposed. Therefore, an examinee that is NOT minimally qualified in this professional field but has gained access to the operational test content will likely have a high estimated ability based on the operational scored items and a low estimated ability based on the security items. The results of DPF analysis would be a low estimated probability of these two measures resulting from the same examinee.

The purpose of the first set of analyses was to demonstrate the outcome of using these "security items" to identify suspect examinees. Items were calibrated using Winsteps filtering out examinees with DPF contrasts greater than 1. The item parameters were anchored based on this analysis for all subsequent analyses. A DPF analysis using all 25 security items for each examinee was conducted. Examinees with a DPF contrast greater than 3 (more than a 3 logit difference between the ability measure based on the 40 operational items and the ability measure based on the 25 security items) and with a probability less than .0001 were identified as suspect. Of 624 examinees, 124 were flagged as meeting these criteria. In addition, there were 22 examinees that had DPF contrasts greater than 3 but probabilities exactly equal to .0001 who were not flagged.

The purpose of the second set of analyses was to determine the number of security items needed to have confidence in the results of this approach to flagging and to see how the difficulty of the items might impact that confidence. We acknowledge that administering 25 security items in a 60-item examination is not feasible for many programs. Therefore, we explored the stability of this system of flagging using smaller samples of items. Specifically, four different sample sizes of security items were randomly sampled: 5, 8, 10, and 15 items. Because missing easier items would result in more extreme person measures and therefore smaller probabilities, we also applied these same sample sizes again by selecting the easiest 5, 8 10, and 15 items administered to each examinee. Six different flagging probabilities (.05, .01, .005, .001, .0005, and .0001) were used for each sampling condition. The percent of consistent decisions as well as Type I and Type II error rates were calculated for each of these 48 conditions (6 flagging probabilities by 4 sample sizes by 2 sampling methods).

Results

Demonstration of process

In the first step of the analysis, the full data set including 40 operational items and 25 security items was used to identify suspect examinees. Figure 2 shows the contrasts in the DPF measures for each of the examinees, highlighting the most egregious examinees. Examinees in the lower right had high ability measures on the operational items and low ability measures on the security items. Examinees with a DPF probability less than .0001[1] and a DPF contrast greater than 3 are identified as "Flagged". These examinees may be considered suspect of having prior knowledge of the operational examination content. The results of this analysis served as the baseline comparison for all analyses in the next section. Specifically, the accuracy of each condition was assessed by comparing the results of the flagging to this initial analysis.

[1] This value was selected as a baseline probability as it represents a likelihood of 1 in 10,000 and is the default for some commercially available software packages that run analyses used to detect cheating.

