AFRL-RH-WP-TR-2019-0042

SUMMARY OF LITERATURE REVIEW OF ITEM EXPOSURE, TEST SECURITY, AND FORENSICS FOR THE WEIGHTED AIRMAN PROMOTION SYSTEM (WAPS)

Gordon S. Waugh
Nicholas J. Walion
Timothy C. Burgoyne
Rodney A. McCloy

Human Resources Research Organization (HumRRO)

May 2019
Interim Report

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
AIRMAN SYSTEMS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2019-0042 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signature//
THOMAS R. CARRETTA
Work Unit Manager
Supervisory Control and Cognition Branch
Warfighter Interface Division

//signature//
MAJ. JOSEPH C. PRICE
Chief, Supervisory Control and Cognition Branch
Warfighter Interface Division

//signature//
LOUISE A. CARTER
Chief, Warfighter Interface Division
Airman Systems Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
19-05-19	Interim	02 JAN 18 – 19 APR 19

4. TITLE AND SUBTITLE
Summary of Literature Review of Item Exposure, Test Security, and Forensics for the Weighted Airman Promotion System (WAPS)

5a. CONTRACT NUMBER
FA8650-14-D-6500, TO 0007

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
Gordon S. Waugh, Nicholas J. Walion, Timothy C. Burgoyne, Rodney A. McCloy

5d. PROJECT NUMBER
5329

5e. TASK NUMBER
09

5f. WORK UNIT NUMBER
H0SA (532909TC)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Human Resources Research Organization (HumRRO)
66 Canal Center Plaza, Suite 700
Alexandria, VA 22314-1578

8. PERFORMING ORGANIZATION REPORT NUMBER
2019 No. 004

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory
711th Human Performance Wing
Airman Systems Directorate,
Warfighter Interface Division,
Supervisory Control & Cognition Branch
Wright-Patterson AFB, OH 45433
Air Force Materiel Command

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
711 HPW/RHCI

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR-2019-0042

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A: Approved for public release

13. SUPPLEMENTARY NOTES
MSC2019-0272; 88ABW-2019-3663 cleared 21 August 2019

14. ABSTRACT

The U.S. Air Force (AF) uses the Specialty Knowledge Test (SKT) and Promotion Fitness Examination (PFE) as part of its Weighted Airman Promotion System (WAPS). These tests are presently administered once per year for pay grades E-5 (Staff Sergeant) and E-6 (Technical Sergeant). Because candidates for promotion not selected during their first year of eligibility can retest as long as they remain eligible for the targeted higher rank (a period of 5 to 8 years), Airmen can complete the SKT or PFE on multiple occasions. Thus, promotion candidates could complete the SKT for E-6 as many as four to seven times. This leads to the likelihood that candidates will see some of the same test items upon re-administration. Elevated item exposure raises questions about the potential for test compromise for later test administrations provided to a given candidate.

To help address AF concerns regarding the effects of item exposure on WAPS test performance, HumRRO examined the testing literature concerning item exposure, test compromise, forensics to detect item/test compromise, and test security in general. These topics are discussed in three sections: (a) Test Security, Item Exposure, and Test Compromise; (b) Test Forensics; and (c) Recommendations.

15. SUBJECT TERMS
Data forensics, item exposure, promotion fitness exam, retesting, specialty knowledge test, test compromise, test forensics, test security

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON (Monitor)
a. REPORT	b. ABSTRACT	c. THIS PAGE			Thomas R. Carretta
Unclassified	Unclassified	Unclassified	SAR	38	19b. TELEPHONE NUMBER *(Include Area Code)* (937) 713-7143

TABLE OF CONTENTS

1.0 OVERVIEW

The U.S. Air Force (AF) uses the Specialty Knowledge Test (SKT) and Promotion Fitness Examination (PFE) as part of its Weighted Airman Promotion System (WAPS). These tests are presently administered once per year for pay grades E-5 (Staff Sergeant) and E-6 (Technical Sergeant). Because candidates for promotion who are not selected during their first year of eligibility can test again as long as they remain eligible for the targeted higher rank (a period of 5 to 8 years), Airmen have the opportunity to complete the SKT or PFE on multiple occasions. Thus, promotion candidates could complete the SKT for E-6 as many as four to seven times. This leads to the likelihood that candidates will see some of the same test items upon re-administration. Elevated item exposure raises questions about the potential for test compromise for later test administrations provided to a given candidate.

To help address AF concerns regarding the effects of item exposure on WAPS test performance, HumRRO examined the testing literature concerning item exposure, test compromise, forensics to detect item/test compromise, and test security in general. These topics are discussed in three sections:

- Test Security, Item Exposure, and Test Compromise;
- Test Forensics; and
- Recommendations.

2.0 TEST SECURITY, ITEM EXPOSURE, AND TEST COMPROMISE

Test security matters because it safeguards test validity, ensures test fairness, and limits test compromise (Ferrara, 2017). Failures in test security create several costs to organizations:

- reduction in test validity, which reduces the quality of employees hired and promoted;
- replacement of compromised test items;
- reduction in the perceived fairness of the selection/promotion system, which makes it more difficult to recruit qualified employees and reduces morale among current staff; and
- risk of litigation.

To minimize these costs, sound test security policies and procedures should be developed and adhered to in all phases of test development and administration. Test security can be compromised in several ways:

- Examinees discover test content before the exam. This is called *pre-knowledge*.
- During the exam, examinees communicate with people or refer to cheat sheets or similar aids. Examples of communication include copying answers and receiving text messages.

Although it is not really a breach in test security, examinees can also benefit by remembering items they took during a previous test administration. After taking the exam, they can write down the content they remember and study that content prior to retaking the test in the future. This can reduce validity when examinees are exposed to some of the same items on multiple test administrations. This can occur when examinees retake an exam using the same test form (or overlapping test forms), or when the test forms for different promotion levels share items.

2.1 Best Practices for Prevention of Test Compromise

As mentioned above, one form of test compromise involves pre-knowledge—that is, examinees discovering some test content before the test administration. Pre-knowledge can occur in several ways:

- During test development, items are leaked (either purposely or inadvertently) by item writers or reviewers.
- Testing staff members leak items (either purposely or inadvertently).
- Electronic files or e-mails containing items are obtained by unauthorized persons.
- Previous examinees tell some future examinees about the test content they remember. These future examinees can then pass the content along to other future examinees.
- Some items on the current test were also on a previous test taken by the examinee.
- People who sell stolen test content hire examinees to memorize or copy (e.g., secretly taking pictures of the items) test items. This is called *item harvesting*.

Preventative measures can be taken to reduce the risk of such test compromise. Regardless of the organization and the test, many of the best practices concerning methods to prevent pre-

knowledge from occurring in the first place remain the same. The best practices outlined below reflect general principles of test security and standards that are applicable across many testing settings.

During test development, it is important to explain security risks and procedures to test contributors (e.g., test developers, SME's) (Wollack & Fremer, 2013). Here is a list of best practices:

- Official letters remind test contributors about security concerns and that they must not be involved with test preparation—formally or informally.
- Confidentiality agreements require that test contributors refrain from discussing test content, have no conflict of interest, and not take the exam within 18 months.
- Security reminders are used to remind test contributors that they (a) must receive any shipments of exam materials in person, (b) package exam materials separate from other materials and be labeled "confidential," (c) lock up secure paper exam materials when not in use, and (d) delete or shred exam materials when the materials are no longer needed.
- Formalized security training is provided for test contributors and/or contributors are asked to take a written security pledge. Security training and security pledges are not often implemented but would have little downside.

Examinees tend to share information with each other before and after a test administration. Best practices limit item exposure due to examinees sharing information (Wollack & Fremer, 2013). These practices are as follows:

- Test administrators clearly convey to examinees that test content is copyright-protected so exam users know there are legal repercussions for sharing materials after the test.
- Test developers become familiar with all websites designed for exchange of test information among likely examinees.
- Test developers are prepared to immediately email a "cease and desist" letter (prepared in advance) to hosts of online posts who reveal too much content about the test.

Finally, to prevent instances of test compromise, organizations can conduct a test security audit (Olson & Fremer, 2013). The audit should be conducted under the following specifications:

- Conduct the audit using expert staff from outside the organization, such as (a) a consultant specializing in test security or (b) the organization's test development and/or delivery vendors.
- Use staff with firsthand knowledge of the organization's testing policies and practices.
- Use formal test security standards to help guide the audit.
- Identify security vulnerabilities as well as recommendations for dealing with the identified vulnerabilities.

2.2 Impact of Test Construction on Test Security

During test construction, there are multiple decisions that affect the level of item exposure and overall test security. These decisions include the size and management of item pools, the construction of test forms, and the type of items chosen for the test. One key decision is whether to use traditional paper-and-pencil testing or computer-based testing (CBT). Paper-and-pencil testing has the added security risk of exposing the content to additional people during printing, distributing, and proctoring. In paper-and-pencil testing, having many test forms and long tests makes the test more resistant to cheating, but some types of CBT such as computerized adaptive testing (CAT) and randomized item sequencing (on the fly) are more resistant to cheating than are paper-and-pencil tests (Guo, Tay, & Drasgow, 2009). The use of paper-and-pencil vs. CBT formats will affect subsequent decisions that can be made regarding test security. The same principles of test security apply to paper-and-pencil tests and CBT, but they are implemented using different methods due to the limited nature of paper-and-pencil tests. CBT offers more freedom with regard to decisions about test security relative to paper-and-pencil tests. Most of the strategies discussed below (regarding item pools, the construction of test forms, and item types) are impossible or difficult to implement using paper-and-pencil testing.

