AFRL-RH-WP-TR-2020-0028

NAVY PROMOTION TESTING RESEARCH: ITEM EXPOSURE ANALYSES

Christean Kubisiak
Michelle Kaplan
Mark Zorzie
PDRI, an SHL Company
1911 N. Fort Myer Drive
Suite 410
Arlington, VA 22209

March 2020
Final 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-2020-0028 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

CARRETTA.THOMAS.R.1228984929 Digitally signed by CARRETTA.THOMAS.R.1228984929 Date: 2020.06.02 09:21:11 -04'00'

THOMAS R. CARRETTA
Work Unit Manager
Collaborative Interfaces and Teaming Branch
Warfighter Interface Division
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

WEBB.TIMOTHY.S.1085032239 Digitally signed by WEBB.TIMOTHY.S.1085032239 Date: 2020.06.02 09:44:02 -04'00'

TIMOTHY S. WEBB, Ph.D., DR-IV
Chief, Collaborative Interfaces and Teaming Branch
Warfighter Interface Division
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

CARTER.LOUISE.ANN.1230249128 Digitally signed by CARTER.LOUISE.ANN.1230249128 Date: 2020.06.02 14:10:58 -04'00'

LOUISE A. CARTER, Ph.D., DR-IV
Chief, Warfighter Interface Division
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

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)*
14-03-20	Final	10-05-18 – 31-03-20

4. TITLE AND SUBTITLE

Navy Promotion Testing Research: Item Exposure Analyses

5a. CONTRACT NUMBER
FA8650-14-D-6500

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
Christean Kubisiak
Michelle Kaplan
Mark Zorzie

5d. PROJECT NUMBER
5329

5e. TASK NUMBER
07

5f. WORK UNIT NUMBER
H0SA

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
PDRI, an SHL Company
1911 N. Fort Myer Drive
Suite 410
Arlington, VA 22209

8. PERFORMING ORGANIZATION REPORT NUMBER

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

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

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR-2020-0028

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A. Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Report contains color; 88ABW-2020-1753; Cleared 13 May 2020

14. ABSTRACT

This report describes work conducted by PDRI to assist the Navy in studying item exposure and other exam development policies related to Navy-Wide Advancement Examinations (NWAE), one of several factors used in the Navy Enlisted Advancement System to determine advancement to ranks E-4 through E-7. We discuss current practices and recommendations regarding ways to ensure that item and exam development processes are fair, valid, and credible. The work consisted of three components: (1) A focused literature review on item exposure; (2) analyses of archival data to assess how item exposure across multiple administrations affected item performance; and (3) a description of best practices and recommendations on exam development policies concerning item exposure, random and randomized equivalent exams, parallel items, SJTs and other measures.This project underscores the need to carefully, consistently analyze item performance and ensure that trends in data do not change in unexplainable ways. If shifts are identified, tools are available to investigate what caused those changes, along with a variety of solutions. However, the application of appropriate tools is not always straightforward, and may have unanticipated secondary effects. Actions taken must be considered in the context of the overall program to ensure fair, valid and accurate testing.

15. SUBJECT TERMS
Military psychology; Industrial psychology; Selection; testing; enlisted promotion

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	165	19b. TELEPHONE NUMBER *(Include Area Code)*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENT

The authors would like to thank Dr. Thomas Carretta, Matthew Shull, Sue Dickerson, and Thomas Updike for their support throughout the project, and for their thoughtful, careful review of, and feedback on, the final report.

EXECUTIVE SUMMARY

This report describes a project to assist the US Navy (USN) in studying item exposure policies and other item and exam development policies as they relate to Navy-Wide Advancement Examinations (NWAE). Performance on the NWAE is one of several factors used in the Navy Enlisted Advancement System (NEAS) to determine advancement to the ranks of E-4 through E-7. The goal was to provide feedback on current practices and recommendations regarding ways to ensure that item and exam development processes are fair, valid, and credible.

The work done in support of this study consisted of three primary components. First, we conducted a focused literature review on topics related to item exposure. Additionally, we conducted analyses on archival data provided by the USN to assess how repeated item exposure across multiple administrations of testing affected subsequent item analyses and Sailor item performance. We also provided a description of best practices and recommendations on exam development policies concerning item exposure, random and randomized equivalent exams, parallel items, situational judgment tests (SJT), and other measures.

Many tools are available to detect possible negative effects of item exposure, but the key is to use multiple techniques over time. This is the only way to ensure consistent and reliable information that accurately identifies, and can be used to address, negative effects of item exposure. We also catalogued many methods of controlling for item exposure, such as modifications to tests and test banks, the addition of alternate forms, and changes to the testing environment. However, each of these methods have different limitations and benefits, so the selection of which to use must be carefully considered in the broader context of the overall testing process. Additionally, there are alternative methods of testing, such as SJTs, which can enhance and add incremental information to the testing process in a reliable, valid manner while mitigating the impact of previous exposure to test takers.

In our analyses of archival NWAE data, a consistent finding was that the proportion of items that became easier was relatively stable across different lengths of time between administrations, whereas the proportion of items that became harder increased. We also found that the percentage of items that became easier was consistently, significantly higher than the percentage of items that became harder. Further, the percentage of items that became easier decreased as the length of time between administrations increased. Finally, we found that among candidates who exhibit performance changes between repeat and non-repeat items, a significantly greater number performed better on repeat items. A variety of interpretations of these results are discussed. Though they provide some unique insight into the current testing program, in and of themselves, they do not provide clear evidence for specific revisions such as increasing the length of time between administrations.

Overall, the work compiled in this report underscores the need to be particularly careful about consistently analyzing item performance and ensuring that trends in the data do not change in unexplainable ways over time. If shifts are identified, tools are available to investigate what may be causing them, and several solutions can be implemented. However, the application of the appropriate tool is not always straightforward, and may have unanticipated secondary effects. Therefore, any actions taken must be considered in the context of the overall testing program to ensure fair, valid, and accurate testing.

1.0 INTRODUCTION

The purpose of this report is to describe work conducted to assist the US Navy in studying item exposure policies and other item and exam development policies as they relate to Navy-Wide Advancement Examinations (NWAE). Performance on the NWAE is one of several factors used in the Navy Enlisted Advancement System (NEAS) to determine advancement to the ranks of E-4 through E-7. The goal of the effort is to provide process feedback and recommendations to ensure that item and exam development processes are fair, valid, and credible.

