Extending Differential Person and Item Functioning to Aid in Maintenance of Exposed Exams

Lisa S. O'Leary
Russell W. Smith

Alpine Testing Solutions

Paper presented at the annual conference of the National Council on Measurement in Education,
San Francisco, CA
April, 2013

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.

Lisa S. O'Leary, Ph.D.
Alpine Testing Solutions
lisa.oleary@alpinetesting.com

Abstract

Rampant test fraud in information technology certification testing programs has lead to the widespread unauthorized release of exam forms and perpetual item exposure. This paper presents evidence of how differential person functioning can be used in conjunction with differential item functioning to minimize the tangible and intangible costs and maximize the measurement integrity and validity associated with exam results that could be affected by test fraud. The methodology presented in this paper can identify suspect candidates based on their own aberrant response patterns, as well as control the influence of item degradation by assessing the extent of exposure for particular items. These results will help to: detect when security breaches have occurred, build more defensible cases to enforce sanctions against candidates, inform exam maintenance and item development, and provide insight into scale and item stability over time. The application of these dual analysis efforts will help to preserve the validity of candidate decisions and the reputation of testing programs operating in an environment of grossly exposed exam content. This paper also contributes to the growing literature on data forensic techniques to gauge the impact of test fraud on testing programs via statistical and psychometric methods.

Introduction

Item exposure is a consistent threat to the validity of certification examinations due to prevalent piracy practices that include the regular theft and unauthorized release of individual items, entire item pools, and exam forms. It has been documented that item exposure is widespread within information technology (IT) certification exams in particular; live exam content and items are routinely being exposed on the Internet. The three basic outlets for legitimate and illegitimate advice and content are: exam preparation sites, Internet auction sites, and braindump sites that either formally sell stolen content or informally encourage candidates to share their own recollections about particular certification exams (Smith, 2004; Foster, 2013). According to research (Maynes, 2009), within some high-volume IT certification testing programs, a majority of candidates (i.e., 85% or more) may have acquired prior item knowledge by purchasing content through braindump sites.

As a result of this problem, IT certification testing programs have been plagued by skepticism about the legitimacy of candidate exam results and resulting inferences about individuals' knowledge, skills, and abilities. Therefore, this paper will focus specifically on addressing and diminishing the influence of test fraud, which is defined as "any behavior that inappropriately or illegally captures test questions and/or answers" (Foster, 2013, p. 47), as opposed to other cheating behaviors such as collusion or proxy test-taking. Test fraud is prominent in IT certification programs for a multitude of reasons, including, but not limited to: the high professional stakes linked to successful certification; advances in technology; the computer-based, continual delivery of many of these exams; and the candidates' familiarity and years of experience with technology (Wollack & Fremer, 2013; Smith, 2004). Given the prevalence of test fraud—particularly piracy—in IT certification testing programs, managing item exposure is a top priority regarding exam security.

Test fraud can damage testing programs on several levels. Intangible costs are a loss of credibility and face validity with key stakeholders (including candidates themselves and employers). Tangible costs are associated with continual efforts geared at item protection and cheating detection, as well as item and test development. Item exposure as a security concern is rampant within certification examinations; it constitutes serious threats to the validity of score interpretation and use. As Impara and Foster (2006) highlight, cheating introduces construct-irrelevant variance; scores may not accurately represent underlying content knowledge but instead "how a particular set of test questions has been answered or tasks performed through inappropriate means" (pp. 91-92). Once items or entire examinations are exposed, candidates' score integrity is compromised and the validity of the inferences being made based on the exam scores can be questioned (Wollack & Fremer, 2013).Consequently, it becomes difficult to discern if candidate performance is due to true ability or cheating through prior unauthorized access to exam content.

Additionally, item degradation occurs as a result of exposure. Such exposure jeopardizes the quality, utility, and functioning of individual items as well as entire item banks. Item degradation is caused by the deterioration of desirable item characteristics over time, including, but not limited to: a reduction in content relevance and representativeness, loss of quality of technical characteristics (i.e., item difficulty and reliability), and a decrease in utility of the correlation between the item and construct of interest (Yang, Ferdous, & Chin, 2007). Maynes (2013) describes the residual effect of a large number of candidates having access to stolen content as a “three-fold issue” (p. 180) that shifts classical item statistics, results in unexpectedly high levels of performance, and presents evidence of collusion through response pattern similarities. This item degradation threatens the statistical assumptions and psychometric models underlying these certification exams, as well as the individual testing outcomes for candidates. The ability to detect and mitigate the negative effects of prior item exposure in an IT certification context where prevention of test theft is not realistic is therefore critically important in maintaining evidence of the validity of these testing programs.