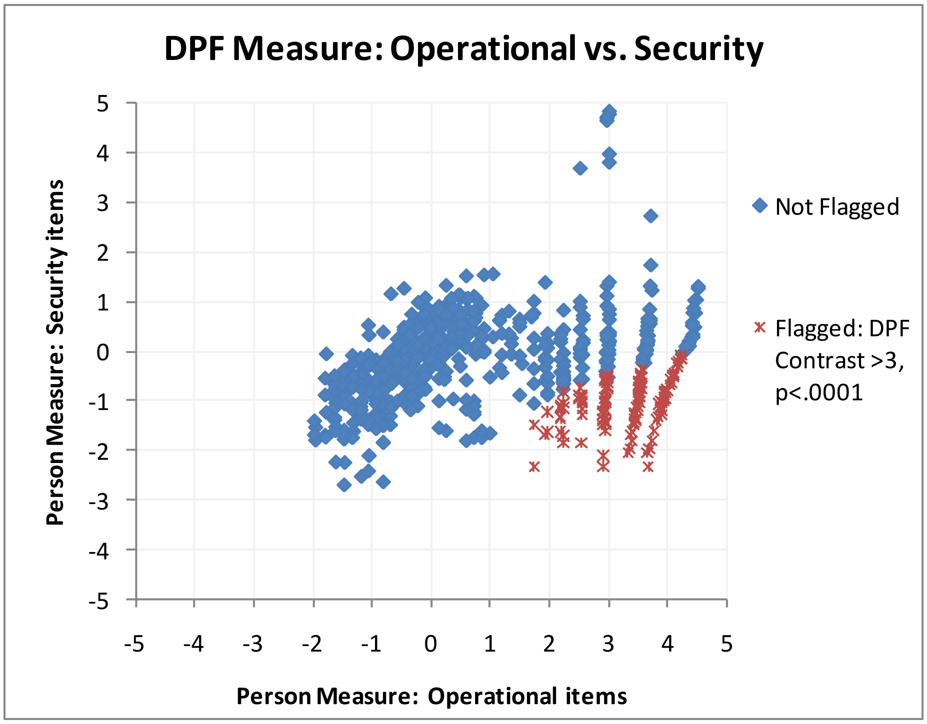

Figure 2. Examinee DPF Measures: Operational vs. Security Items

Investigation of stability across sample sizes

All 48 conditions of flagging criteria were run using the same sample of data and a flagging criterion of DPF contrast greater than 3. Table 1 shows the consistency and error rates based on randomly selecting each of the security item sampling sizes across DPF probabilities (also shown graphically in Figures 3 and 5). Table 2 shows the consistency and error rates based on selecting the easiest items administered to each examinee for each of the sample sizes and across DPF probabilities (also shown graphically in Figures 4 and 6). Consistency is the proportion of agreement of flagged examinees to those identified in the initial DPF analysis with 25 security items and a probability less than .0001. Type I error is the proportion of examinees flagged by the subset analyses that were not flagged by the initial analysis (i.e., over flagging). These examinees should not have been flagged but were marked as suspect in the subset analysis. Type II error is proportion of examinees flagged based on the initial analysis that were not flagged in the subset analysis (i.e., under flagging). These examinees should have been flagged as suspect but were not.

Table 1. Consistency and error rates based on random sampling

	Security	Flagging probability					
	Items	0.05	0.01	0.005	0.001	0.0005	0.0001
Consistency							
	15	0.883	0.888	0.902	0.930	0.935	0.900
	10	0.877	0.903	0.916	0.917	0.903	0.841
	8	0.852	0.883	0.905	0.883	0.866	0.807
	5	0.885	0.810	0.807	0.807	0.807	0.807
Type I errors							
	15	0.103	0.098	0.084	0.051	0.026	0.003
	10	0.114	0.086	0.069	0.014	0.003	0.000
	8	0.123	0.089	0.058	0.016	0.003	0.000
	5	0.044	0.000	0.000	0.000	0.000	0.000
Type II errors							
	15	0.014	0.014	0.014	0.019	0.039	0.097
	10	0.009	0.011	0.016	0.069	0.093	0.159
	8	0.025	0.028	0.037	0.101	0.131	0.193
	5	0.072	0.190	0.193	0.193	0.193	0.193

Table 2. Consistency and error rates based on sampling the easiest items

	Security	Flagging probability					
	Items	0.05	0.01	0.005	0.001	0.0005	0.0001
Consistency							
	15	0.931	0.944	0.945	0.945	0.939	0.907
	10	0.925	0.936	0.933	0.900	0.885	0.824
	8	0.935	0.928	0.910	0.843	0.829	0.807
	5	0.903	0.813	0.807	0.807	0.807	0.807
Type I errors							
	15	0.062	0.050	0.045	0.033	0.026	0.003
	10	0.048	0.030	0.026	0.009	0.005	0.000
	8	0.036	0.022	0.011	0.002	0.000	0.000
	5	0.011	0.000	0.000	0.000	0.000	0.000
Type II errors							
	15	0.006	0.006	0.009	0.022	0.034	0.090
	10	0.026	0.034	0.040	0.090	0.111	0.176
	8	0.030	0.050	0.079	0.156	0.171	0.193
	5	0.086	0.187	0.193	0.193	0.193	0.193