According to test standards stipulated by Caveon Test Security (Olsen & Fremer, 2013), tests should be designed to "discourage memorization and sharing and make common methods of cheating less effective. They should limit item exposure, thereby prolonging the usefulness of test items and test results" (p. 59). This can be done by using a variety of methods and following various rules of thumb. With regard to the item pool, the minimum item pool size is affected by factors such as the length of the testing window, the number of administrations per year, and the number of examinees per window. Under the best of circumstances (e.g., administered one day per year, small candidate volume, one test form per administration), item pools should, at a minimum, contain twice the number of items that are currently in the operational test form (Wollack & Fremer, 2013). Rules about the size of item pools also vary by test type. For example, for a CAT, it is recommended to have item pools that are 8 to 10 times as large as the tests (Mills & Stocking, 1996). More items in the overall item pool reduce the risk of test compromise by limiting item exposure (particularly when item exposure algorithms are employed to ensure equal exposure of items throughout the pool).

Item pools also should be strategically managed to reduce item exposure and the risk of test compromise. One such strategy is to give items periodic "vacations" from operational test forms and shuffle item pools unpredictably to substantially increase the number of exposed items that an examinee would have to memorize to gain an advantage (Mills, Potenza, Fremer, & Ward, 2005). Another strategy for item pool management, especially with large item pools, is to break up the item pools into more manageable subsections (Mills & Stocking, 1996). Separate subdivisions of item pools would decrease the amount of item exposure that a single security break could impose.

Randomizing item order within a given test form is another way to enhance test security. If items are presented in a different order to every examinee, it will be more difficult for examinees to communicate the correct answers, copy each other's answers, or use a stolen answer key. Randomizing item order would also make it more difficult for an examinee to recall correct answers on the test. If test items retain the same order, then at a later retesting date the examinee

may have greater recall of the test items if they cue one another in memory. An advanced form of item randomization is called linear on-the-fly-testing (LOFT). LOFT generates randomized equivalent test forms by randomly selecting the item to be presented to the examinee from the total item pool (Gibson & Weiner, 1998). LOFT further increases the security of the test by giving each examinee a unique test form. LOFT is a compromise between multiple fixed forms testing and computerized adaptive testing (CAT), and it can be implemented using statistical equating through Item Response Theory (IRT) statistics or Classical Test Theory (CCT) statistics. LOFT is preferred over CAT when the test must have a specific number of items in each topic area (or in other item characteristic categories such as cognitive level).

Some testing programs use automated item generation to generate test items during test development or even during a test administration. The use of computer technology that automatically generates item variants from a parent item enhances test security (Bennet, 1999). *Item cloning* takes advantage of this approach. Item cloning is used to create items that mirror each other on item statistics and the construct assessed but differ slightly in the actual wording or content of the item. "Parent items"—also known as "item forms," "item templates," or "item shells" —are created along with algorithms to derive families of clones (Osburn, 1968). Computer algorithms determine the item content based on substitution sets governed by rules. For instance, replacement-set procedures can pick distractors randomly from a list of possible wrong answers or substitute random elements in open places in the stem of the item and adjust the alternatives accordingly. Item clones created from parent items should be created with substitution sets that artificially differ in content but do not substantively change the meaning of the construct the item assesses. Upon retesting with item cloning, an examinee would be less likely to recognize that two items are assessing identical content, prolonging the shelf life of the items and reducing the rate of item exposure. Clearly, automated item generation is much more appropriate for some types of tests (e.g., arithmetic) than others (e.g., reading comprehension).

Discrete Option Multiple Choice (DOMC) items also decrease the rate of item exposure by changing the way that the examinee is presented with the item. With DOMC items, examinees are presented with each item response option one at a time and must assess whether the response option is correct or incorrect independently of other response options. The option that is presented to an examinee is chosen at random by an algorithm. Response options will no longer continue being presented to examinees for a given item if the examinee provides either (a) the incorrect answer for a distractor or (b) the correct or incorrect answer for the keyed response. For example, consider an examinee presented with a sample item that has four response options (A, B, C, and D) where option C is keyed as the correct answer. The computer randomly chooses option B for presentation to the examinee with the stem. If the examinee indicates (incorrectly) that option B is the right answer, then the entire item is scored as incorrect for the examinee, and they are not presented with any more response options for that particular item. If the examinee indicates (correctly) that option B is not the correct answer, the computer would randomly select one of the three other response options (A, C, or D) to present to that examinee. Suppose the computer next selects option C to present to the examinee, and the examinee (correctly) indicates that this is the right answer. Then the entire item is scored as correct, and the examinee moves on to the next item. If the examinee (incorrectly) indicates that option C is not the right answer, then the item is scored as incorrect and the examinee moves on to the next item. Using this approach, not every examinee will see all the response options for every question, thus limiting the test

content that is exposed. DOMC items are a relatively new innovation in testing (Foster & Miller, 2009), but initial research indicates they may reduce the impact of test-wiseness, which is the ability to find subtle cues towards the solution by the simultaneous comparison of available response options (Willing, Ostapczuk, & Musch, 2015). However, there are some disadvantages to DOMC items, because different examinees see different versions of each item. As a result, the item's difficulty differs for different examinees, and an item might measure different things for different examinees.

2.3 Effects of Retesting on Validity and Subgroup Performance

Several issues arise when retesting in operational employment settings. First, there is the issue of the validity of retest scores and whether they are as valid as scores from the initial administration. Are score gains in retesting due largely to construct-irrelevant factors such as increased test-wiseness, coaching, or pre-knowledge of some items? A second issue with retesting concerns differences in subgroup performance on retests. Offering multiple retesting opportunities may level the playing field for some subgroups that might be at a disadvantage in terms of test-wiseness or test anxiety. Conversely, retesting opportunities could exacerbate existing subgroup differences in test performance.

Research has demonstrated that retest scores for job knowledge tests are more valid than initial tests (Lievens, Buyse, & Sackett, 2005; Van Iddekinge, Morgeson, Schleicher, & Campion, 2011). Of particular relevance to the AF, retesting did in fact enhance criterion-related validity for job knowledge tests. That is, the scores on the retest were slightly better predictors than the scores on the initial test. These effects of retest gains in validity for job knowledge tests have also been shown to generalize to real-world employment settings (Van Iddekinge et al., 2011). In addition to retest scores on job knowledge tests being better predictors of performance than initial scores, retest scores are typically higher than initial test scores.

Gains in test scores are not necessarily equivalent across subgroups, however. Research has investigated how retesting affects subgroup performance (Randall, Villado, & Zimmer, 2016; Schleicher, Van Iddekinge, Morgeson, & Campion, 2010; Van Iddekinge et al., 2011). Schleicher et al. (2010) investigated the effects of retesting and differences in subgroup performance, including Black-White differences, White-Hispanic differences, White-Asian differences, Gender differences, and Age differences. They also looked at how these results may vary across different test types, including verbal ability tests, job knowledge tests, biodata tests, interviews, a leaderless group exercise, and a case analysis exercise. When comparing relative score gains among subgroups, they found that Whites' test scores improved more than Blacks' or Hispanics' test scores. Whites improved more than Hispanics and Blacks on what were classified as written tests: job knowledge (White $d = 0.21$, Hispanic $d = 0.10$, Black $d = 0.08$), biodata (White $d = 0.37$, Hispanic $d = 0.22$, Black $d = 0.17$), and verbal ability (White $d = 0.12$, Hispanic $d = 0.06$, Black $d = 0.04$).[1] No retest differences by race were found for performance-based tests (i.e., interviews, a leaderless group exercise, and a case analysis exercise). In addition, women and applicants under 40 years of age showed larger improvements with retesting than did men

[1] The *d* statistic is Cohen's *d* – the standardized mean difference between test scores from a reference group (say, Whites or Males) and a focal group (say, Blacks or Females). The difference is expressed in units based on the standard deviation as calculated on either the reference group or the entire sample (i.e., a pooled standard deviation).

and applicants over 40. Subsequent investigations assessed the effects of retesting on validity and subgroup differences, specifically in a job knowledge context (Van Iddekinge et al., 2011). Consistent with other research, evidence was found that retesting may reduce the likelihood of adverse impact against some subgroups (e.g., female candidates) but increase the likelihood of adverse impact against other subgroups (e.g., older candidates).

Investigating how subgroup differences among retest scores may contribute to adverse impact, Randall et al. (2016) contended that even though some groups (e.g., whites and women) may have greater score gains than others, the general increase in scores and the increase in validity of retests may play a role in decreasing adverse impact. This could occur if score increases result in a higher number of applicants from protected groups meeting established cutoff scores. Therefore, although some groups may outperform others on retests, even the groups being outperformed still have score gains that could decrease the rate of adverse impact if the score gains result in higher rates of applicants being promoted or selected. However, if the number of applicants who are ultimately selected does not change regardless of score gains due to retesting, (e.g., rank order selection) then adverse impact will not be ameliorated. Green and McCloy (2018) provide a recent review of best practices regarding testing policy.

2.4 Test Security in Police and Fire Departments

In police settings, it is common for promotion testing to be performed with only one testing window, so as to limit the amount of information-sharing that occurs after the test. We are unaware of instances in police promotion testing where the test is completed more than once by any eligible candidates, unless there are extreme extenuating circumstances that would warrant a retest. Indeed, we are aware of no other setting where candidates take the same promotion test multiple times throughout their career to be promoted to different pay grades. However, we investigated how police and fire departments have dealt with test security issues and present these examples below.