Item exposure negatively impacts item statistics. Han and Hambleton (2008) noted that proactive, systematic efforts should be made to identify exposed items, retire, and replace those items with less exposed (or ideally new items) before the impact on the item statistics is too drastic. While Han and Hambleton (2008) indicate that concerted security efforts should be put in place to initially protect from the test fraud and piracy, they recognize that continuous data forensics efforts are also necessary to assess the extent of compromised content. Therefore, it is important for testing programs with perpetually at-risk exam content to develop procedures around data forensics to quantify the degree of unauthorized released content, as well as policies around item exposure controls to gauge the extent of compromise in item banks and retire and refresh content as necessary (Impara & Foster, 2006).

For the purposes of this paper, item exposure is defined as “the number or percent of people with pre-knowledge of the item before taking the test” as opposed to the more traditional “number of naturally occurring presentations of the questions” (Foster, 2013, p. 79). A solid research foundation has developed in recent years detailing how to detect and statistically control the standard item exposure that can be expected from the use of computer-adaptive testing algorithms and logic (Han & Hambleton, 2004; Lu & Hambleton, 2003; Han, 2003; Veerkamp & Glas, 2000). The authors of this research recommend that extensive item banks should be created to enable the regular replacement of items once over-exposure and compromise has been detected (Han & Hambleton, 2008; Foster, 2013). While the authors of this paper advocate for developing large item banks to address issues of item exposure, the focus will be more on identifying compromised items and replacing those with new items as necessary to allow for the continual administration of some unexposed content amongst largely exposed item pools instead of controlling the routine usage of particular items. Timeliness is of the essence to minimize the impact of item exposure and maximize measurement integrity in this testing environment; the extended use of exposed test items increases the extent of sharing and opportunity for prior knowledge of exam content (Carson, 2013). Test fraud can occur immediately following the initial release of exam forms, sometimes within days or even hours (Maynes, 2009; Smith, 2004). Therefore, it has become of utmost importance that new items are continuously produced to support content refreshing and replacement.

Testing programs have had to develop methods through which they can identify exposed items and aberrant response patterns within candidates to reduce the influence of test fraud on the validity of their testing programs. Several common practices are used within the IT testing community to detect irregular testing behavior, such as investigation into item response latencies (Maynes, 2013), comparisons of total exam performance by total exam time (Smith & Davis-Becker, 2011) and score patterns on Trojan Horse items (Maynes, 2009). These practices help detect suspicious candidate behavior (i.e., candidates who answer items quickly but receive high scores and/or consistently answer correctly according to intentionally miskeyed items). However, these techniques are limited in their enforceability. Candidates can adjust their response patterns to avoid detection. Additionally, these methods offer little assistance to identify specific compromised content despite giving some indication of the overall extent of exposure.

Many testing programs have thus incorporated the application of statistical and psychometric methods on their exam data to detect cheating on examinations into their security analyses. Cizek (1999) supports the use of data forensics to detect cheating despite some limitations to the use of methods because “the conclusion that cheating has occurred is almost always probabilistic and requires inference” (p. 150). It is therefore common practice to rely on multiple sources of evidence and further investigation into statistical anomalies prior to taking action against any particular candidate or group of candidates. While there are still on-going discussions regarding the legitimacy of enforcement of sanctions against candidates based on the results of data forensics, Maynes (2013) has suggested that these actions can be considered defensible provided that the methods employed implement proper error control and rely on accurate data, credible measurements, consistent procedures, scientific methods, probability statements, and well-reasoned findings. However, current data forensics range in their levels of effectiveness and sophistication; techniques have been proposed to detect unusual score gains, collusion among candidates, aberrant wrong and right answer patterns, suspicious erasures and answer changes, and test retake violations (Fremer & Ferrera, 2013).

While research in the field of data forensics has emerged and developed in recent years, more advanced methodologies to assess the impact of test fraud and item exposure “are in their infancy and little is known about how well the few methods work in practice” (Wollack & Fremer, 2013, p. 8). In Cizek’s (1999) review of statistical methods to identify students copying from one another, he noted that initial attempts to apply the Rasch person-fit measures to identify suspect candidates based on misfitting response patterns yielded little valuable information. However, more recent work in the application of IRT models—particularly Rasch—to generate statistics indicative of cheating behavior other than collusion have proven more successful. For example, Maynes (2011) provided useful background on how IRT methods could be applied to compute score differences within a candidate’s test responses through precise probability statements within single exam instances.