The work done in support of this study consists of three primary components. First, we provide a literature review of the effects of item exposure, forensic tools used to detect overexposure of items, random vs. non-random equivalent exams, randomized vs. non-randomized equivalent exams, parallel item development, and SJTs. Next, we describe our analyses of how repeated item exposure across multiple administrations of testing and how it affects item analysis and Sailor item performance. Finally, we provide a description of best practices and recommendations on exam development policies concerning item exposure, random and randomized equivalent exams, parallel items, SJTs, and other measures.

2.0 LITERATURE REVIEW

In this section, we review academic and practitioner literature that describes item exposure policies and other item and exam development processes as they relate to the NWAE. We focus on item exposure in personnel selection and advancement exams as it impacts test validity, overexposure of items, and exam compromise. Further, we summarize the literature regarding various forensic tools used to detect overexposure of items as well as methods available for controlling overexposure of items and describe the feasibility and utility of these different tools and practices. Further, we review the literature on parallel item writing, discussing different methods of parallel item and test development across various examination situations, and describe the methods available for using parallel items and tests to control for item exposure. Last, we give a brief overview of the literature on SJTs including a discussion of the feasibility of implementing and applying SJTs to the NEAS process. Relevant literature was identified by searching databases of academic and practitioner literature, including Google Scholar, PsycInfo, and DTIC databases.

2.1 Item Exposure

Item exposure refers to the extent that an item has been used in past administrations of an examination. It is generally used as an indicator of the amount of risk there is that the item has been compromised for future administrations (Robin, 2005) as the possibility of examinees gaining specific knowledge about any item increases with each exposure. This is a key concern for test administrators because examinees gaining knowledge of items for future administrations due to item overexposure can lead to individuals' scores being compromised on their examinations. In order to avoid overexposing items, test administrators must establish a limit for the number of times they are willing to use an item and implement effective test development procedures that maintain exposure below this amount (Robin, 2005; Way, 1998). Test developers need to be mindful of item usage to prevent item overexposure and the issues it leads to, such as exam compromise and impacts on test validity.

When some examinees have access to a subset of the items used for an administered test prior to examination, the exam is considered to be compromised (Belov, 2014). When exam compromise exists, the validity of the exam is at risk due to individuals having preknowledge of the items they will be responding to (Belov, 2014), thus impacting the ability of examiners to accurately gauge examinee ability. An item can be considered compromised when there is evidence that it has become less difficult over time and it is reasonable to believe that this change is due to content being distributed beyond authorized usage or due to overexposure to examinees (Zara, 2006). Individuals can be considered compromised when there is impersonation, advanced information (preknowledge) about the content of the exam, or when there is evidence of test collusion (Belov, 2012; Wainer, 2014). Impersonation occurs when an examinee has another individual take the examination on their behalf. Preknowledge occurs when examinees have pre-existing knowledge of the answers to an item or items on an examination. Test collusion can include illegal coaching by a teacher or mentor, examinees accessing stolen test content on the internet, examinees communicating and informing one another of test answers during an exam, and test tampering such as changing answers after tests have been administered.

In order to identify level of item exposure, there should be consistent monitoring of item-level properties and examinees over test administrations as well as monitoring of examinee information networks (Robin, 2005). However, it is important to note that most studies looking

at issues of item exposure in repeat testing have not found a strong indication of such issues, even in high stakes environments such as medical exams and licensure examinations (Feinberg, Raymond, & Haist, 2015; Geving, Webb, & Davis, 2005; Giordano, Subhiyah, & Hess, 2005; Hertz & Chinn, 2003; O'Neill, Sun, Peabody, & Royal, 2015; Raymond, Kahraman, Swygert, & Balog, 2011; Raymond, Neustel, & Anderson, 2007; Wagner-Menghin, Preusche, & Schmidts, 2013; Wood, 2009). Even in the military, research has previously looked into the potential of exam compromise for the Armed Services Vocational Aptitude Battery (ASVAB). In comparing item and test difficulty across groups as well as test/retest scores, it was found that the ASVAB did not exhibit any signs of exam compromise (Alf & Stapleton, 1981; Guo, Tay, & Drasgow, 2009; McBride, 1997).

For example, Wagner and colleagues (2013) looked at the effects of reusing written test items on a newly introduced medical exam and found that there was no significant difference in perceived difficulty for examinees on new items compared to reused items. Moreover, students who took the test later in the examination process had lower scores than individuals who took the test earlier in the examination process (Wagner et al., 2013). Of courses, an alternative interpretation would be that weaker students may have delayed testing to facilitate additional preparation. Additionally, Feinberg and colleagues (2015) looked even more specifically at repeat effects of individuals who retake high-stakes examinations. However, when they looked at the data of repeat examinees on a credentialing exam, they found that examinees who answered incorrectly on an item in their initial examination attempt were likely to select the same incorrect response option 68% of the time on their second examination attempt (Feinberg et al., 2015). Thus, it is likely that repeat examinees are more likely to be misinformed rather than uninformed or compromised.

However, Qian, Staniewska, Rechase, and Woo (2016) looked at two high stakes examinations, one in the financial industry and another in the healthcare industry where they found some indications of potential item exposure issues. These two examinations utilized different techniques to administer their exams, which led to differing results between the two. Specifically, the financial industry examination used a traditional testing format with multiple-choice items while the healthcare industry examination used an adaptive testing format. For both of these examinations of potential item exposure, the adaptive test showed no indication of item exposure issues, while the other exam showed that two out of 111 items were potentially exposed with two out of 1,172 individuals indicating some amount of preknowledge on these items. Qian and colleagues (2016) cautioned against blind conclusions based on statistically significant results, as errors can easily occur and evidence from a statistical analysis alone is not enough evidence to invalidate an examinees test scores. As such, they recommended that when such results are found, a careful qualitative analysis should be performed to look at potential irregularities such as behavior during the given testing sessions or at the given testing center (Qian et al., 2016).