2.4.1 The Los Angeles Sheriff's Department

The Los Angeles Sheriff's Department uses a paper-and-pencil test with one test administration period. They use one testing window to prevent items from being exposed to other candidates after the administration. The paper-and-pencil testing precludes them from using CAT, because they test 1,500 candidates at a time and do not have a testing facility that could accommodate 1,500 CATs at one time.

To improve their test security, the Los Angeles Sheriff's Department implemented a stand-alone server which is accessible only to the test developers. Since the Department implemented this server, the mean score for their promotion test has fallen 1.5 to 2.0 standard deviations and has remained at that level. Before this change, they often had one or two candidates earning a close to perfect score; now, no one scores higher than in the low 90% range. Additional controls were implemented to prevent dissemination of test material.

2.4.2 Jefferson County Fire Department

The Jefferson County (Alabama) Fire Department discovered that a small group of SMEs was leaking content for their promotional exam, even though they had security measures in place. They subsequently made changes to their test development and test administration procedures to enhance test security. During test development, the Department began (a) to use out-of-state SMEs rather than local SMEs, (b) to question SMEs about their relationships with local fire personnel, (c) to inform SMEs of prior cheating issues and require them to sign a confidentiality agreement, and (d) not to disclose the test development efforts to the local fire departments. They also made changes to test administration procedures, which included confiscating cell phones during testing, assigning testing rooms and seating, and not allowing any outside non-fire materials to be brought into testing. They found that the new procedures and controls did not result in adverse impact against any group, and test scores declined.

3.0 TEST FORENSICS

Test compromise and fraud are major concerns for organizations. Each year, millions of dollars are spent on test development (Chingos, 2012; Meinert, 2015). When cheating occurs, or items become compromised, a burden is placed on the organization to respond to these test integrity concerns, which in turn requires additional time, money, and resources (O'Leary & Smith, 2017). But the expense of developing tests is not the only reason cheating is an issue. Test compromise undermines the validity of the assessments (Cizek & Wollack, 2017). Test scores are used to determine that examinees have obtained a certain level of proficiency in the area being tested. In the presence of cheating or item compromise, we can no longer be confident in the conclusions drawn from these scores. This is especially troublesome when important outcomes like job promotion are affected by test performance.

To support the validity of inferences made from test scores, many organizations incorporate cheating detection methods into their testing programs. The forensic tools and practices available can be categorized into five distinct types of methods that function to detect the following: (a) answer copying between pairs of examinees, (b) aberrant responding, (c) examinees who had pre-knowledge of the item content (usually due to item compromise), (d) unusual answer-changing behavior, and (e) unusual gains in scores between two testing periods (Cizek & Wollack, 2017; Kingston & Clark, 2014). In most situations, statistical methods and practices provide evidence rather than conclusive proof of test fraud. They are tools that assign a probability that cheating occurred and can offer reasoned conclusions based on statistical evidence (Cizek & Wollack, 2017). Cheating by a group (e.g., an entire school) is easier to detect than cheating by a single person. Researchers disagree whether statistical evidence alone is sufficient to conclusively prove that an individual has cheated. Organizations typically ask an individual to retest, under greater scrutiny using an uncompromised test form, when there is statistical evidence of cheating.

3.1 Answer Copying

A standard practice to detect test fraud is to examine response data for evidence of answer copying. To detect answer copying, it is important to understand what response behavior looks like in the absence of answer copying. It is expected that, to some extent, examinees who share similar experiences (e.g., received the same training, had the same instructors) would also have a similar understanding of the test material and that this common understanding might manifest in their responses to exam items (Allen, 2014). Accordingly, a baseline needs to be established for similarity in response patterns so that those who have similar responses because of non-independent test-taking behavior can be distinguished from those who have similar responses for reasons unrelated to test fraud.

Response similarity indices are the quantitative tools used by organizations to detect these true instances of answer copying. They are used to evaluate the probability of agreement between the response vectors for two examinees given the assumption that the examinees were taking the exam independently (Zopluoglu, 2017). Generally, response similarity indices vary in two primary ways: (a) their definition of agreement between two response vectors, and (b) the statistical distributions they depend on to evaluate what is considered anomalous response behavior.

Agreement between response vectors can be defined as matching incorrect responses between examinees. Indices like the K index and a few of its variants (i.e., K_1, K_2, S_1) endorse this definition (Zopluoglu, 2016). Approaches like the ω index and S_2 consider agreement between vectors in terms of matching both incorrect and correct responses. In addition, some indices take into account all responses and consider matching responses to imply copying and non-matching responses to imply the absence of copying (Zopluoglu, 2017).

Response similarity indices can also be differentiated by the distributions on which they are based. Some indices are based on empirical null distributions (i.e., the distribution when there is no copying), whereas others rely on already established statistical reference (i.e., theoretical) null distributions (e.g., normal, binomial; Zopluoglu, 2017). To create an empirical null distribution for the number of shared responses between two examinees, an organization matches examinees on a third variable (e.g., geographical location, test site) so that the data reflect pairs of examinees for which answer copying is not possible. Creating this empirical null distribution requires access to a large dataset so that the probabilities of response matches can be estimated accurately (Sotaridona, Wibowo, & Hendrawan, 2014). When an empirical distribution is used, confidence in the results is increased when a method based on a theoretical distribution finds similar results.

After a null distribution is developed, goodness-of-fit analyses must be conducted to ensure that an index's statistical assumptions have been met (Sotaridona et al., 2014). Methods based on an empirical null distribution are too cumbersome for some testing programs because of the additional analyses required. With WAPS, for example, several empirical null distributions would have to be developed. Because matched responses between examinees will likely vary depending on the test items used, abilities of the examinees, and myriad other variables, a distribution would not be able to be used interchangeably across different test forms, job functions, or administration times. A separate distribution must be created for each specific testing instance (Zopluoglu, 2017). For these reasons, this approach has not received much attention in the literature. Accordingly, we do not see it as a useful approach for the WAPS testing program.

As mentioned above, an alternative to developing an empirical null distribution is to use response similarity indices that are based on existing reference statistical distributions (Zopluoglu, 2017). To evaluate the performance of these indices, the Type I error rate and statistical power should be considered. With response similarity indices, Type I error rate is defined as the proportion of honest examinee pairs who are incorrectly identified as copiers (i.e., false positives), whereas power is defined as the proportion of copying examinee pairs who are correctly identified as copiers (i.e., true positives). Most response similarity indices perform sufficiently well and have empirical Type I error rates below the theoretical level for both dichotomous (e.g., multiple-choice) and nominal response (e.g., 1-5 rating scale) outcomes. However, the ω index (Wollack, 1997) performs best in terms of controlling nominal Type I error rate as indicated by the empirical Type I error rates being closest to the nominal (i.e., true) levels (Zopluoglu, 2017). With respect to statistical power, the ω index is also the best method for correctly detecting examinee pairs who copied. Other methods like the K index and GBT

index have relatively low empirical Type I error rates, but their conservative nature comes at the expense of power, leading to poorer detection of true copying pairs than the ω index provides.

The ω index is a copying statistic that is based on the normal distribution. It compares the observed agreement of both incorrect and correct responses between two examinees to the expected agreement (Zopluoglu, 2017; Sunbul & Yormaz, 2018). The formula for the ω index is derived from item parameters based on the nominal response model (Bock, 1972)[2] of item response theory (IRT). This model is used to estimate the expected values (i.e., probability of endorsement) of each alternative response while taking into account the ability of each examinee in a pair of examinees (Sunbul & Yormaz, 2018). However, this model requires creating and using an option matrix that can be difficult to manage. Therefore, many organizations use dichotomous IRT models like the 1-parameter, 2-parameter, 3-parameter, and Rasch models when using the ω index (Assess Systems, 2016).

Although response similarity indices like the ω index are not necessarily difficult to employ, it may not be worth the effort to use them if answer copying is not a concern for the AF. The AF Instruction (AFI) 36-2605 lists specific requirements for test facilities. For example, the facility should have a minimum space of 15 square feet per examinee. This, combined with the requirement that the test facility be setup in a way that ensures the test examiner can see and hear examinees at all times, suggests that necessary measures have been taken to discourage answer copying. Assuming these guidelines are adhered to, it seems unlikely that answer copying would be a widespread issue for the WAPS testing program unless examinees communicate via mobile devices during the exam. Nevertheless, formulas for computing the ω index can be found in Appendix A.

3.2 Aberrant Responding

Another forensic approach involves detection of aberrant responding. For example, it is normal for an examinee to do better on easy items than on hard items. If an examinee does just as well (or better) on many of the hard items as on the easy ones, that is evidence for some type of cheating, such as getting some of the test content ahead of time. Aberrant responding can also occur when test items are field tested using non-candidates who have little or no reward for doing well on the test. In that case, some candidates can become unmotivated during the exam and start responding randomly or carelessly. Aberrant responding is an issue because organizations want to be confident that a test is reliable and valid in assessing examinees' ability on the construct being measured. The presence of irregular response patterns is a potential indicator of test fraud; it undermines the credibility of the test (Cizek & Wollack, 2017) and leads to some unqualified candidates being hired, promoted, etc.