Item exposure calls into question the validity of candidates’ test scores; therefore, researchers argue that analyses must compare the performance of candidates on both exposed and unexposed test content (Carson, 2103; ATP, 2013; Maynes, 2009). For example, Maynes’ research (2011) purported the use of his score differential method to enable comparisons between new and old items, scored versus unscored items, and multiple choice and performance-based items to detect unusual score patterns within candidates that could indicate prior access to exam content. This paper presents further methodology to enhance these data forensics techniques in terms of identifying aberrant performance by candidates in testing programs within largely exposed item pools. Namely, this paper extends the existing research by Smith and Davis-Becker (2011) on how to detect candidates likely to have item pre-knowledge through the use of differential person functioning (DPF). It supports the utilization of DPF to identify candidates who likely gained prior access to exam content through illicit means by comparing performance on unscored pilot items along with scored items. It then furthers the approach by following the DPF with subsequent differential item functioning (DIF) analyses to detect compromised items due to exposure. While this research does not address the current dearth of research “that has demonstrated the efficacy of DIF analysis for detecting group-based security breaches" (Maynes, 2013, p. 192), it does present evidence of how DPF can be used in conjunction with DIF to highlight individual candidate incidences of item exposure and reduce the impact of test fraud on item banks.

Data

The data in this study were 8,350 administrations of a large-scale IT certification exam with substantial item-exposure issues. Candidates were randomly administered one of three pre-equated, parallel forms, each consisting of 80 scored items. Additionally, candidates were randomly administered 20 unscored newly written pilot items that were proportionately representative of the content detailed in the exam blueprint. The item pool consisted of 641 total items, 227 scored and 414 unscored. Figure 1 shows candidates' total scored item scores and their total test time; the high incidence of candidates with high exam scores in low time resulted in exposure concerns for the 227 scored items. Specifically, 95 candidates achieved a perfect score on the scored items (80 out of 80) in a median time of 23.2 minutes, which equates to 17.4 seconds per item. Furthermore, 141 candidates were identified as having spent less than 5 seconds on 20% or more of the items.

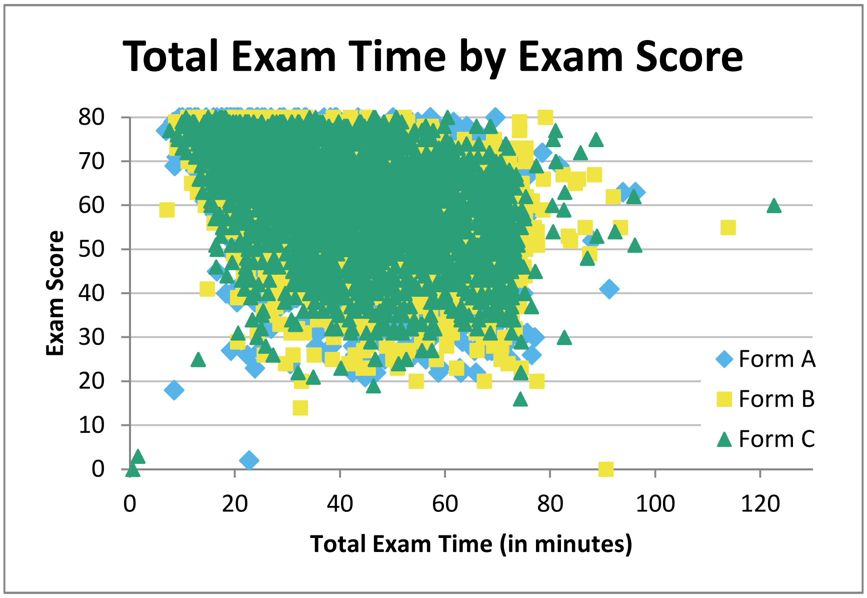

Figure 1. Total Exam Time by Total Score

Methods

Differential Person Functioning (DPF)

Differential person functioning (DPF) was initially conducted using Winsteps (Linacre, 2009) to identify candidates likely to have had prior knowledge of exam content by comparing candidates' performance on scored and unscored items. DPF is a statistical analysis approach for comparing the performance of candidates on subsets of items while holding the item and person parameters constant, except for the person for whom DPF is being calculated. A candidate's ability measures are estimated on each subset of items, along with a calculation of the log-odds estimate of the difference between the two ability measures. A probability is then calculated for each candidate that indicates the likelihood of a particular combination of scores based on the joint standard error between the measures.