2.1.1 Forensic Tools

Several tools have been created to detect the various effects of item exposure. Across different testing mediums, detection methods can be grouped into several categories, identifying:

- Items that have been compromised
- Individuals who have preknowledge of compromised items
- Individuals who have been compromised and the items of which they have preknowledge

- Groups of individuals who are working together to expose themselves to test items (Eckerly, 2016a)

Methods for detecting data that has suspicious patterns of item exposure include checking for repeated response patterns, change in means and passing rate, response changes, and individual score increases and decreases (Kantrowitz & Gutierrez, 2013). These methods can be broken down into the following categories:

- Response Pattern Modeling
- Response-Time Modeling
- Speed/Ability Distributions
- Item Compromise Probabilities
- Utilizing Item Response Theory (IRT)

Two of the simplest methods for identifying compromised items include moving averages and the Log Odds Ratio Statistic (Eckerly, 2016a; Han, 2003; Han & Hamleton, 2007; McLeod, Lewis, & Thissen, 2003).

- *Moving averages*. With moving averages, examiners look at the proportion of correct responses for each item and then look at these proportions across different testing administrations to see if the average proportion of correct responses changes over time (Han, 2003; Han & Hambleton, 2007). With this method, we can infer that an item is compromised if the proportion of correct responses on that item drastically increases over time without any other apparent cause (Han & Hambleton, 2007).
- The Log Odds Ratio Statistic. Conversely, the Log Odds Ratio Statistic gives individual probabilities for each item being compromised. This statistic is based on Bayes' theorem, which postulates that the probability of something occurring can be determined through the use of knowledge and information about the conditions related to the thing in question. In this particular application, observed responses and other information about the prior conditions related to the item, such as prior usage and exposure rates, are collected. This information is used to calculate the probability that an item is compromised by looking at the likelihood that a response pattern differs from what is to be expected (McLeod et al., 2003).

The Log Odds Ratio Statistic, like many techniques, can be expanded to utilize IRT for greater accuracy. IRT is commonly used in test development and analysis to look at the relationship between individual examinees' performances on given items, tests, and test-takers. In IRT, there are different functions used to examine specific individuals taking the test and specific items on the test. To evaluate specific individuals, person-fit statistics are used to determine whether an individual's results on a test are valid. When looking at specific items, item-fit statistics focus on the item parameters that are estimated and use them to give information regarding two specific dimensions of the item: difficulty and discrimination. The difficulty parameter is used to estimate how difficult or easy the item is. The discrimination parameter estimates how much the correct answers on the item vary dependent on the person's knowledge of the concept that the item measures. There are several commonly used tools that utilize IRT, including:

- *Final Log Odds Ratios (FLOR).* One way of expanding the Log Odds Ratio Statistic with IRT is through FLOR, a series of seven log odds ratios (FLOR1-FLOR7) with varying degrees of constancy, difficulty, and frequency (McLeod et al., 2003). FLOR indices use

IRT parameters along with examinee response data to create probabilities showing the likelihood of an item being compromised (McLeod et al., 2003). This index uses Bayes' theorem similarly to the traditional Log Odds Ratio statistic, but takes it a step further by applying Bayesian calculations to person-fit statistics in IRT. That is, it looks at probabilities of items being compromised while person-fit statistics look at how individuals are performing across items in an exam. Thus, in FLOR, information gained from prior exposures and observed responses can be used to create the probability of an item being compromised. This is done differently across the seven indices that look at the probability of any given item being exposed to examinees. FLOR1 and FLOR2 are based on constant values for the probability that an item has been exposed. FLOR3 and FLOR4 use equations to derive this probability based on the difficulty of the item. FLOR5, FLOR6, and FLOR7 base this probability on empirical data that compares the relative item frequency from two, four, and eight sources respectively (McLeod et al., 2003). Using these indices, test makers can detect if examinees have some sort of item preknowledge/exposure. Because it utilizes Bayesian methods, which are adaptable and continually adjust to new/additional data provided, this probability can be continually updated as more information becomes available (such as through subsequent testing administrations).

- *Deterministic Gated Item Response Theory Model (DGM).* Another method of utilizing IRT to identify compromised individuals is the DGM, otherwise known as the Shu method (Shu, Henson, & Luecht, 2013). In traditional DGM, one of the parameters used in item-fit statistics, difficulty, is estimated prior to running the model. Then, a function is created to take an individual's performance on an exam and split it into two distributions: one that shows the individual's responses on normally appearing items and one that shows the individual's responses to items that have been compromised. Once these two distributions are created, they are then compared to see if there are differences in performance between the normal item distribution and the compromised item distribution in order to flag individuals that may have preknowledge (Shu et al., 2013).
- *Scale Purified Deterministic Gated Item Response Theory Model (Scale Purified DGM).* The DGM has also been modified to create the Scale Purified Deterministic Gated Item Response Theory Model which uses similar techniques, but with more precision, to more cleanly estimate which individuals have item preknowledge (Eckerly, 2016a; 2016b). In DGM, precision is conceptualized in the context of sensitivity and specificity. Sensitivity refers to the proportion of compromised individuals who are correctly detected as compromised, and specificity refers to the proportion of non-compromised individuals who are correctly detected as non-compromised (Shu et al., 2013). In Scale Purified DGM, the parameter estimates for difficulty are pre-set to predetermined, fixed values. Once the difficulty parameters have been set to these fixed values, then the DGM is run. This allows test-makers to look at the response data from a test group to find any potentially compromised individuals. These individuals are then removed, and the difficulty parameters are re-estimated using the cleaned response data (instead of the predetermined values used previously). This process of using fixed difficulty parameters and then re-estimating the parameters using the response data creates "purified" item difficulty estimates that can then be used to run through the DGM with the response data from all examinees. By "purifying" the parameter estimates for difficulty parsing out for

potentially compromised individuals, there is an overall greater amount of precision, with more sensitivity and more specificity than traditional DGM (Eckerly, 2016b).

Another technique used to determine whether individuals' responses are compromised is modeling their speed. The concept is that individuals who have preknowledge of specific items will answer those items more quickly than the items for which they do not have preknowledge (Boughton, Smith, & Ren, 2017; van der Linden, 2006; van der Linden & Guo, 2008). Numerous techniques utilize response-time modeling, including lognormal modeling of response times and hierarchical modeling (Boughton et al., 2017; van der Linden, 2006; van der Linden & Guo, 2008).

- *Lognormal Modeling*. In a lognormal model, response and response-time distributions are assumed to be determined by several different parameters (van der Linden, 2006). IRT is then used to set these different parameters, with a parameter for the speed of each person, another parameter for time intensity, and a third parameter for the discrimination of each item.
- *Hierarchical Modeling*. In hierarchical modeling, a lognormal model is combined with response pattern approaches (van der Linden & Guo, 2008). When doing this, the parameters are set according to the lognormal modeling approach, while also taking into account examinee ability.