Many organizations use person-fit indices as a method of detection. Unlike response similarity indices, which compare the response patterns of two examinees, person-fit indices compare an examinee's response vector to the response patterns of other examinees under a statistical model of interest (Zopluoglu, 2017). For example, an examinee should tend to do much better on the

[2] The nominal response model is a polytomous IRT model that is an extension of the 2PL IRT (2-parameter logistic item response theory) model. A nominal test item contains unordered categorical options.

easier items. If an examinee's responding deviates from what is expected under the model, it is flagged for further review.

There are two types of person-fit indices: parametric and non-parametric. The former identifies when an examinee's responding is inconsistent with the response pattern expected from an IRT model. Examples of parametric person-fit indices include the l_z index, the l_{z*} index, and Cumulative Sum (CUSUM). Non-parametric person-fit indices derive expected responses using classical test theory or Guttman scales (Guttman, 1944). U3 and H^T are two commonly used nonparametric person-fit indices.

Using person-fit indices to identify aberrant responding has received little attention in the literature (Zopluoglu, 2017). This can be attributed to the fact that aberrant response patterns can be the result of many factors that are unrelated to test compromise or fraud (e.g., careless responding, guessing). Therefore, the usefulness of the information derived from these indices is questionable. When person-fit indices are used to flag aberrant responding alone, the l_z, l_{z*}, and U3 indices tend to flag the same examinees (Kim, Woo, & Dickison, 2017). The main difference between these indices is that the first two require IRT parameter estimates, whereas the U3 index does not. Despite being variations on the same index, the CUSUM indices (i.e., CUSUM C^+ and CUSUM C^-) tend to flag very different examinees as aberrant. The CUSUM indices tend to be more liberal in flagging examinees than other indices. The CUSUM C^- index tends to flag examinees whose performance is better toward the latter part of the test (i.e., it takes these examinees time to "warm up"). Although there are many person-fit indices to choose from, it is unclear which performs best because all have been shown to flag only a subset of the examinees flagged by test data providers.

Person-fit indices have also been used to detect answer copying but have been criticized for being underpowered when used for this purpose (Zopluoglu, 2017). Overall, they perform worse than response similarity indices when used to classify honest pairs and copying pairs. D_Θ and H^T (Sijtsma, 1986) are two indices that have been shown to perform acceptably. However, the effectiveness of person-fit indices can vary depending on the amount of copying that occurs and the abilities of the examinee pairs. For example, in one simulation, D_Θ had comparable performance to response similarity indices when a low-ability examinee copies 40% or 60% of items from a high-ability examinee, but at other variations in ability and percentages of items, it did not perform as well (Zopluoglu, 2017). In contrast, H^T is a bit more stable in its detection across conditions and has been found to perform better relative to the other person-fit indices (Dimitrov & Smith, 2006). H^T is also particularly easy to interpret. The H^T statistic ranges from 0 to 1, with aberrance indicated when $H^T < .3$ (Sijtsma & Meijer, 1992). The formula for computing H^T can be found in Appendix A.

Because person-fit indices are considered underpowered when used for answer copying, some argue that instead of using person-fit indices alone, a combination of person-fit indices and response similarity indices should be used in forensic detection (Belov & Armstrong, 2010). Specifically, a two-step approach has been recommended: use a (a) person-fit index to flag potential copiers (i.e., flag examinees whose response pattern differs from the typical examinee) and (b) response-similarity index to identify if agreement between examinees within close proximity is unusual. For the WAPS testing program, though, we do not feel that using person-fit

indices would be particularly productive. They perform worse than response similarity indices in the detection of answer copying and do not yield specific information about the possible cause of the aberrance when used for aberrant response detection. As we discuss in subsequent sections, other detection methods identify aberrance and provide information that is potentially more useful (e.g., aberrant responding that suggest pre-knowledge, aberrant gain scores that suggest answer changing post-exam).

3.3 Item Compromise and Pre-Knowledge

Up to this point, we have discussed methods that analyze response patterns for non-independent test taking or irregularities. These issues are usually localized to an individual examinee or pair of examinees. However, test compromise can also occur because of a more widespread issue, such as when many examinees have pre-knowledge of test items. This can be the result of a variety of factors, including

- items being overexposed because of continuous testing windows that use the same test form,
- instructors sharing items with examinees before the test,
- examinees sharing information about items with other examinees who have not taken the test yet, or
- items being fraudulently acquired and posted to a "brain dump" website (Eckerly, 2017).

Pre-knowledge is a serious issue that needs to be addressed. The methods used by organizations to detect item pre-knowledge and compromise can be categorized into four distinct approaches that have methods to detect suspect individuals, groups, or items. Each method detects one of the following: (a) individual examinees who may have had prior knowledge of item content, (b) the specific items that may have been compromised, (c) the individual examinees who may have benefited from pre-knowledge *and* the items that may have been compromised, or (d) groups of examinees who may have had prior knowledge of item content.

The Deterministic Gated Item Response Theory Model (DGM) and the Scale-Purified Deterministic Gated Item Response Theory Model are two approaches that have been used to detect individual examinees who may have had pre-knowledge of item content (Eckerly, 2017). These methods are employed when an organization is fairly certain that a subset of test items have been compromised. This is because the formulas used with these methods require that items be specified as either *compromised* or *secure*. This information, in conjunction with examinee performance and item parameters, is used to classify examinees into two groups: those who performed (a) better on the compromised items (i.e., the pre-knowledge examinees), and (b) equivalent or better on the secure items. Through simulations, the proportion of the time an examinee is assigned to each of the two groups is determined. Examinees who are assigned to the pre-knowledge group above a specified proportion of the time are flagged. The primary difference between the two methods is that the scale-purified DGM uses a scale-purification procedure to reduce biases in the item and person parameter estimates. This adaptation of the original DGM decreases the false positive rate (i.e., proportion of examinees incorrectly identified as having pre-knowledge) and, under most conditions, increases the true detection rates (i.e., proportion of examinees correctly identified as having pre-knowledge). However, the

performance of the scale-purified DGM is markedly better than the original DGM only when there is a high base rate of examinees benefiting from prior item knowledge. At low base rates, the scale-purified DGM performs only marginally better than the original DGM. The equations for the DGM and scale-purified DGM can be found in Appendix A.

A concern with using either of these methods is that they rely on the user having accurate knowledge of which items have and have not been compromised (Eckerly, 2017). In some cases, an organization can be fairly certain that a group of items has been compromised (e.g., the organization discovers test content online). However, there might be additional compromised test content unbeknownst to the organization. Moreover, it is possible that items may have been compromised but not necessarily exposed to a large number of examinees. Therefore, it seems unwise to make designations about an item's state of compromise. Nevertheless, it is our understanding that WAPS testing program staff has, or will have, a fairly strong sense of items that have been overexposed (e.g., knowledge of how long items have been operational, the number of administrations and forms that have used the items, results of the archival data analysis) and is aware of brain dump websites (if any) that exist. Therefore, we believe these methods are still likely to be worthwhile.

If either of these methods is used, simulations need to be conducted (Eckerly, 2017). Expected false positive rates are sensitive to the number of items that have been identified as compromised. Therefore, to correctly interpret and draw conclusions using these detection methods, we need to understand what these rates look like in the absence of pre-knowledge. The analyses would involve simulating data that reflect conditions under which no compromise has occurred but for which the item and person parameters are derived from historical data for the exam size. This is necessary to compare the false positive rates under these simulated conditions with those that are observed. If the rates between these conditions are similar, it would provide support for item flags being errant. If there is a large difference in these rates, it would support the idea that items were compromised and would also give an indication of the extent of the compromise.

There are other methods used to identify individuals benefiting from pre-knowledge that are simpler and less time-intensive. For example, some organizations use Trojan-horse items to flag individuals who have benefited from prior knowledge of the answer key (Eckerly, 2017). This particular practice involves miskeying several items on the exam scoring key and relies on the principle that if the key is lost or stolen, examinees will put stock in the idea that the key is unequivocally correct. For each examinee, the probability that he/she answered a particular Trojan-horse item incorrectly given his/her ability (i.e., score on the operational items) is estimated. Examinees who were likely to answer the item correctly (given their relatively high ability) but got it wrong are flagged. Although this seems like a relatively simple practice to employ, it only benefits the organization if the most egregious form of test compromise occurs—exposure of the answer key. Furthermore, examinees with high ability may recognize when an item is miskeyed and respond based on their knowledge of the content being tested instead of depending on the key. However, if the analyses and simulations associated with the DGM and scale-purified DGM seem a bit overwhelming, this may be a useful alternative.

More commonly though, knowledge about which items have been compromised is unavailable. Fortunately, there are detection methods to identify these specific items. One of the most prominent ways is using the method of moving averages to detect changes in an item's difficulty (Han, 2003) over time. The moving-averages method shows how the average *p*-value for an item changes over time for a pre-determined and fixed sample size. This method allows items to be flagged soon after they have been compromised because it provides more precise information about when the *p*-value began to change. The underlying idea is that as items become compromised, they should become easier, and this will manifest in their *p*-values. A moving average is used to smooth out random fluctuations in the average caused by sampling error. The formulas for using the method of moving averages can be found in Appendix A.

The method of moving averages relies on the assumption that the distribution of examinee ability is stationary over time (Eckerly, 2017). Therefore, it might not be the most appropriate method for detecting item compromise for the WAPS testing program. When this assumption is not met, the moving averages method performs poorly and is too liberal in flagging items as compromised (Han & Hambleton, 2004). Because WAPS examinees are able to take an exam multiple times, it is logical to conclude that the distribution of examinee ability will change over time, because unsuccessful examinees (i.e., those low in ability) are the ones re-testing during subsequent administrations. On the other hand, the moving averages could be limited to first-time examinees.