For this analysis, the item subsets for the DPF are the 80 scored, operational items and 20 unscored, pilot items administered to each candidate. DPF is beneficial because the unscored items are randomly administered to candidates; the estimated probability and error is individually based on the specific items each respective candidate answered. This process enables flagging individual candidates with aberrant scores on the operational versus pilot items. Given the underlying assumptions of the Rasch model (i.e., sample independence), the precision of the ability estimates is not impacted by the comparative sample sizes of scored versus unscored items. However, practical decisions around the ratio of scored to unscored items administered to candidates and corresponding flagging criteria should be determined by the purpose of the exam, subsequent security analyses, and the testing program's capacity for follow-up investigation and enforcement.

This methodology is based on the notion that candidates with prior knowledge of the item pool would likely have a high estimated ability on the scored items and a low estimated ability on the unscored items; this results in a low estimated probability of these two measures resulting for the same candidate. This presupposes that only the operational, scored items have been exposed and that the unscored, pilot items have not yet been subject to test fraud. If this condition is met, this DPF analysis provides evidence for a validity argument for or against candidates' exam scores by identifying candidates likely to have had prior content knowledge—intended or not. This assumption was confirmed for the purposes of this analysis, with the average Rasch item difficulty being easier for the scored items than the unscored items (average of -0.43 and 0.23 for scored and unscored items, respectively).

The DPF analysis was conducted for all candidates on all 80 scored and 20 unscored items to detect candidates with an unexpectedly low probability of the combination of their two ability estimates. Candidates were flagged as possibly suspect if they had more than a 2 logit difference between their respective ability measures on the 80 scored items and 20 unscored items (DPF contrast greater than 2) and a probability of less than .01. These flagging criteria, which were more liberal than those suggested by Smith and Davis-Becker (2011), were considered appropriate for these analyses because the focus was identifying a cohort of candidates that likely had pre-knowledge of items for subsequent use as a subgroup in the DIF analyses—not necessarily providing probabilistic-based data to support enforcement cases against particular candidates. If detection and additional evidence for enforcing sanctions are primarily goals of security analyses, more conservative criteria would be suitable (e.g., the DPF contrast greater than 3 and probability less than .0001 utilized by Smith and Davis-Becker [2011]).[1]

Differential Item Functioning (DIF)

Differential item functioning (DIF) was subsequently conducted using Winsteps (Linacre, 2009) to assess the extent to which candidate prior knowledge of exam content impacted item

[1] Also see this research for practical guidance on decision consistency (as well as Type I and Type II error rates) which can be expected for a variety of unscored item sample sizes that could be useful to selecting flagging criteria.

performance by comparing item difficulty based on candidates' DPF results. The DIF procedure implemented in Winsteps is based on the same theoretical properties as the Mantel-Haenszel method (Linacre & Wright, 1987). Of particular interest was the extent of item degradation that resulted from the unauthorized item exposure due to test fraud. Additionally, gathering information to drive exam maintenance—including identifying items in need of content refreshing or replacement as well as those appropriate to be utilized as anchor items—was a top priority.

DIF is a statistical approach for comparing item difficulty across subgroups while controlling for candidate ability and item difficulty, except for the item for which DIF is being calculated. Item difficulties are calculated for each subgroup, along with a calculation of the log-odds estimate of the difference between the two difficulty measures. A probability is then calculated for each item that indicates the likelihood of a particular combination of difficulties based on the joint standard error between the difficulty measures.

For this DIF analysis, the subgroups of candidates were those flagged through the DPF versus those without flags, or candidates with likely prior item exposure versus those with response patterns not indicative of having item pre-knowledge. The intent of the DIF is to determine the extent to which items have been exposed by comparing the item difficulty measures for flagged versus non-flagged candidates. The assumption of this paper is that non-exposed items would be of similar difficulty for both groups of candidates and that exposed items would favor flagged candidates. Again, given the underlying assumptions of sample invariance in the Rasch-based DIF model, the use of flagged candidates based on aberrant scores on scored versus unscored items should not affect model fit. The estimation of the parameters must be invariant across sub-samples of candidates. In terms of establishing flagging criteria to identify compromised items, the overall size of the item pool, tolerance for retiring and replacing items, ability to refresh content with new, unscored items, and budget and resources for continued item development should be considered.