Response-time techniques use Bayesian techniques to estimate the models, but are limited in the respect that they do not account for factors such as test-taker time management and other possible causes of speedy test-taking. In high-stakes testing, individuals may move quickly due to insufficient time so that they are able to answer every question before their allotted time runs out, while in low-stakes testing, individuals may move quickly because they are guessing randomly (Wang & Xu, 2015). As such, it is important to model accuracy along with speed in order to account for other possibilities. One example of this would be:

- *The Wang and Xu (WX) Model*. Wang and Xu (2015) suggest using a hierarchical model (the WX model) to examine the differences between behaviors on responses and response time to account for both the speed and accuracy of test-takers when attempting to detect individuals with preknowledge. In the WX model, an individual's responses are used to generate their proficiency distribution, which indicates their overall ability on the items. Then, a predictive distribution for the individual's responses on possibly compromised items is developed. Individuals with item preknowledge can subsequently be identified by comparing their observed responses on their proficiency distribution to the generated predictive distribution (Wang, Liu, & Hambleton, 2017).

Some of these techniques were developed specifically to detect individuals with item preknowledge, while others are traditional statistics techniques that have been modified for this purpose. O'Leary and Smith (2017) suggest using techniques within IRT to see if individuals or items are compromised:

- *Differential Person Functioning (DPF) with Differential Item Functioning (DIF)*. In IRT, tests can be conducted to look at differential person functioning (DPF) and differential item functioning (DIF). In O'Leary and Smith's (2017) approach, examinee performance on a subset of items that were only used once would be compared with examinee performance on a subset of items that were used multiple times. If individuals or items

perform differentially from the overall group, then there is a likelihood that the individuals and/or items are compromised (O'Leary & Smith, 2013; 2017). They recommend looking at DPF to see which individuals might be compromised, followed by sequential DIF to see the extent to which individuals' prior knowledge of item content affected item performance by comparing item difficulty and individuals' DPF results. In addition, Segall (2002) described an IRT method that can be used to detect the effects of test compromise when there are known secure items (items that do not have repeated exposure). This model evaluates the impact of test compromise on test scores through the estimation of two different test score distributions, one for the examinee's responses on secure items and another for the examinee's responses on non-secure items. These distributions are then compared to see if there are high score-gains on the non-secure items compared to the secure items to detect whether there has been test compromise.

There are fewer detection statistics that focus on identifying groups of individuals, due to the complexities involved. However, Zhang, Searcy, and Horn (2011) proposed a method of detecting group collusion by use of factor analysis:

- *Factor Analysis*. Factor analysis is a statistical technique that identifies latent factors (underlying patterns) that can be used to group items together. In their application, Zhang and colleagues (2011) suggested using this technique to identify underlying response patterns on items within a test. If they are found, they look to see if individual responses fit into those patterns using person-fit indices. If an individual's responses fit into a particular item response pattern that is aberrant, that individual may be compromised.
- *The l_z statistic*. The specific person-fit index used, which is commonly applied to item exposure detection is the l_z statistic (Drasgow, Levine, & Williams, 1985), which is a standardized version of statistic l (Levine & Rubin, 1979). The statistic l is a likelihood ratio index that measures aberrant responding based on a probability distribution of the ability in the population of examinees. Because this is a normal distribution, the probabilities of a high-ability individual incorrectly answering an easy item or a low-ability individual correctly answering a hard item are very low. Therefore, depending on an individual's ability on the overall distribution, the likelihood of them correctly or incorrectly answering items of varying levels of difficulty can be determined. If individuals answer items in a manner inconsistent with their ability, this suggests the individual is compromised.

There are also several statistics that specifically look to understand distributions of performance on items across individuals to search for anomalies that could indicate compromise:

- *Kullback-Leibler divergence (KLD)*. One method that looks at these distributions is the h statistic, or KLD (Kullback & Liebler, 1951; Sinharay, 2017). Similar to other methods, this statistic uses distributions of performance on items to see if individuals have knowledge of items prior to the exam. Specifically, the h statistic compares the distributions of speed or ability for an individual test-taker on compromised items versus non-compromised items (Kullback & Liebler, 1951). If there is a significant difference in performance between the two distributions, then it is likely that the individual has been compromised. Both the h statistic and the l_z statistic, in addition to being commonly used themselves as compromise detection tools, are the basis of many other compromise detection statistics.

- *L_s and R_s statistics*. A few statistics have been developed that combine principles used in the *h* statistic with principles used in IRT. Two techniques that use these principles are the L_s and R_s statistics which both compute ability estimates and compare performance on compromised and noncompromised items (Sinharay, 2017). They compute the maximum likelihood estimate for an individual's ability on a given test and then check to see if they outperform their given ability distribution (Sinharay, 2017). The primary difference between the two is how the ability distribution is computed. Otherwise, they both serve the same function as the *h* statistic. Among the three statistics, the *h* statistic is the least likely to falsely categorize someone as compromised when they are not (low Type I error rate), but is also the least likely to catch someone who is compromised (high Type II error rate, low power). All three of these techniques rely on test administrators first knowing which items have been (or could have been) compromised. However, there are several ways to detect potentially disclosed items, as discussed previously, including moving averages, time dependent IRT models (such as lognormal and hierarchical models models), procedures that detect DIF, and analysis of item fit statistics in IRT (Hatfield, 2007).

Table 1 contains a conceptual breakdown of the discussed methods, grouping them by type of statistical strategy utilized.

Table 1. Forensic Methods for Detecting Item Exposure

Strategies	Methods
Response Pattern Modeling	• Moving Averages • Factor Analysis
Response-Time Modeling	• Lognormal Modeling • Hierarchical Modeling • WX Model
Speed/Ability Distributions	• l_z statistic • *h* statistic/KLD • L_s statistic • R_s statistic
Item Compromise Probabilities	• Log Odds Ratio Statistic
Utilizing IRT	• FLOR • DGM/Shu Method • Scale Purified DGM) • DPF with DIF

2.1.2 Control Methods across Testing Formats

Examinations are most commonly given either using a paper-and-pencil format with written responses or through computer-based testing (CBT) with electronic responses. Previous research has shown that comparable scores are found across paper-and-pencil testing and CBT (Fisher, 2018; Mead & Drasgow, 1993; Paek, 2005), with only some minor effects attributed to the medium (Eignor, 1993; Paek, 2005; Pommerich, 2004). As such, many examinations across contexts are shifting to CBT (Folk & Smith, 2002; Moreno, 1997; Stocking, Smith, & Swanson, 2000; Wolfe, Moreno, & Segall, 1997). In paper-and-pencil testing programs, the primary

method used to control for the exposure of test questions is by developing parallel forms (Hetter & Sympson, 1997). Several alternative methods have been proposed to control for item compromise and individual preknowledge in CBT systems.