Compared to moving averages of *p*-values, moving averages of item residuals or of standardized item residuals have been shown to perform better in these instances, except when items are particularly easy or at low levels of compromise. Similar to DGM and scale-purified DGM, we recommend that more research needs to be done with moving averages to identify what the distribution of examinee responses might look like in the absence of compromise. Because compromise can take a variety of forms (e.g., examinees sharing information about the exam content, scoring keys being stolen), it is not easy to specify what this variation in response patterns will look like. Therefore, simulations under a variety of conditions need to be conducted to better allow for the accurate detection of compromise and to avoid making improper conclusions.

There are also methods for detecting both individual examinees who may have benefited from item pre-knowledge and the specific items that were compromised. These approaches use a two-step process that entails (a) first screening the data for examinees who are likely to have benefited from item pre-knowledge and (b) then identifying the specific items that are likely to have been compromised. One well-known method involves a combination of Differential Person Functioning (DPF) and Differential Item Functioning (DIF; O'Leary & Smith, 2017). With DPF, examinee performance on two subsets of items (i.e., operational items and unscored experimental pretest items) is compared. Examinees who are aberrant based on their scores to the two subsets of items are flagged. For example, examinees who perform particularly well on the operational items but do not perform well on the pre-test items would be a cause for concern. This method assumes that candidates who had pre-knowledge of the items would likely have a higher score on operational items and a lower score on pre-test items, and that the operational items were the only items that were compromised.

After DPF is used to classify examinees into those likely to have had pre-knowledge and those not likely to have pre-knowledge, DIF analysis is conducted (O'Leary & Smith, 2017). The intent is to compare the item difficulty measures of those flagged as having pre-knowledge with those who were not flagged. Items that were not subject to compromise should be similar in difficulty for both groups of examinees, whereas items that were compromised should favor the group of pre-knowledge examinees. It is through these analyses that specific items in need of being replaced or retired are identified. The benefit of using this approach is that it can be easily customized to the Air Force. For example, with DPF, the thresholds used to indicate a meaningful difference in performance between operational and pre-test items can be established to meet the goals of WAPS. If the aim of the testing program is to enforce the Air Force policy directive regarding military testing, it would be advisable to set conservative flagging parameters such that only the most extreme cases would be flagged (e.g., examinees who have the largest discrepancies between their scores on the operational and pre-test items). Further, the Air Force's budget and resources for item development, ability to replace and retire items, number of test forms, and status of the item bank could all be considered when deciding the thresholds used for flagging suspect items. In the case of limited resources, multiple test forms, and/or a meager item bank, the AF could set the DIF thresholds higher so that these constraints are factored in, but item exposure concerns are still addressed.

Some organizations look for evidence of pre-knowledge not only through response patterns but also through response times (RTs). RT models are used to identify whether an examinee's RT for a test item differs significantly from what the model predicts given the examinee's overall performance and rate of responding. This can be used to pinpoint whether examinees had pre-knowledge of the items (Kim, Woo, & Dickison, 2017). Examinees who had pre-knowledge of items are likely to answer those items more quickly than they would have without the pre-knowledge (Eckerly, 2017).

RT models can be applied to both items and examinees. That is, they can be used to flag items for which a large number of examinees responded quickly or to flag people who have brief RTs for multiple items. Some examples of RT models are the effective response time (ERT) model (Meijer & Sotaridona, 2006), the hierarchical framework for response times and item responses (van der Linden, 2007), the lognormal RT model (van der Linden, 2006), and the hierarchical lognormal RT model (van der Linden & Guo, 2008).

The ERT model proves to be particularly promising, because it has been cross-validated with other methods used for detecting aberrance (Liu, Primoli, & Plackner, 2013). This approach involves estimating the ERT (i.e., the time an examinee needs to answer an item correctly) for each item so that the effects of item characteristics are removed. Given the definition of ERT, its equation includes only examinees who responded to the item correctly and whose probability of a correct response on the item is larger than a predetermined value. This effectively eliminates guessers when establishing the predicted ERT for each item. The difference between examinees' observed response times and predicted ERTs are then used to flag those who might have had item pre-knowledge. The formulas for the ERT model can be found in Appendix A.

Nevertheless, the usefulness of RT models for identifying test compromise remains unclear. When the same flagging criterion was used for the ERT model and the hierarchical framework,

the results were inconsistent (Kim, Woo, & Dickison, 2017). For ERT, the criterion was too liberal (i.e., a large proportion of cases were flagged); for the hierarchical model, it was too conservative. Both methods use chi-square tests that involve summing the squares of the residuals. This may make it difficult to determine aberrance in a subset of items. When the lognormal RT model (van der Linden, 2006) and the hierarchical lognormal RT model (van der Linden & Guo, 2008) were used to flag both individuals with pre-knowledge and items that had been compromised, these approaches also yielded results that were inconclusive. Both methods had only a modest number of flagged items and examinees in common with a licensure company's determination of the compromised items and examinees with pre-knowledge (Boughton, Smith, & Ren, 2017). Because more research is needed to determine the effectiveness of RT models in CBT, and because many things can contribute to aberrance in RT (e.g., poor time management during the exam), we recommend that RT analyses be used only as a complement to other detection approaches.

Finally, there are methods used to detect *groups* of examinees who may have benefited from pre-knowledge. With these approaches, examinees are grouped based on a common variable (e.g., test center, region, country). Two of the more promising methods include detection of collusion using cluster analysis (Wollack & Maynes, 2017) and the divergence algorithm, which stems from the Kullback-Leibler divergence statistic (Belov, 2017). The former does not require a grouping variable to be specified a priori, whereas the latter does. However, one of the issues with using cluster analysis to detect collusion is that the item parameter estimates are less accurate because the data set includes contaminated examinees (i.e., individuals who are part of the collusion group; Wollack & Maynes, 2017). Although this represents a realistic scenario, it inevitably affects the probability of answer matches and would require more examinees to share more answer matches in common before being flagged as atypical, resulting in less power. To overcome this issue, we would need data from secure test administrations (i.e., no cheating, no examinees have pre-knowledge) to estimate the item parameters. It is often the case that we cannot be entirely confident that the test data exclude data from examinees who colluded. Several methods have been suggested to overcome this limitation and increase power, including fitting the model using all examinees and then removing contaminated examinees and re-estimating the item parameters without them, and analyzing clusters in smaller batches or separately by location or by other grouping variables. However, these also bring with them their own drawbacks, including additional data work and limited ability to detect collusion across groups.

Yet, cluster analysis does not attempt to identify different subsets of compromised items for different groups of examinees. In reality, it is very likely that different groups of examinees have been exposed to different groups of compromised items. Thus, we believe it is prudent to consider a method that does address this scenario. The only method that does is the divergence algorithm introduced by Belov (2015). This approach has been able to identify compromised subsets of items with relatively high precision and performs fairly well at identifying aberrant examinees flagged by the test provider. Based off the Kullback-Leibler divergence statistic, it measures the similarity of the distributions of compromised and uncompromised items. However, a critical step in the algorithm involves performing a process called *simulated annealing*, which involves an iterative search to find the compromised subset of items by adding, swapping, and removing random items from a subset with the pool of eligible items to obtain the

optimal solution (Kirkpatrick, Gelatt, & Vecchi, 1983). This can be a rather tedious process, depending on the size of the operational test section and/or the size of the compromised subset. In the context of the WAPS, it is difficult to see the benefit in this approach for such a robust testing program with a sizable pool of operational items. The formulas and steps for the divergence algorithm and simulated annealing can be found in Appendix A.

3.4 Unusual Answer-Changing Behavior

Earlier, we discussed the use of RT models as a forensic detection method unique to CBT. Accordingly, it is only appropriate that we also discuss the detection methods that are specific to paper-and-pencil tests. In particular, erasure analysis is a forensic practice used to detect frequent erasures on a test. When an answer sheet contains many erasures, some patterns of erasures are consistent with cheating. Although some erasure patterns are due to cheating by an individual examinee during testing (Qualls, 2001), most test fraud involving erasures is performed post-exam by someone other than the examinee (Bishop & Egan, 2017). This is most prevalent in an academic setting, particularly for schools that historically have not met performance expectations. In these situations, a teacher or administrator may feel pressure to alter incorrect responses to improve the test scores of a class or school. However, erasure analyses could be useful outside of a school setting—particularly if there is some other reason to suspect that someone is changing answer sheets post-exam (e.g., test administrator).

Many organizations perform erasure analyses because the technology is widely available and relatively inexpensive (Bishop & Egan, 2017). Most organizations use an optical mark recognition (OMR) device or an image scanner with the OMR software. When an answer sheet is processed, information about the darkness of the mark and the amount of coverage inside response bubbles is captured for each item. These two pieces of information are then compared to a set of rules to decide which responses are considered legitimate and which are deemed erasures. However, there are issues with this technology. For example, erasures may be so light that they do not register as erasures, or stray marks on the answer sheet could be incorrectly reported as erasures. Despite these potential issues, optical scanners remain a cost-effective way of detecting erasures and could be employed in the WAPS testing program.