This DIF analysis was conducted for all 641 total items (227 scored and 414 unscored) to support exam maintenance in the context of a grossly exposed item bank, including the determination of items that are fair to administer to all candidates, have been affected by exposure, are viable for inclusion in future iterations of the exam, and are appropriate for assignment to an anchor set. Items were flagged as exposed if they had more than a 2 logit difference between their respective difficulty measures for the flagged versus non-flagged candidates (DIF contrast greater than 2) and a probability of less than .05. These flagging criteria were considered appropriate for these analyses; the purpose of the DIF was to detect exposed items for retirement as well as well-functioning items to serve as anchor items in preparation for an exam upgrade involving widespread content refreshing and item replacement. In circumstances with fewer resources to support wide-scale item development and form re-assembly, more conservative DIF contrast and probability flagging criteria would likely be better to reduce the pool of compromised items identified through the analyses.

Results

Differential Person Functioning (DPF)

Of 8,350 candidates, 531 candidates (6.4%) were flagged for DPF based on the established criteria. The critical probability to identify candidates with DPF was set to *p*<0.01, with a DPF contrast (absolute difference between the estimated Rasch ability measures) > 2.0. In this analysis, a positive DPF contrast suggests that the candidate scored significantly better on the scored items than the unscored items. The flagged candidates scored unexpectedly well on the scored items as compared to the unscored items (e.g., a score of 79/80 on the scored items and 2/20 on the unscored items). For example, the likelihood of Candidate 508d having scored 100% (80 out of 80 possible points) on the scored items and only 60% (12 out of a possible 20 points) on the unscored items is 1 in 10,000. These candidates were therefore considered likely to have had item pre-knowledge and were suspected of accessing exam content prior to exam administration via test fraud. Figure 2 displays the contrasts in the DPF ability measures for each of the candidates. The highlighting displays those candidates with the most differential results by scored versus unscored Rasch measures.

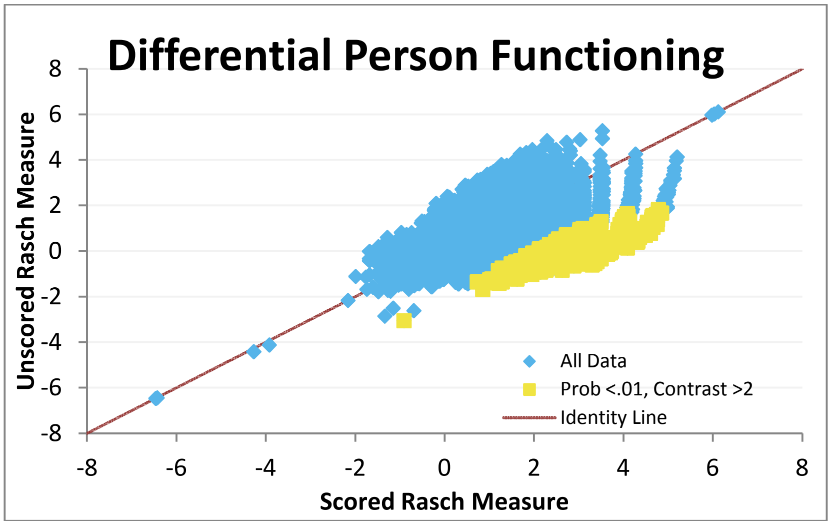

Figure 2. Differential Person Functioning by Scored v. Unscored Items

Differential Item Functioning (DIF)

Of 641 items, 138 items (20.2%) displayed significant DIF based on the set parameters. The critical probability to identify items with DIF was set to *p*<0.05 (higher than the *p*<0.01 DPF critical value) due to the lower stakes associated with identifying exposed items as opposed to suspect candidates that might be subjected to sanctions based on the results of the security analyses. The DIF contrast (i.e., absolute difference between the estimated Rasch item difficulties) was ≥ 2.0. In these analyses, a positive DIF contrast suggests that the flagged candidates performed better on the item; a negative DIF contrast indicates that the non-flagged candidates scored higher on the item. See Table 1 for the breakdown of DIF results (positive or negative) by item status.

Table 1. DIF Results by Item Type

	Significant DIF		No DIF
Item Status	Positive	Negative	
Scored	57	1	169
Unscored	0	80	334

Figure 3 displays the DIF results by candidate flagging status. As seen in Figure 3, the DIF results showed 57 scored items that were statistically significantly easier for flagged candidates, such that flagged candidates performed better on these items than non-flagged candidates. These items were therefore considered to be the most grossly exposed and were marked for retirement and replacement. In contrast, 80 unscored items (and 1 scored item) displayed DIF in the opposite direction; these items were statistically significantly harder for flagged candidates. This was likely due to the secure nature of these items because they had not yet been subject to piracy. These unscored items were considered viable for inclusion as scored items on future iterations of the exam.