2.1.2.1 Traditional Testing Controls: Parallel Forms

In the more traditional cases of paper-and-pencil testing, item exposure is usually controlled through the use of alternate examination formats, otherwise referred to as parallel forms (Luecht, 2003). Generally, parallel forms refer to alternate versions of a test that can be considered equivalent to one another, wherein scores on one test can be translated and equated to scores on another test. Traditional tests administered by paper and pencil or non-adaptive computer-based testing generally attempt to control for the exposure of test questions by developing parallel forms (Hetter & Sympson, 1997; Stocking, 1993). Using different forms at the same time and then discarding them after a certain amount of uses allows examiners to limit the degree to which an item is exposed to various test-takers. Further, developing parallel tests allows test makers the ability to monitor exposure at the assembly stage by choosing which items are on which forms and how much each form is used (Ariel, Veldkamp, & Breithaupt, 2006). As such, parallel forms are one way of limiting item exposure and are commonly used in military testing (Hetter & Sympson, 1997; Oswald, Shaw, Farmer, 2015; Segall, Moreno, & Hetter, 1997).

Tests can have different types of parallelism, including:

- *Item-by-Item Parallelism* (Clause, Mullins, Nee, Pulakos, & Schmitt, 1998). In item-by-item parallelism, alternate forms of an exam are created by making them equitable at the item level (Clause et al., 1998). This form of parallelism is useful for test makers when individual items capture multiple dimensions simultaneously, since parallel items should capture all of the same dimensions as their parallel equivalent. If parallel items are created, then all dimensions captured in the original form of an examination will also be captured in the alternate form. However, this method of parallel test development is often highly complex and time-consuming to create. Thus, other options are often used.
- *Item-Set (Testlet) Parallelism* (Luecht, 2003; Ariel et al., 2006). One common method is to bundle items together into testlets to meet various statistical targets and categorical constraints (Luecht, 2003). Because similar items are clustered together, creating parallel testlets, the primary areas of interest to examiners will be ensuring equivalence in testlets across the parallel forms.
- *Exam Parallelism* (Mead & Meade, 2010). Exam parallelism is employed when neither of the previous methods are viable. With exam parallelism, the overall exam aims to capture the overarching dimensions of interest with similar degrees of accuracy to one another.

This can also be conceptualized by looking at the different degrees of parallelism that can be utilized when creating parallel forms, such as:

- *Strictly Parallel*. The highest degree of parallelism is found when tests are constructed to be strictly parallel before implementation. In strictly parallel tests, items are developed to behave identically, with the same means and variances between them. One problem with this is that even when designing tests in this way, many parallel tests have small statistical differences between their measurement properties (Wyse, 2011). However, research has shown that these small differences between measurement properties only

produces differences in accuracy (similarity between parallel items on different forms) less than 1.5% (Wyse, 2011). Unfortunately, given the complexity of making every item identical across tests, it is usually prohibitively difficult to design tests that are strictly parallel.

- *Weakly Parallel*. Another method of parallel test developments is constructing weakly parallel tests (any tests that are developed to be parallel while not being strictly parallel). Weakly parallel tests are easier to develop (Samejima, 1977) and are therefore much more commonly used. There are several ways to develop weakly parallel tests, considering test length and reliability. Among them are the minimization and maximization models (Sanders & Verschoor, 1998). In the minimization model, parallel tests are developed to have the shortest possible length, while in the maximization model, tests are developed to have the highest possible test reliability (Sanders & Verschoor, 1998).

We address different methods by which these types of parallel forms can be developed below.

2.1.2.1.1 Parallel Form Development

Parallel forms can be developed through several statistical approaches, the two most common being Classical Test Theory (CTT) and IRT. CTT is a measurement theory that uses a series of assumptions regarding the relationships between observed test scores and true test scores (Brown, 2013, Muñiz, 2005). The true test score refers to the individual's true ability on the overall topic of the examination in absence of any measurement error, while the observed test score refers to the score they actually receive. In CTT, true scores and observed scores are compared, looking at the factors that cause a difference in these scores– sources of error (Brown, 2013, Livingston, 1972). In parallel forms development, statistical analyses are conducted to allocate items to different forms that are similar in content and statistical qualities. Then these forms are pre-tested prior to being put into use and if observed scores means and variances are the same, then the forms are considered parallel.

CTT can be used for parallel test development and is easier than other commonly used methods as it allows for simple interpretation of examinee scores and item facility estimates (Armstrong, Jones, & Wu, 1992; Brown, 2013; Gibson & Weiner, 1998). This relative ease of use can be leveraged to automate test form construction. This is done by identifying item groups and randomly selecting a predetermined number of items from each group (Ackerman, 1989; Gibson & Weiner, 1998). Information is then derived about the key statistics for the preliminary test form by looking at which items discriminate the most between individuals (Armstrong, Jones, & Wang, 1994). Once this has been determined, test screening is used to develop parallel forms with the same key statistics such as score means and variances (Ackerman, 1989; Armstrong et al., 1994; Gibson & Weiner, 1998).

The primary focus when using CTT for parallel test development is optimizing test reliability (Armstrong et al., 1992), but it should be noted that CTT is a test-level approach, rather than an item-level approach (DeMars, 2018). As such, CTT can be used to develop parallel forms at both the testlet and full exam levels, but cannot be used at the item-by-item level. In absence of the ability to equate tests at the item-by-item level in CTT, the recommendation is to use CTT at the testlet level. Testlet parallelism allows test makers to cluster together similar items to ensure that there is equivalence across each category being tested (Luecht, 2003; Ariel et al., 2006). Conversely, with full exam parallelism, equivalence is only found at the full exam level and is

not recommended unless neither item-by-item nor testlet parallelism are viable options (Mead & Meade, 2010), even when employing CTT techniques. CTT methods are based on evaluating examinee test score and comparing overall observed test scores and overall true test scores (Brown, 2013; DeMars, 2018; Muñiz, 2005). IRT, conversely, looks at individual item functioning within the larger subset of items on a test. As such, when looking to take an item-level approach, IRT is preferred.