Several methods have been used for erasure analyses. Some organizations analyze erasures at the examinee level, whereas others aggregate the data at other levels (e.g., test center; Bishop & Egan, 2017). This information is then used to establish flagging procedures. A standard in the field is to flag cases that are four, five, or even eight standard deviations above the mean (Primoli, Liassou, & Bishop, 2011; Schiliro, 2010). Some methods like the *Z*-test for Population Means and Deviations from the Mean are based on a single dependent variable (e.g., the sum of wrong-to-right [WR] erasures, the ratio of WR erasures to total erasures [TE]), whereas other methods are regression-based approaches (e.g., student-level and group-level joint distributions) that take into account the conditional relations between WRs and TEs. A drawback of all of these methods is that they are single-level models, which means they can lead to aggregation bias and overlook relations that exist across nesting groups (Raudenbush & Bryk, 2002). We discuss a hierarchical approach later when discussing unusual gain scores.

However, both the *Z*-test for Population Means and the Deviations from the Mean approaches are appealing methods, because they indicate when a group's erasures are too high to be

attributed to random sampling alone (Bishop & Egan, 2017). This alignment with the basic principles of statistics allows them to be interpreted by stakeholders with ease. The primary difference between the two approaches is that the *Z*-test for Population Means uses examinee-level standard deviations to calculate the standard error, whereas the Deviations from the Mean approach uses standard deviations over the group means. Both methods have also been criticized because they assume that erasures follow a normal distribution.

For the purposes of the WAPS testing program, either of these approaches could be incorporated into the testing program with ease if erasure analyses are something the AF feels are worth conducting. Again, this is largely contingent upon whether the AF suspects there is a reason why erasures might occur post-exam and whether the WAPS testing program intends on continuing to use paper-and-pencil format in the future. As a crude flagging method, both these approaches are acceptable as long as it is understood that there are known issues with assuming normality of erasure distributions. The formulas for the *Z*-test for Population Means and Deviations from the Mean approaches can be found in Appendix A.

It is important to note that answer-changing behavior is not limited to paper-and-pencil tests, pertaining to CBT as well. This area of research is relatively sparse. There is some research that has explored wrong-to-right answer changes during CBT and the time spent on an item's screen when changing answers (Tiemann & Kingston, 2017). However, the research highlights the need for additional studies on what aberrant answer-changing behavior looks like. We can assume that a single wrong-to-right answer change made quickly might be suspicious (i.e., examinee did not spend much time re-considering a response) or that several wrong-to-right answer changes toward the end of an exam in a brief period of time are unusual. However, there could be alternative explanations for this test-taking behavior that do not reflect cheating.

3.5 Unusual Gain Scores

The final method we will discuss is detection of unusual score gains. This approach is closely associated with answer-changing behavior. If answers are changed from wrong to right, this will reflect a large gain in the score that may be improbable given historical performance. Gain score analyses are used to detect cheating by evaluating the differences between scores from two points in time. Groups who experience a large increase in their test score from Time 1 to Time 2 are flagged for review (Bishop & Egan, 2017).

One approach to gain score analyses involves using nonlinear regression (Clark, Skorupski, & Murphy, 2017). However, this method is best suited for when complete data are available for each examinee for Time 1 and Time 2. In real world settings, this is rarely the case. Without a complete data set, listwise deletion of cases that systematically relate to test compromise or high aptitude can occur. This would be particularly problematic to use for WAPS testing given that examinees who re-test for a promotional test (i.e., those who have data for both Time 1 and Time 2) would be the same individuals who were not promoted previously. Hence, data for those who are promoted would be systematically missing from analysis.

One method that seems relatively promising is the Bayesian Hierarchical Linear Model (BHLM) for detecting aberrant growth at the group level (Skorupski, Fitzpatrick, & Egan, 2017). This approach models examinee scores, nested within groups, over time. After considering group- and

time-level effects, groups are flagged based on aberrant growth in scores. One benefit of this method is that the examinees need not come from intact groups: the group-level information is incorporated only at Time 2. Examinees' scores at Time 1 are used to set a baseline of achievement, but these examinees do not need to be part of that same group at Time 2. This flexibility may be especially important for WAPS testing if it is expected that examinees may transition to different groups in the Air Force between test administrations. BHLM has performed fairly well as demonstrated by its ability to pinpoint known aberrant groups and to correctly exclude known groups of honest examinees. Although this method could be applied at the examinee level to standardize individual growth and evaluate aberrant growth for individual examinees (Skorupski et al., 2017), further research needs to explore its use for that purpose. The formulas for BHLM can be found in Appendix A.

The utility of BHLM is largely predicated on the extent to which known group membership is likely to be a cheating determinant in WAPS testing. We see this as potentially meaningful if there are specific groups within the Air Force for which large gain scores are observed. This could be the result of operational items being circulated to examinees throughout an AFSC and, thereby, contributing to a spike in test scores results. This could also be the result of instructors "teaching to the exam," but given these tests are for promotional purposes, it is unknown to us whether this is a plausible explanation.

3.6 Commercially Available Products

Several companies offer commercial products for detecting test compromise and fraud. However, because of the proprietary nature of these products, many companies understandably do not explicitly state which detection methods are being used. Although the exact statistical approaches are often not specified by these companies, some do provide insight into the quantitative methods used in manuals or other publicly available information (Assess Systems, 2016).

Caveon, Pearson, Questionmark, and Assess are a few of the companies that offer these commercial forensic products for statistical detection of cheating. Some are particularly vague in their description of products on their websites. For example, Caveon provides a brief description of their Secure Exam Interface (SEI) product that is intended to deliver exams protected against cheating and theft (Caveon, n.d.), but the mechanisms by which it achieves these goals are unclear based on the publicly available information. They also market their data forensic consulting services and claim that they can provide guidance on implementing a data forensics program, but they do not furnish much detail on their approach or methods for doing so. Pearson offers a product called IntelliVUE, which has reporting capabilities that allow organizations to understand and manage their testing program, including potential security issues (Pearson, n.d.). But the exact data they provide in the reports or the methods used are also unknown. Another vendor, Questionmark offers a product that has similar reporting capabilities. They provide more information, including screenshots of the different reports that are generated to identify potential cheating, flag content theft, and determine if allotted time to complete the assessment is sufficient (Questionmark, n.d.). Assess is one of the few companies that makes specific details of their product public (Assess Systems, 2016). They offer a test-compromise detection software application called Software for Investigating Test Fraud (SIFT). This software uses indices to detect answer copying, item pre-knowledge, and aberrant patterns in reaction time data. They

even go as far as to describe and provide the formulas used for the statistical indices in the product's online user manual.

Because of the limited information available, it is difficult to compare the vendors and their products. However, if the AF is leaning toward using a vendor for these forensic detection methods, it is might be particularly helpful to set-up demonstrations and/or speak with representatives from these companies so that the utility of the products for WAPS testing can be better evaluated. If developing a forensics program within the AF is desired, Caveon's consulting services may merit further consideration.

4.0 RECOMMENDATIONS

Ideally, recommendations should be based on a thorough test security audit. However, we can still make some tentative recommendations and suggest areas to investigate. Several suggested actions for any testing program have already been mentioned. Some specific recommendations for WAPS are highlighted below. Obviously, some recommendations might not be feasible at this time. Some other recommendations might already be implemented or be planned for the near future.

1. *Use a short testing window.* A 1 or 2-week window is common for professional certification programs. The shorter the window, the less chance that test content will be communicated to examinees in that window. Longer windows might be necessary, but the length should be no longer than needed.
2. *Administer the exam electronically*. It is much more difficult to control access to paper exams. It also allows you to collect item response time data which has various uses—including detecting test fraud and determining optimal testing time limits. Finally, CBT allows on-the-fly randomization of the item order.
3. *Create multiple test forms for each test administration*. The benefit of a candidate obtaining the test content on one test form is greatly reduced if he/she is administered a different test form. This might not be feasible for the SKT because each AFSC has its own test. If paper forms are used, answer-copying or pre-knowledge of the answer key could be reduced if two or more versions are created that contain the same items but are in different orders.
4. Consider banning cell phones from the testing room.
5. Formally train the test proctors.
6. Train item writers and other test contributors in test security. Require that they sign a non-disclosure agreement.
7. *Carefully control test materials during test development.* This includes strict controls over paper materials and encrypting of electronic materials.
8. *Consider using a web-based item-banking application for test development*. User permissions can be set up so that item authors, reviewers, and others so that contributors see only the item content they need to see.
9. *Communicate to examinees that sharing item content is forbidden*. Require them to sign a non-disclosure agreement. Clearly communicate the consequences for test fraud.
10. Consider converting the PFE to a computerized adaptive test (CAT).
11. *Compute some forensic statistics*. The appropriate statistics depend considerably on the whether the test is CBT or paper, whether item response latency data is obtained, the testing volume, resources, and the statistical expertise of the staff. It also depends on the forensic goals and the type of test fraud that is most likely. It might be best to hire a consultant such as Caveon to perform the forensic services or, at least, to help determine what types of forensics would be appropriate.

12. Analyze item parameter drift (e.g., changes in item difficulty over time) to help identify items that might be compromised.

5.0 REFERENCES

Allen, J. (2014). Relationships of examinee pair characteristics and item response similarity. In N. M. Kingston & A. K. Clark, *Test fraud: Statistical detection and methodology* (pp. 23–37). New York: Routledge.

Assess Systems. (2016). *User's manual for SIFT: Software for investigating fraud in testing*. Retrieved from http://www.assess.com/wp-content/uploads/2017/01/SIFT-1.0-Manual.pdf.

Belov, D. I. (2015). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, *40*, 83-97.