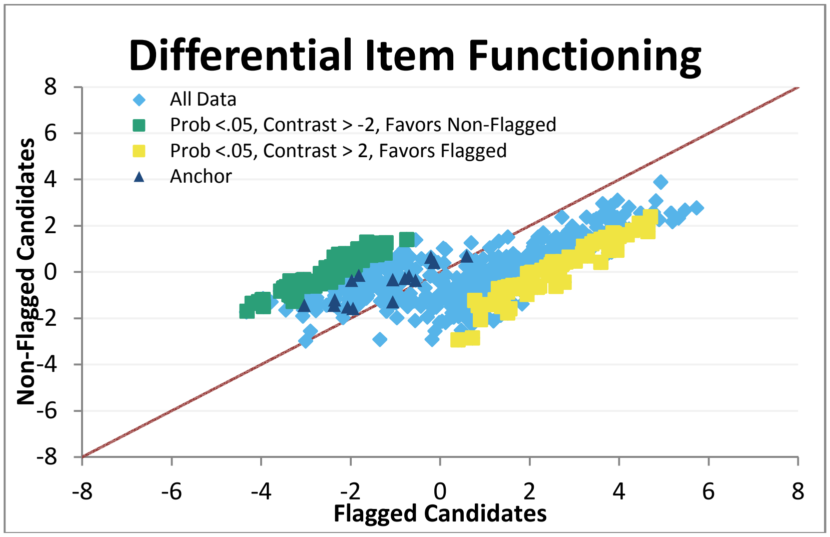

Figure 3. Differential Item Functioning by Flagging Status

In order to be able to equate the planned upgrade version of the exam to the current live version of the exam, the 169 scored items that did not display significant DIF (*p*>0.05, contrast<2.0) were considered viable for anchoring because prior knowledge of exam content was not shown to impact performance of those items. Fifteen of these items were selected to be retained as anchor items that were proportionately representative of the blueprint[2] and well-fitting to the model. Table 2 shows the item statistics for the anchor items and highlights their range in both content distribution and item difficulty.

[2] The exam contains six sections, each respectively with 2 to 11 associated objectives, with content representation as follows: Section 1 (27.5%), Section 2 (20%), Section 3 (18.75%), Section 4 (15%), Section 5 (8.75%), and Section 6 (10%).

Table 2. Anchor Item Statistics with DIF Results

Item		Rasch	P-	Item-Score	Flagged Candidate		Non-Flagged Candidates		DIF	
ID	Section	Measure	value	Correlation	DIF Measure	DIF S.E.	DIF Measure	DIF S.E.	Contrast	Prob.
6	1	0.61	0.60	0.53	-0.21	0.28	0.64	0.04	-0.85	0.003
20	1	0.42	0.64	0.49	-0.14	0.27	0.42	0.04	-0.56	0.045
32	1	-0.33	0.77	0.33	-1.06	0.39	-0.33	0.05	-0.73	0.063
47	1	-0.27	0.76	0.33	-0.79	0.29	-0.27	0.05	-0.52	0.085
57	1	-1.43	0.90	0.27	-2.37	0.7	-1.43	0.07	-0.94	0.185
66	2	0.69	0.58	0.34	0.59	0.16	0.69	0.04	-0.1	0.560
79	2	-1.57	0.91	0.28	-1.95	0.59	-1.57	0.07	-0.38	0.520
102	2	-1.43	0.90	0.25	-3.04	0.89	-1.43	0.07	-1.62	0.073
130	3	-1.29	0.89	0.25	-1.06	0.39	-1.29	0.06	0.23	0.565
151	3	-0.18	0.74	0.53	-1.82	0.42	-0.14	0.05	-1.68	0.000
160	4	-0.16	0.74	0.44	-0.7	0.28	-0.16	0.05	-0.54	0.063
175	4	-0.4	0.78	0.52	-1.98	0.42	-0.36	0.05	-1.62	0.000
182	4	-0.35	0.77	0.39	-0.55	0.27	-0.35	0.05	-0.2	0.473
200	5	-1.2	0.88	0.31	-2.36	0.69	-1.2	0.06	-1.16	0.093
212	6	-1.51	0.91	0.31	-2.07	0.51	-1.51	0.07	-0.55	0.283

An item drift analysis on the anchor items was conducted by investigating displacement from their anchored item measures in Winsteps (Linacre, 2009) on the upgraded exam version with refreshed content and item replacements for those designated as exposed. These initial analyses indicate that the anchor set's item difficulty estimates are stable within the new item bank. Table 3 displays that the drift displacement in the respective Rasch item calibrations was all less than the 0.3 logit threshold noted by Wright and Stone (1979). Furthermore, the minimal amount of item difficulty drift displayed by the anchor items followed an expectantly balanced pattern, with six items drifting in a negative (easier) and nine items moving in a positive (harder) direction.