One of IRTs features is that it allows test developers to look at individual item information functions and thus see how each item contributes to the overall score of the individual. The item information functions are found by looking at the properties of items compared to the variance of the items. These functions can then be used to see how much information a given item is contributing. Because each item information function can be considered locally independent from one another, these functions are additive. Therefore, these individual item information functions can be summed together to create the test information function (TIF) that gives a general overview of the information regarding the exam. The TIF can be used to establish the maximum likelihood estimator of ability for any given individual across items (Ackerman, 1989; Lord, 1977), showing the most likely outcome for the individual across parallel forms of an exam. Tests are considered to be weakly parallel if their TIFs are identical (Boekkooi-Timminga, 1990). However, when using IRT to develop strictly parallel tests, different test characteristics are used (Boekkooi-Timminga, 1990).

Specifically, there is a function within IRT that allows test developers to look at the relationship between item response probabilities and underlying properties of the items such as their difficulty and discrimination. These item response functions (IRFs) can be used to examine properties of items. These IRFs can then be summed or averaged together to create an overall test response function (TRF), otherwise known as test characteristic curves (TCC). The TCC is most commonly used to develop strictly parallel forms, as tests are considered to be strictly parallel if they have both the same test length and the same TCC (Boekkooi-Timminga, 1990).

There are several methods that utilize TIF to create weakly parallel forms. One way to develop weakly parallel test forms is to use the heuristics, such as:

- *Minimization of Differences (DIFMIN).* In DIFMIN, items are assigned to a specific test one-by-one. Items are selected based on their individual item information functions. The goal with this heuristic is to reduce the largest difference in the information functions across all versions of the test (Ackerman, 1989; Adema, 1992).
- *Minimax (MI).* In MI, the primary objective is to minimize the difference between the TIF for the original test form and the TIF of the parallel form being built. This is done to optimize the similarity of information gathered from either test form (Adema, 1992). However, with each successive test developed, the TIF for the newest form will be increasingly different from the original test form, causing a decrease in statistical quality for each successive form.
- *Minimax with Minimization of Differences (MIDI)*. One way to overcome some of the issues present in both MI and DIFMIN is to combine them. In the MIDI method, the DIFMIN heuristic is applied to the minimax method. Because of the heuristic's item-by-item approach, it helps to increase the statistical quality of successive tests built (thus preventing the problems that arise when solely using minimax). In the MIDI method, content areas that need to be represented are pre-specified prior to test development along

with the TIF. Then, items are selected onto a test one-by-one, with the item that has the highest information value selected first. If the specific content area of a given item is already represented on the test, then the next best item is selected. This process is repeated until an item is found for a content area that has not reached its pre-specified number of items (Adema, 1992).

- *Maximin (MA)*. MA looks to maximize the amount of information for every given item. This is done by looking at the statistical value of each item and discerning how much information would be lost if a given item were not included. Then, items are selected based on how great their loss is. Items whose maximum loss is greater than the minimum loss of all other items within a content area are the first to be selected (Adema, 1992).
- *Maximin with Minimization of Differences (MADI)*. Similar to the way in which DIFMIN is applied to MI to create the MIDI method, DIFMIN can also be applied to MA to overcome some of its issues. The MADI method selects items following the same basic principles as MIDI, with items chosen within pre-specified content areas. The key difference between the MIDI and MADI methods are the manner in which items are selected. While MIDI follows the MI format of optimizing information values, MADI follows the MA method of minimizing information loss in the MADI method (Adema, 1992).

Overall, IRT and CTT have comparable outcomes statistically and conceptually across various types of tests, and both have been widely used in the military (Fan, 1998; Hwang, 2002; Lin, 2008; Mead & Meade, 2010; Oswald, Shaw, & Farmer, 2015; Stage, 2003). However, IRT performs slightly better when a method using a predetermined TIF is used (Mead & Meade, 2010). There are several criteria for choosing between the two statistical methods for test development. CTT should be used when sample sizes are small and data has multiple dimensions (Zickar & Broadfoot, 2009), because developing parallel forms at the test level can help capture across numerous dimensions more easily. Furthermore, because IRT is a more complex method, CTT is preferable with small samples due to its simplicity. IRT, however, should be used when the test is focused on specific constructs of interest to the examiner (Zickar & Broadfoot, 2009), since it allows test developers to focus on each construct and area of interest. Segall et al. (1997) recommend developing parallel forms with a large range of item difficulties since this eases the ability of examiners to capture differences in individual ability, functionally independent items that assess different characteristics of the examinees, unidimensionality of items so that it is easier to create parallel equivalents, supplemental item banks so that more items are available as needed, and regular item reviews to best utilize them.

2.1.2.1.2 Additional Approaches to Parallel Forms

Another approach to parallel tests involves not requiring them to be developed as parallel before administration. Rather, the tests are equated after they are administered (Algina & Penfield, 2009; Kolen & Tong, 2005). However, while scores on tests designed to be parallel can be equated, the process is much more difficult with tests that were not originally designed to be parallel. The three most commonly used methods of test equating are random groups, single group design, and common-item non-equivalent groups (Kolen & Tong, 2005). In a random groups design, different test forms are randomly assigned to examinees and compared. In a single group design, the same examinees are given both forms of the test and then scores across both exams are compared and equated. In common-item non-equivalent groups, different test forms are given to separate (non-random) groups, but the test forms have some common items on

each exam. Then, comparison of the common and unique items is done to see if there is a difference between individuals' performance on items that are common to both tests and items that are unique to their specific test. Then, if performance is different on unique items compared to common items, adjustments in scoring can be made for the unique items of the given tests.

With statistical equating, there are two primary methods: linear equating and equipercentile equating (Kolen & Tong, 2005).