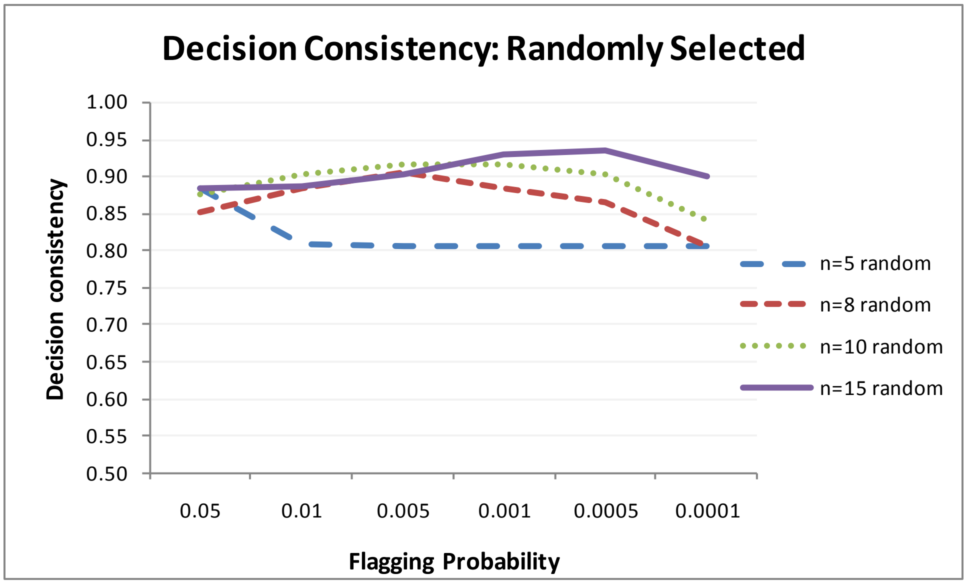

Figure 3. Decision consistency, random selection

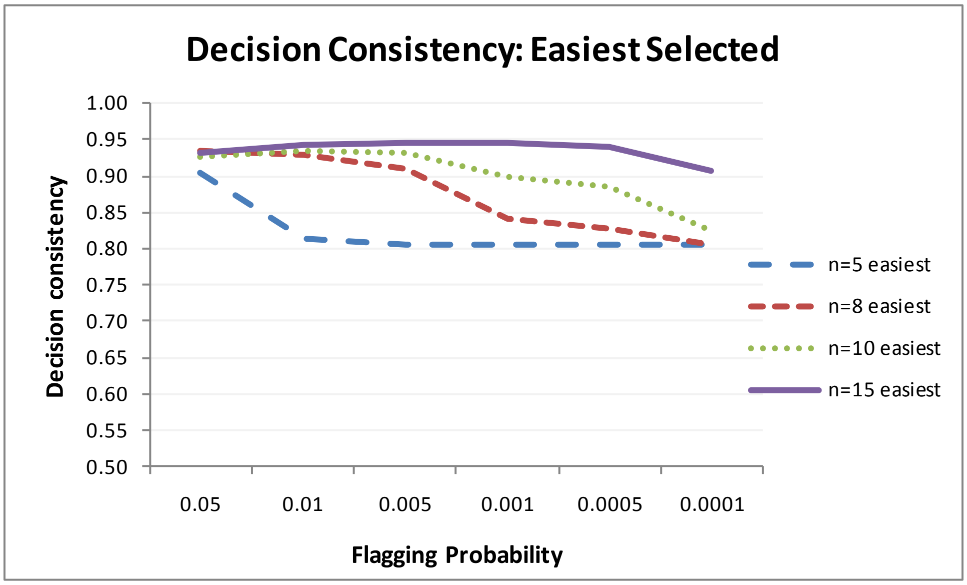

Figure 4. Decision consistency, easiest selected

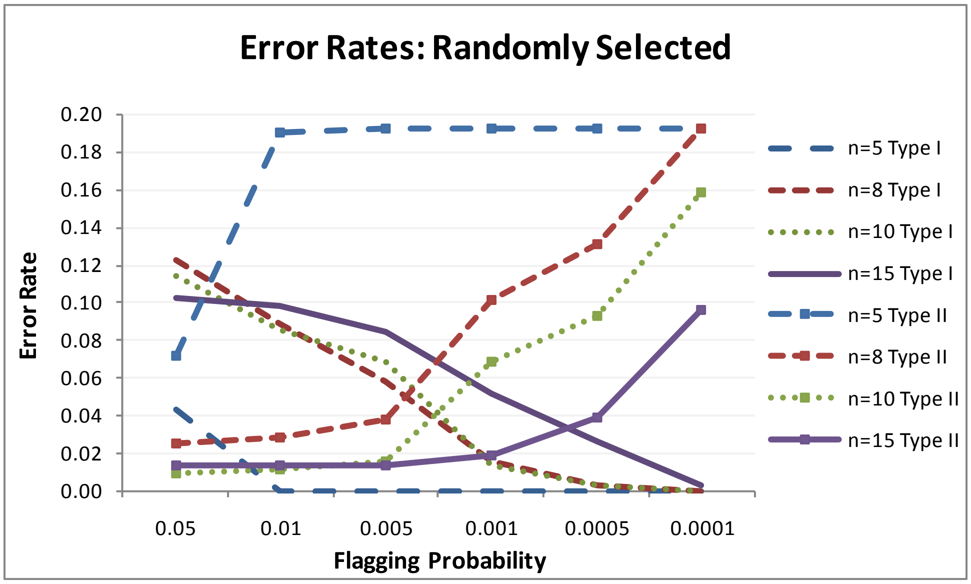

Figure 5. Error rates, random selection

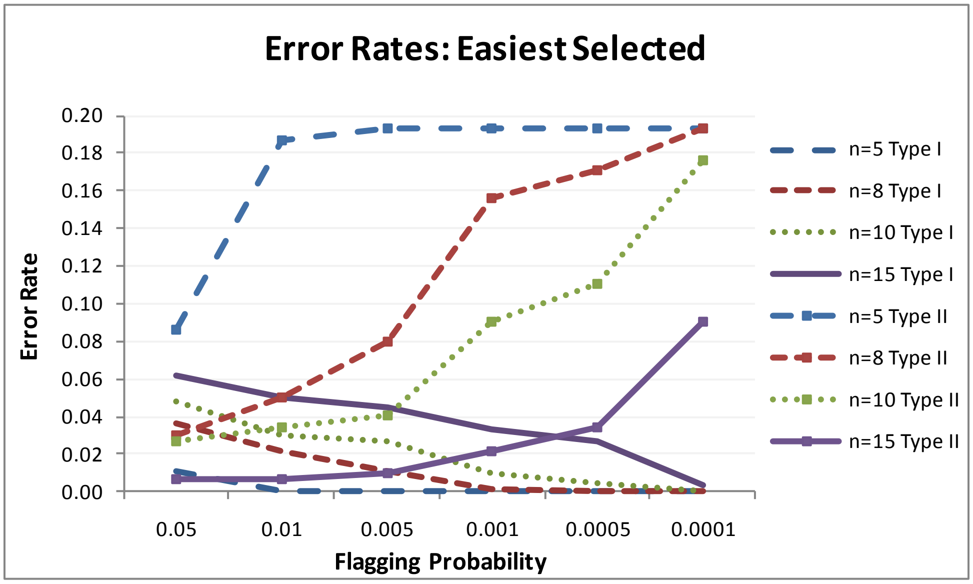

Figure 6. Error rates, easiest selected

In terms of decision consistency, higher values were observed when more security items were used, particularly at lower probability values. Using larger probability values, higher decision consistency estimates resulted from using the easiest items seen by each examinee in contrast to randomly selecting the security items. In the random selection condition, it is possible that some examinees were not administered many easy items.

With respect to Type I errors (over flagging), lower error rates were observed when smaller samples of items were used along with smaller probably values and the easiest items each examinee saw were selected. However, these results come with a caution that using a smaller sample of items and lower probability values does reduce Type I error but will also cause a decrease in power, or the ability to flag candidates who should be flagged.

The maximum possible Type II error rate (under flagging) in this study is just under 20% because that is the proportion of examinees flagged by the initial DPF analysis. When looking at Type II error, lower error rates were observed when the items were selected randomly, a larger sample of security items was used, and a larger flagging probability was used.