Table 3. Anchor Item Statistics with Displacement

Item ID	Section	Number of Responses	P-value	Rasch Measure	Displacement
6	1	404	0.59	0.61	0.08
20	1	411	0.62	0.42	0.13
32	1	388	0.73	-0.33	0.29
47	1	399	0.75	-0.27	0.14
57	1	424	0.92	-1.43	-0.3
66	2	414	0.58	0.69	0.02
79	2	397	0.91	-1.57	0.09
102	2	444	0.91	-1.43	-0.01
130	3	444	0.90	-1.29	-0.02
151	3	436	0.79	-0.18	-0.24
160	4	416	0.78	-0.16	-0.14
175	4	412	0.77	-0.4	0.13
182	4	401	0.79	-0.35	-0.06
200	5	418	0.87	-1.2	0.19
212	6	395	0.89	-1.51	0.24

Conclusions

The rampant test fraud in IT certification testing programs has lead to the widespread unauthorized exposure of exam forms and perpetual item exposure. As shown in this paper, the combination of DPF and DIF can strengthen the security monitoring of a testing program with known exposure issues. The tight timeframes for piracy (within days or weeks) within IT certification testing programs gravely reduce the benefit of investing the budget and resources into developing an entirely new item bank for exam with egregious exposure issues; that content would likely be stolen and exposed too quickly for the testing program to benefit from its efforts. Instead, realistic approaches to exam and item maintenance are needed to reduce threats to the validity of score interpretation and use in continually administered exams with exposure issues by highlighting specific compromised content.

This paper presented how the use of DPF in conjunction with DIF can minimize the tangible and intangible costs of test fraud, in addition to maximizing the measurement integrity and validity associated with exams with this known security issue. Additionally, these procedures could deter future test fraud by enhancing the data forensics associated with a testing program. Wollack and Fremer (2013) note that well-articulated procedures are “a very effective way to communicate to candidates that cheaters leave behind irregular patterns of responses, and that even if they are sufficiently clever to successfully cheat on the exam, they will be unearthed by sophisticated statistical procedures being run in the background” (p. 11). It should be noted that these techniques rely on the hypothesis that only operational items have been exposed, and that unscored, pilot items have not been subject to test fraud. This assumption should be investigated and accepted by a given testing program prior to implementing these security analyses.

With that caveat, the combined use of DPF and DIF enhances a test program’s ability to maintain credible exams within a context of prevalent test fraud by identifying suspect candidates based on their own aberrant response patterns and exposed items based on bias towards candidates with item pre-knowledge. Since DPF incorporates probabilistic-statements regarding discrepancies in individual candidates’ comparative scores within a single testing instance on a respected measurement model (Rasch), the results do align with Maynes’ (2013) requirements for enforceable security analyses and could lead to defensible sanctions against candidates when combined with other evidence. In these situations, candidates’ scores could be invalidated or candidates could be banned from testing for a set period of time, among a myriad other actions. Likewise, the DIF analysis as described presents probabilistic-statements about apparent bias in particular items towards those candidates with item pre-knowledge. Therefore, the extent of item degradation can be controlled by regularly checking the extent of exposure for particular items in compromised item banks. Once exposure is detected for particular items, these items can either be replaced immediately with new, unscored pilot items or be monitored for changes in their item statistics and replaced as necessary and feasible given the status of item development and availability of new items.