- *Linear Equating*. In linear equating, the data gathered from both forms are transformed so that both have the same mean and standard deviation. In equipercentile equating, scores on alternative forms are made to have the same distribution across a population of examinees. These statistical methods, however, can be very susceptible to sampling error as individuals who take the different forms of the examination may not be representative of the overall population.
- *Equipercentile Equating*. Traditionally, when using equipercentile equating, smoothing methods are used to reduce the sampling error (Kolen & Tong, 2005). Smoothing generally refers to statistical processes that take a dataset and capture patterns and other key elements within the data by approximating functions that reduce error and extract the most information possible from the data. These methods can be used to ease equipercentile equating in two ways. The first is through presmoothing methods, which smooth the overall score distribution (Kolen & Tong, 2005). The second is through postsmoothing methods, which smooth the specific equipercentile function (Kolen & Tong, 2005). Equipercentile equating and linear equating are done to make the forms easier to compare to one another, but they are susceptible to a few types of errors in equating.

Translating scores on one test to a different test can result in errors of bias for several reasons (Kolen & Tong, 2005). Random errors can occur with any statistical analysis due to unpredictable fluctuations in individuals and testing tools. With equating forms, however, this issue can often be overcome by using a technique called bootstrapping (Angoff, 1957; Kolen & Tong, 2005). Bootstrapping takes a given dataset and continually resamples data from it, thus reducing the effects that would otherwise be found from random sampling errors. Systematic errors can occur when there is a common error occurring for all individuals in one case, but not another. In parallel forms, this would occur when there are systemic differences between results on one form of an exam compared to another. When these errors occur, it is an indication of some issue with the development of the two forms failing to make them equitable. Other types of systematic errors that can occur in parallel forms after development when equating them. Systematic equating errors and errors of interpretation can occur when there is a violation of assumptions made in the equating method used. For example, if using an equating method that requires regression, then regression-based assumptions must be satisfied to avoid systematic equating errors (Kolen & Tong, 2005). Equating must be done following established statistical protocols to avoid these errors (Angoff, 1957; Kolen & Tong, 2005).

2.1.2.2 CBT Controls: Multi-Stage Testing

The most common method for controlling item exposure when utilizing CBT is changing the format of the examination from a standard fixed exam to some form of multi-stage testing (MST) (Davis & Dodd, 2003; Hetter & Sympson, 1997; Davey & Parshall, 1995). See Figure 1. MST interactively selects item sets for test-takers dependent on their individual ability. Traditionally

in MST, all test-takers are first given a set of items of intermediate difficulty. If they do well on this first item set, they are given a more difficult item set. If they do poorly on the first item set, they are given an easier item set. If they do neither well nor poorly, then they are given an intermediate item set. This process then continues and item sets are selected based on individual ability, building out the overall test in stages for the specific test-taker.

Thus, in MST, with the exception of the first item set, individuals will be exposed to different items dependent on their ability. If they are high-ability, they will see more difficult item sets; if they are intermediate-ability, they will see more intermediate item sets; and if they are low-ability, they will see easier item sets. The total number of item sets viewed is generally fixed and depends on the examination. While adding more item sets allows examiners to have more flexibility, it also increases the complexity of test assembly (Yan, Lewis, & von Davier, 2016).

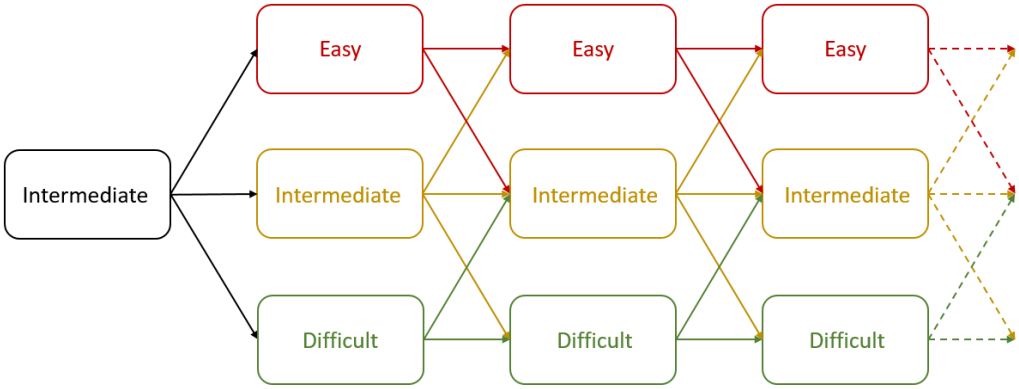

Figure 1. Multi-Stage Testing Design

As such, it is the decision of the test developer to determine what the appropriate number of item sets for a given examination will be depending on how flexible they want the test to be and how much complexity they are able to engage in.

Multi-stage testing can mitigate problems with item exposure because it limits the number of individuals who are exposed to various items, among other benefits. A specific type of examination based on MST that is commonly used to limit item exposure issues is computerized adaptive testing (CAT). Similar to MST, the test is tailored to an individual's ability level, presenting different items to different individuals. However, with CAT, items are selected one at a time instead of in sets. Numerous studies have looked at CAT compared with traditional testing environments and have found it has the same level of performance as traditional tests (Eignor, 1993; Haley, Coster, Andres, Kosinski, & Ni; 2004; Mead & Drasgow, 1993; Stocking et al., 2000; Segall, Moreno, Kieckhaefer, Vicino, & McBride; 1997). Due to its adaptive nature, CAT is able to accomplish this while administering fewer items (Ayhan, 2015; Moreno, Wetzel, McBride, & Weiss, 1984).

Further, the utility of CAT in controlling for item exposure has been shown in the military context. For example, the ASVAB was converted from a traditional paper-and-pencil format to a CAT format. This conversion was conducted using score equating development (SED) and score equating verification (SEV) to make sure that individuals who took either the traditional format or the CAT were evaluated equally (Moreno, 1997; Segall, 1997). Research on this conversion has shown that there is little significant difference in performance between the traditional test

format and the CAT format (Moreno et al., 1984; Segall et al., 1997; Wolfe et al., 1997). However, CAT provides a slightly more precise measurement of aptitudes over the previous paper and pencil method and traditional non-adaptive CBT (Divgi & Mayberry, 1991; McBride, 1997).

Similarly, the military has used CAT in the Tailored Adaptive Personality Assessment System (TAPAS). This assessment was created to measure twelve personality facets that were relevant to attrition and training performance in the Army (Drasgow, Stark, Chernyshenko, Nye, Hulin, & White, 2012; Nye, Drasgow, Chernyshenko, Stark, Kubisiak, White, & Jose, 2010). Research done by Personnel Decisions Research Institutes (PDRI) and others has shown that the TAPAS is a valid measure of the non-cognitive characteristics involved in selection of new soldiers (Drasgow et al., 2012; Nye et al., 2010). The ASVAB and TAPAS are often used in conjunction with one another to measure the cognitive and non-cognitive abilities of new soldiers (Drasgow et al., 2012; Nye et al., 2010).