Belov, D. I. (2017). Identification of item preknowledge. In G. J. Cizek & J. A. Wollack, *Handbook of quantitative methods for detecting cheating on tests* (pp. 164–176). New York: Routledge.

Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback-Leibler divergence and K-index. *Applied Psychological Measurement*, *34*(6), 379-392.

Bishop, S., & Egan, K. (2017). Detecting erasures and unusual gains. In G. J. Cizek & J. A. Wollack, H*andbook of quantitative methods for detecting cheating on tests* (pp. 193–213). New York: Routledge.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

Boughton, K. A., Smith, J., & Ren, H. (2017). Using response time data to detect compromised items and/or people. In G. J. Cizek & J. A. Wollack, *Handbook of quantitative methods for detecting cheating on tests* (pp. 177–190). New York: Routledge.

Caveon. (n.d.). Secure exam interface. Retrieved from https://sei.caveon.com/

Chingos, M. M. (2012, November 29). Strength in numbers: State spending on K-12 assessment systems. https://www.brookings.edu/research/strength-in-numbers-state-spending-on-k-12-assessment-systems/

Cizek, G. J., & Wollack, J. A. (2017). Exploring cheating on tests. In G. J. Cizek & J. A. Wollack, Handbook of quantitative methods for detecting cheating on tests (pp. 3–19). New York: Routledge.

Clark, J. M., Skorupski, W., & Murphy, S. (2017). Using nonlinear regression. In G. J. Cizek & J. A. Wollack, Handbook of quantitative methods for detecting cheating on tests (pp. 245–261). New York: Routledge.

Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (2005). *Computer-based testing: Building the foundation for future assessments*. New York: Routledge.

Dimitrov, D. M., & Smith, R. M. (2006). Adjusted Rasch person-fit statistics. *Journal of Applied Measurement*, *7*(2), 170.

Eckerly, C. A. (2017). Detecting preknowledge and item compromise. In G. J. Cizek & J. A. Wollack, *Handbook of quantitative methods for detecting cheating on tests* (pp. 101–123). New York: Routledge.

Ferrara, S. (2017). A framework for policies and practices to improve test security programs: Prevention, detection, investigation, and resolution (PDIR). *Educational Measurement: Issues and Practice*, *36*(3), 5-23.

Ferrara, S., & Fremer, J. J. (2013). Security in large-scale paper-and-pencil testing. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 31-52). Routledge.

Foster, D., & Miller Jr, H. L. (2009). A new format for multiple-choice testing: Discrete-Option Multiple-Choice. Results from early studies. *Psychological Test and Assessment Modeling*, *51*(4), 355.

Gibson, W. M., & Weiner, J. A. (1998). Generating random parallel test forms using CTT in a computer-based environment. *Journal of Educational Measurement*, *35*(4), 297-310.

Green, J. P., & McCloy, R. A. (2018). *Best practices for testing policy: White paper* (2018 No. 056). Alexandria, VA: Human Resources Research Organization.

Guo, J., Tay, L., & Drasgow, F. (2009). Conspiracies and test compromise: An evaluation of the resistance of test systems to small-scale cheating. *International Journal of Testing*, *9*(4), 283-309.

Guttman, L. A. (1944). A basis for scaling qualitative data. *American Sociological Review*, *91*, 139–150.

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

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

Kim, D., Woo, A., & Dickison, P. (2017). Identifying and investigating aberrance. In G. J. Cizek & J. A. Wollack, *Handbook of quantitative methods for detecting cheating on tests* (pp. 70–97). New York: Routledge.

Kingston, N. M., & Clark, A. K. (2014). *Test fraud: Statistical detection and methodology*. New York: Routledge.

Kirkpatrick, S., Gelatt, C.D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671–680.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retesting effects in operational selection settings: Development and test of a framework. *Personnel Psychology*, *58*, 981–1007.

Liu, X. L., Primoli, V., & Plackner, C. (2013, October). *Utilization of response time in data forensics of K-12 computer-based assessment*. Paper presented at the Annual Conference on the Statistical Detection of Potential Test Fraud. Madison, WI.

Meijer, R. R., & Sotaridona, L. (2006). *Detection of advance item knowledge using response times in computer adaptive testing*. (LSAC research report series; No. CT 03-03). Newton, PA: Law School Admission Council

Meinert, D. (2015, June 1). What do personality tests really reveal? Retrieved from https://www.shrm.org/hr-today/news/hr-magazine/pages/0615-personality-tests.aspx

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, *9*(4), 287-304.

O'Leary, L. S., & Smith, R. W. (2017). Detecting candidate preknowledge and compromise content using differential person and item functioning. In G. J. Cizek & J. A. Wollack, *Handbook of quantitative methods for detecting cheating on tests* (pp. 151–163). New York: Routledge.

Olsen, J. F., & Fremer, J. J. (2013). *TILSA test security guidebook: Preventing, detecting and investigating test security irregularities*. Washington, DC: Council of Chief State School Officers.

Osburn, H. G. (1968). Item sampling for achievement testing. *Educational and Psychological Measurement*, *28*(1), 95-104.

Pearson. (n.d.). Reporting. Retrieved from https://home.pearsonvue.com/Test-Owner/Manage-your-program/Reporting.aspx

Primoli, V., Liassou, D., & Bishop, N. S. (2011, April). *Erasure descriptive statistics and covariates*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Qualls, A. L. (2001). Can knowledge of erasure behavior be used as an indicator of possible cheating? *Educational Measurement: Issues and Practice*, *20*(1), 9-16.

Questionmark. Reporting - item and assessment analysis. Retrieved from https://www.questionmark.com/content/reporting-item-and-assessment-analysis#section3

Randall, J. G., Villado, A. J., & Zimmer, C. U. (2016). Is retest bias biased? *Journal of Personnel Psychology*, *15*, 45-54.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd ed.)* Thousand Oaks, CA: Sage.

Schiliro, K. (2010, February 18). Questionable CRCT answer sheets at MCES? "Minimal concern" category by two-tenths of a percent, Morgan County Citizen. Retrieved from http://www.morgancountycitizen.com/?q=node/12946

Schleicher, D. J., Van Iddekinge, C. H., Morgeson, F. P., & Campion, M. A. (2010). If at first you don't succeed, try, try again: Understanding race, age, and gender differences in retesting score improvement. *Journal of Applied Psychology*, *95*(4), 603.

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden*, *7*(22), 131-145.

SIjtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's non-parametric IRT model. *Applied Psychological Measurement*, *16*(2), 149-157.

Skorupski, W. P., Fitzpatrick, J., & Egan, K. (2017). A Bayesian hierarchical model for detecting aberrant growth at the group level. In G. J. Cizek & J. A. Wollack, *Handbook of quantitative methods for detecting cheating on tests* (pp. 232–244). New York: Routledge.

Sotaridona, L. S., Wibowo, A., & Hendrawan, I. (2014). A parametric approach to detect a disproportionate number of identical item responses on a test. In N. M. Kingston & A. K. Clark, *Test fraud: Statistical detection and methodology* (pp. 38–52). New York: Routledge.

Sunbul, O., & Yormaz, S. (2018). The effect of augmented reality applications in learning process: A meta-analysis study. *Eurasian Journal of Educational Research*, *74*, 207-226.

Tiemann, G. C. & Kingston, N. M. (2017). An exploration of answer changing behavior. In N. M. Kingston & A. K. Clark, *Test fraud: Statistical detection and methodology* (pp. 158–172). New York: Routledge.

van der Linden, W. J. (2006). A lognormal model for response times on test items. Journal of *Educational and Behavioral Statistics*, *31*(2), 181–204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

van der Linden, W. J., & Guo, F. M. (2008). Bayesian procedures for identifying aberrant response time patterns in adaptive testing. *Psychometrika*, *73(*3), 365–384.

Van Iddekinge, C. H., Morgeson, F. P., Schleicher, D. J., & Campion, M. A. (2011). Can I retake it? Exploring subgroup differences and criterion-related validity in promotion retesting. *Journal of Applied Psychology, 96*(5)*,* 941.

Willing, S., Ostapczuk, M., & Musch, J. (2015). Do sequentially-presented answer options prevent the use of testwiseness cues on continuing medical education tests? *Advances in Health Sciences Education*, *20*(1), 247-263.

Wollack, J. A., & Fremer, J. J. (Eds.) (2013). *Handbook of test security*. New York: Routledge.

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

Wollack, J. A., & Maynes, D. D. (2017). Test collusion detection by clustering. In G. J. Cizek & J. A. Wollack, *Handbook of quantitative methods for detecting cheating on tests* (pp. 124–150). New York: Routledge.

Zopluoglu, C. (2016). Classification performance of answer-copying indices under different types of IRT models. *Applied Psychological Measurement*, *40*(8), 592-607. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5978724/

Zopluoglu, C. (2017). Similarity, answer copying, and aberrance. In G. J. Cizek & J. A. Wollack, *Handbook of quantitative methods for detecting cheating on tests* (pp. 25–46). New York: Routledge.