The combination of these methodologies provides several practical options for testing programs trying to maintain the validity of their candidate decisions in the context of grossly exposed item banks. The comparison of scored to unscored items has been shown to be useful in identifying candidates likely of prior exam knowledge. However, specific decisions regarding the number of each of these item types to administer can be customized to the purpose of the exam, planned subsequent security analyses, and the testing program’s capacity for follow-up investigation and enforcement. If detection and enforcement are primarily goals of the DPF analyses, conservative flagging criteria can be utilized to identify candidates with the vastest discrepancies in their scores on the compromised and non-compromised content. With regard to detecting exposed items, the overall size of the item pool, tolerance for retiring and replacing items, ability to refresh content with new, unscored items, and budget and resources for continued item development should be considered. For example, testing programs should take into account the following questions, among others: How many items are in the item pool, and what percent of those are exposed? To what extent has item degradation impacted item statistics and candidate results? Do sufficient unexposed items exist to retire and replace compromised items, or is new content necessary? What is the timeline of new item development, including review? Will the production and maintenance schedule allow for the piloting new items? In circumstances with fewer resources to support wide-scale item development and form re-assembly, more conservative flagging criteria should be selected for the DIF analyses to reduce the pool of compromised items identified through the analyses.

Research still needs to be conducted to confirm these results and explore extensions of these methods, such as scale stability over time. This paper presents how these methods can be used to (1) detect when security breaches have occurred; (2) determine the extent of item exposure; (3) build cases against suspect candidates; (4) collaborate with other evidence to support the enforcement of sanctions against candidates; (5) highlight specific items with compromised content, and (6) evaluate appropriate next steps for particular items and entire item banks while discussing the relevant psychometric and policy issues for each of these areas. The application of these dual analysis efforts will help to preserve the validity of candidate decisions and the reputation of testing programs operating in an environment of grossly exposed exam content. This paper also contributes to the growing literature on data forensic techniques to gauge the impact of test fraud on testing programs via statistical and psychometric methods.

References

Association of Test Publishers (ATP) Security Committee (2013, January). Assessment Security Options: Considerations by Delivery Channel and Assessment Model. Retrieved April 11, 2013 from http://www.testpublishers.org/assets/assessment_security_options-considerations_by_delivery_channel_and_assessment_model_1-23-13.pdf.

Bergstrom, B. A., Stahl, J. A., & Netzky, B. A. (2001, April). *Factors that influence item parameter drift*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Carson, J. D. (2013). Certification/licensure testing case studies. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 261-283). New York, NY: Routledge.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it and prevent it.* Mahwah, NJ: 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, NY: Macmillan.

Foster, D. (2013). Security issues in technology-based testing. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 39-83). New York, NY: Routledge.

Fremer, J. J. & Ferrara, S. (2013). Security in large-scale paper and pencil testing. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 17-37). New York, NY: Routledge.

Han, N. (2003). Using moving averages to assess test and item security in computer-based testing. *Center for Educational Assessment Research Report No. 468.* Amherst, MA: University of Massachusetts, School of Education.

Han, N. & Hambleton, R. K. (2004, April). *Detecting exposed test items in computer-based testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Han, N. & Hambleton, R. K. (2008). Detecting exposed items in computer-based testing. In C. L. Wild & R. Ramaswamy (Eds.), *Improving testing: Applying process tools and techniques to assure quality* (pp. 423-448). New York, NY: Lawrence Erlbaum Associates.

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

Linacre, J. M. & Wright, B. D. (1987). Item bias: Mantel-Haenszel and the Rasch model. *Memorandum No. 39 MESA Psychometric Laboratory*. University of Chicago, Department of Education.

Lu, Y., & Hambleton, R. K. (2003). Statistics for detecting disclosed item in a CAT environment. *Center for Education Assessment Research Report No. 498*. Amherst, MA: University of Massachusetts, School of Education.

Maynes, D. (2013). Educator cheating and the statistical detection of group-based test security threats. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 173-199). New York, NY: Routledge.

Maynes, D. D. (2011, April). *A method for measuring performance inconsistency by using score differences.* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Maynes, D. (2009). Caveon speaks out on IT exam security: The last five years. Retrieved April 10, 2013 from http://www.caveon.com/articles/IT_Exam_Security.pdf.

Smith, R. W. & Davis-Becker, S. (2011, April). *Detecting suspect candidates: An application of differential person functioning analysis*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Smith, R.W. (2004, April). *The impact of braindump sites on item exposure and item parameter drift*. Paper presented at annual meeting of the American Educational Research Association, San Diego, CA.

Veerkamp, W. J. J. & Glas, C. A. W. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics, 25,* 373-389.

Wollack, J. A. & Fremer, J. J. (2013). Introduction: The test security threat. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 1-13). New York, NY: Routledge.

Wright, B. D. & Stone, M. (1979). *Best test design.* Chicago, IL: MESA Press.

Yang, Y., Ferdous, A., Chin, T. Y. (2007, April). *Exposed items detection in personnel selection assessment: An exploration of new item statistic.* Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.