Conclusions

There are some distinct advantages for using security items along with DPF analysis for flagging suspect examinees. First, though based on statistical probability, the approach is conceptually simple and logical. One can easily explain to test sponsors and policy makers that getting a very high score on one set of items and getting a very low score on another set of easy items designed to measure the same domain is unlikely. For example, one examinee from this analysis scored 40 out of 40 on the operational test and answered 0 of 8 security items correctly. A practitioner could easily communicate this scenario to policy makers and explain that the probability of these two scores coming from the same examinee (assuming that the person did not have prior knowledge of the operational content) would be 1 in 5000 (i.e., .0002). Second, without prior knowledge of the continuously refreshed items, cheating examinees cannot alter their behavior in such a way as to not be detected. Third, DPF contrast and associated probability values are provided for each examinee which can be used to corroborate other evidence of cheating. Finally, this approach may be presented as a validation of examinees' abilities instead of being accusatory, which may be appealing to practitioners as well as test sponsors.

Based on the results of this study, there is no recommended right number of security items. The results are sample and test dependent for an exam that has very clearly been compromised and likely will not generalize. However, the results do allow us to better understand the trade offs of power and Type I and II errors. This is an important consideration for a test sponsor if examinee performance on a set of items is going to be used for decision making at the examinee level. For example, if a test program wants to avoid Type I errors (over flagging) and is willing to lose power (not being able to detect some of the most egregious offenders) then they might be able to use eight security items. For a particular testing program, simulation studies could be conducted that would estimate consistency and error rates for different numbers of items with different parameters.

Based on the data in this study, using 8 security items, a DPF contrast greater than 3, and flagging probabilities less than .005 would result in 91% decision consistency, 1.1% Type I error rate, and a 7.9% Type II error rate. The flagged examinees scored 37 or higher out of the 40 operational items and 3 or fewer out of the 8 security items. If the operational and security items were delivered in a fixed form, rather than selected randomly, and had known item parameters, it would not be necessary to run a post-administration DPF analyses to flag examinees. Rather, DPF contrast and probabilities could be used to determine raw scores that would indicate cheating that could then be used to make decisions about the examinee at the time of administration based on those raw scores.

This is the first step in testing this methodology as a means of identifying suspect examinees. With sufficient support and measurement precision, policy makers may choose to take specific action based on these results. Such action could be at the exam level (e.g. rebuild the exam forms with new unexposed content) or the examinee level (e.g. withhold certification).

Real data rather than simulated data were used in part to show the reality of the challenge we are facing and in part because we would not know how to begin to simulate cheating as dimension, which it clearly is when exposure is this extreme. We recognize that the severity of this high of frequency of examinees having prior knowledge of content may be unique to testing within some professional fields and may not generalize to other testing programs. However, it is reasonable to think that similar problems may exist in other programs to a lesser extent or that as test takers around the world gain more access and skills with the Internet, this type of cheating may emerge.

Many certification exams are available on demand worldwide and many retake policies allow examinees to take an exam immediately or almost immediately after an administration. Perhaps this convenience, along with increasing access to the Internet, has come at the cost of examination content being exposed. There are other approaches, such as windowed testing, that might have a much larger impact on deterring cheating and possibly even at a lesser cost. We encourage test sponsors to evaluate the benefits and the costs of such approaches. In the mean time, continuously refreshing a set of security items and using that information to flag examinees, may potentially thwart and deter some cheaters.

References

Cizek, G.J. (1999). *Cheating on tests: How to do it, detect it and prevent it*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Cohen, A.S. & Wollack, J.A. (2006). Test Administration, Security, Scoring and Reporting. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 355-386). New York: Macmillan.

Impara, J. C., & Foster, D. (2006). Item development strategies to minimize test fraud. In S.M. Downing & T.M. Haladyna (Eds.). *Handbook of test development* (pp. 91-114). Mahwah, NJ: Lawrence Erlbaum Associates.

Linacre, J.M. (2009). *WINSTEPS® Rasch measurement computer program*. Beaverton, Oregon: Winsteps.com.

Messick , S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Smith, R W. (2005). The Impact of Internet Sites on Item Exposure and Item Parameter Drift. *Clear Exam Review*, 16(2), 12-15.