APPENDIX A: FORENSICS STATISTICAL FORMULAE

ω index:

$\widehat{w}_{ij} = \beta_0 + \beta_1 W_{ij}$

$\widehat{w}_{ij}$ = observed number of identical incorrect responses observed between two examinees

W_{ij} = observed number of items both the i^{th} and j^{th} examinees answered incorrectly

β_0 = regression intercept

β_1 = regression slope coefficient

$\widehat{R}_{ij} = \beta_0 + \beta_1 (R_i * R_j)$

$\widehat{R}_{ij}$ = observed number of identical correct responses observed between two examinees

R_i = observed number of correct responses for the i^{th} examinee

R_j = observed number of correct responses for the j^{th} examinee

β_0 = regression intercept

β_1 = regression slope coefficient

$E_{ij} = \sum_{k=1}^{K} P_{iko}$

E_{ij} = expected agreement between two examinees

P_{iko} = probability of selecting the o^{th} response alternative of the k^{th} item for the i^{th} examinee

$\sigma^2 = \sum_{k=1}^{K} P_{iko} * (1 - P_{iko})$

σ^2 = the variance of E_{ij}

$\omega = \frac{(w_{ij} + R_{ij}) - E_{ij}}{\sigma}$

H^T:

$H^T = \frac{\sum_{k \neq j} \beta_{jk} - \beta_j \beta_k}{\sum_{k \neq j} \beta_j (1 - \beta_k)}$

σ_{ij} = covariance between item scores of examinees *i* and *j*

β_j = the proportion of correct items for examinee j

β_k = the proportion of correct items for examinee k

β_{jk} = the proportion of correct items for both examinee j and examinee k

DGM and Scale-Purified DGM:

$$P(U_{ij}=1 \mid \Theta_{tj})^{1-T_j} * [(1-I_i) * P(U_{ij}=1 \mid \Theta_{tj}) + I_i * P(U_{ij}=1 \mid \Theta_{cj})]^{T_j}$$

$$P(U_{ij}=1 \mid \Theta_{tj}, b_i) = \frac{\exp(\Theta_{tj} - b_i)}{1+\exp(\Theta_{tj} - b_i)}$$

$$P(U_{ij}=1 \mid \Theta_{cj}, b_i) = \frac{\exp(\Theta_{cj} - b_i)}{1+\exp(\Theta_{cj} - b_i)}$$

$$\sum b_i = 0$$

T_j = 1 when $\Theta_{tj} < \Theta_{cj}$

Θ_{tj} = the jth examinee's true ability

Θ_{cj} = the jth examinee's cheating ability as estimated by examinee's performance on items specified as compromised

b_i = item difficulty for item i

I_i = the gating mechanism for item i such that I_i= 0 when assumed by the user to be secure and I_i = 1 when it is assumed to be compromised

Moving Averages:

Sequence of moving *p*-values is denoted as:

$(p_{100}, p_{101}, \ldots, p_{n-100})$

Where $p_{100} = \frac{1}{k}(u_{i,1} + u_{i,2} + \cdots + u_{i,100})$

$p_{101} = \frac{1}{k}(u_{i,2} + u_{i,3} + \cdots + u_{i,101})$

$p_{102} = \frac{1}{k}(u_{i,3} + u_{i,4} + \cdots + u_{i,102})$

$p_{n-100} = \frac{1}{k}(u_{i,n-99} + u_{i,n-98} + \cdots + u_{i,n})$

k = window size (100 in this example)

ERT Model:

$$\ln T_{ij} = \mu + \delta_i + \tau_j + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

μ = parameter for the mean of log response time for n items and N persons

δ_i = parameter for the response time required for item i

τ_j = parameter indicating the slowness of person j

$\epsilon_{ij} =$ residual assumed to follow a normal distribution with a mean of 0 and a standard deviation of σ

Slowness parameter of person j is calculated as:

$\tau_j \equiv E_i(ln\, T_{ij}) - E_{ij}(ln\, T_{ij})$

$E_i(ln\, T_{ij})$ = mean of person j's log response time over n items

$E_{ij}(ln\, T_{ij}) = \mu$

$\ln T_{ij} = \beta_0 + \beta_1 \Theta_j + \beta_2 \tau_j + \epsilon_j$

With $\epsilon_j \sim N(0, \sigma^2{}_i)$

$\beta_0, \beta_1, \beta_2$ = regression coefficients

ϵ_j = error term

Θ_j and τ_j = known regressors coming from the ability estimation process using known item parameters

$$z_{ip} = \frac{\ln T_{ip} - \widehat{\ln T_{ij}}}{\sigma_i}$$

Where $\sigma_i{}^2 = (N_i - 1)^{-1} \sum_j^{N_i} (ln\, T_{ip} - \widehat{\ln T_{ij}})^2$ is the variance of the log response time for item i

N_j = the number of persons selected for item i using the two criteria (i.e., person selected the correct answer to item i and a person whose probability of a correct response to item i is above a specified value

Divergence Algorithm:

$d_{j,S}$ = D * $(F_{T_j \setminus S} || F_{T_j \cap S}) \sum_{y \in Y} F_{T_j \setminus S}(y) \ln \frac{F_{T_j \setminus S}}{F_{T_j \cap S}}$

$E_\Omega[d_{j,S}]$ = expectation of the divergence statistic over examinees in the collection of random subsets

If for at least one examinee, $E_\Omega[d_{j,S}] > C(\alpha_1)$, then the group is affected. Then, each examinee from the affected group with $E_\Omega[d_{j,S}] > C(\alpha_2)$ is included in suspicious subgroup *J*.

$C(\alpha_1)$ and $C(\alpha_2)$ = critical values computed form simulated examinees drawn from *N*(0,1) population, where $\alpha_1 < \alpha_2$

$E_J[d_{j,S}]$ = expectation of $d_{j,S}$ over examinees from the suspicious subgroup J

S = subset

$j = j^{th}$ examinee

T_j = test administered to examinee j

Ω = collection of random subsets

$F_{T_j \setminus S}$ = posterior distribution of ability for examinees on a subset of items on a test administered to examinee *j*

$F_{T_j \cap S}$ = posterior distribution of ability for examinees on a subset of items that intersects with items on a test administered to examinee *j*

For each operational section of an exam (i.e., O_i, i = 1, 2,…,n), take the following steps:

1. Detect affected groups and suspicious subgroups using the statistic, $E_\Omega[d_{j,S}]$

2. For each affected group, detect the compromised subset (i.e., $S^* \subset O_i$) by finding the max $E_j[d_{j,S}]$, where S is the optimization variable and J is the suspicious subgroup detected.

3. For each affected group and compromised subset $S^* \subset O_i$, detect the aberrant subgroup using the statistic $d_{j,S*}$

Simulated Annealing:

1. Set the best solution $S^* = arg_{S \in \Omega} \max E_J[d_{j,S}]$, the current solution $S_0 = S^*$, and the temperature $t = t_0$.

2. Set subset S = S_0, then simulated random variable $\delta \in \{1, 2, 3\}$ according to the discrete distribution (7) and modify S, respectively.

3. If $E_J[d_{j,S}] > E_J[d_{j,S*}]$, indicating that an improvement to the best solution has been found, then set $S_0 = S$ and $S^* = S$ (update the best solution) and go to step 5; otherwise continue to step 4.

4. Simulate a uniformly distributed $\gamma \in [0,1]$. If $\gamma < \exp\left(\left[E_J[d_{j,S}] - E_J[d_{j,S_0}]\right]/t\right)$ (the probability of accepting a modification to the current solution, S_0, that did not improve the best solution, S *) then set $S_0 = S$ (update the current solution).

5. If $t > t_1$, then $t = t$ x d and go to Step 2 (perform more iterations to improve the best solution); otherwise stop (S^* is detected compromised subset).

***Z*-test for Population Means:**

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}/\sqrt{n}}$$

$\bar{X}$ = mean of the group

μ_0 = null population value

$\sigma_{\bar{X}}/\sqrt{n}$ = standard error

$\sigma_{\bar{X}}$ = standard deviation over all examinees

n = number of people in the group

Deviations from the Mean:

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}}$$

$\bar{X}$ = mean of the group

μ_0 = grand mean over all means for the groups in the testing program

$\sigma_{\bar{X}}$ = standard deviation calculated over the group means

Bayesian Hierarchical Linear Model:

Individual scores are simulated at Time 1 using a standard normal distribution:

$Y_{ig1} \sim N(0,1)$

i = individual

g = group

1 = time 1

Individual scores at Time 2 are simulated by introducing group-level growth conditional on scores at Time 1 and individual random error and a possible cheating effect for classrooms simulated to be aberrant.

$$Y_{ig2} = \mu_{g2} + \beta + \rho Y_{ig1} + \epsilon_{ig2}$$

μ_{g2} = mean increase of scores for examinees within group g at Time 2

β = cheating effect

P = correlation of scores between Time 1 and Time 2

ϵ_{ig2} = random individual error

Nested models:

$$Y_{igt} = \mu_{gt} + \epsilon_{igt}$$

$$\mu_{gt} = \mu_t + \tau_{gt}$$

Y_{igt} = the score of examinee i in group g at time t, with i = 1,…,$N(g)$

$N(g)$ is the number of individuals in group g

g = 1,…,300

t = 1, 2

Individual error variance-covariance structure is estimated separately for each group:

$\epsilon_{igt} \sim MVN(\underline{0}, \sum g)$, and

$\tau_{gt} \sim MVN(\underline{0}, \psi)$

For growth aberrance, calculate difference between effect sizes:

$$GA_g = \frac{\grave{\mu}_{g2} - \grave{\mu}_2}{\sqrt{\sigma_{g2}^2}} - \frac{\grave{\mu}_{g1} - \grave{\mu}_1}{\sqrt{\sigma_{g1}^2}}$$

μ_{gt} = mean of group g at time t

μ_t = marginal mean at time t

σ_{gt}^2 = variance of scores for group g at time t