2.1.2.2.1 Forensic Tools in CAT

Although CAT is a recommended solution for item overexposure, it is not a complete solution. Though only a few of the methods that exist to test for compromised individuals in traditional testing use response time, the majority of the techniques used specifically for CATs use speed as a factor in identifying compromised individuals, due to the computer-based format and the ease with which response time and response patterns can be measured (Choe, Zhang, & Chang, 2017; Lee, 2018). These methods include:

- The loglinear response time model
- Effective response time model (ERT)
- Lognormal response time model (Ln-RT)
- Joint model within a hierarchical framework (H-IRTRT)
- The mixture Rasch model (MRM)
- The mixture lognormal model of response times (MRM-RT)
- The WX model (Choe et al., 2017; Lee, 2018)

According to Lee (2018), in the lognormal response time model, observed response times for each item for each person can be averaged across items and people to look for item and person effects. Aberrant responding is identified by looking at the amount of time elapsed for an individual on an item. In ERT, the comparison is done between observed and predicted response time (Lee, 2018). In the Ln-RT, response times and item responses (whether they are correct or incorrect) are incorporated with one another to determine item and person specific parameters which are then incorporated into an IRT framework (Lee, 2018). The H-IRTRT is a two-level model which looks first at IRT and response time models and then compares the distribution between an individual's ability and speed (Lee, 2018). The MRM is a predecessor of the DGM and looks at item responses through the use of Rasch modeling (Karabatsos, 2003; Shu et al., 2013).

Rasch models operate on similar principles to IRT with a few key theoretical differences. While IRT approaches focus primarily on fitting a model to the data, this is only a secondary requirement for Rasch modeling. Rasch modeling approaches instead focuses primarily on meeting three measurement requirements. The measurement of any item is unrelated to any

person factors and the measurement of any person is unrelated to any item factors. The measurement of items and persons should function according to conventional rules of arithmetic, and any combinations of measurements should be plausible combinations of items and persons (Andrich, 2004; Rasch, 1960; Smith, 1990).

The Rasch model is often considered to be the One-Parameter Logistic (1PL) IRT model. However, proponents of Rasch modeling prefer to view it as a completely different approach to conceptualizing the relationship between data and theory. Like other statistical modeling approaches, IRT emphasizes the primacy of the fit of a model to observed data, while the Rasch model emphasizes the primacy of the requirements for fundamental measurement, with adequate data-model fit being an important but secondary requirement to be met before a test or research instrument can be claimed to measure a trait. Operationally, this means that the IRT approaches include additional model parameters to reflect the patterns observed in the data (e.g., allowing items to vary in their correlation with the latent trait), whereas in the Rasch approach, claims regarding the presence of a latent trait can only be considered valid when both (a) the data fit the Rasch model, and (b) test items and examinees conform to the model. Therefore, under Rasch models, misfitting responses require diagnosis of the reason for the misfit, and may be excluded from the data set if one can explain substantively why they do not address the latent trait. Thus, the Rasch approach can be seen to be a confirmatory approach, as opposed to exploratory approaches that attempt to model the observed data.

The MRM-RT expands on the MRM to classify examinees' response times and observed responses into either solution-finding behaviors or rapid guessing behaviors (Lee, 2018). Finally, the WX model operates in the context of CAT in the same way that it does in the context of paper-and-pencil tests and traditional CBT (WX, 2015).

Table 2 contains a conceptual breakdown of the discussed methods, grouping them by type of statistical strategy used and application to traditional testing vs. CAT.

Table 2. Comparison of Forensic Methods for Detecting Item Exposure in Traditional Testing vs. CAT

Strategies	Methods for Traditional Testing	Methods for CAT	Methods for both Traditional Testing and CAT
Response Pattern Modeling	• Factor Analysis		• Moving Averages
Response-Time Modeling	• Hierarchical Modeling	• Loglinear Response Time Model • ERT • MRM • MRM-RT	• Ln-RT • H-IRTRT • WX Model
Speed/Ability Distributions			• l_z statistic • *h* statistic/KLD • L_s statistic • R_s statistic
Item Compromise Probabilities			• Log Odds Ratio Statistic
IRT	• DPF with DIF		• FLOR • DGM/Shu Method • Scale Purified Deterministic Gated Item Response Theory Model (Scale Purified DGM)

2.1.2.2.2 Control Methods in CAT

There are two primary strategies for controlling for item exposure in CAT: randomization and conditional selection (Chang & Twu, 1998; Georgiadou, Triantafillou, & Economides, 2007). The former adds a random component based on some criteria (dependent on the method) to the pre-specified item selection method. One common criterion is to look at how informative an item is, or how much information can be gained about the individuals' abilities. For example, a common randomization strategy is 5-4-3-2-1, where the first item is chosen randomly from the five most informative items in the pool, the second item is chosen randomly from the next four most informative, the third item is chosen randomly from the next three most informative, the fourth item is chosen randomly from the next two most informative, after which the random component ends and all subsequent items are chosen based on how informative they individually are. Most randomization strategies utilize components of IRT to determine which items are most informative and randomly select items for any given examinee based on those principles (Davis, 2002, 2004; Davis & Dodd, 2005; Revuelta & Ponsoda, 1998; Stocking, 1993).

Conditional selection strategies involve assigning a parameter based on some criteria for each item to control the number of times that the item can be used (Georgiadou et al., 2007). They are primarily based on the Sympson and Hetter (SH) method, which was one of the first attempts to

address item exposure issues in CAT (Chang, Ansley, & Lin, 2000). In the SH method, an exposure control parameter is assigned to each item based on a series of statistical simulations. All of the other conditional selection methods, such as the Davey and Parshall (DP) method, the Stocking and Lewis Multinomial (SL) method, and Targeted Exposure Control (TEC) are based on the SH method, with some modification to either parameter determination or item selection (Chang, et al., 2000; Chen, 2010; Chen, Ankenmann, & Spray, 1999; Chen & Lei, 2005; Davey & Parshall, 1995; Pastor, Dodd, & Chang; 2002; Stocking, 1993; Stocking & Lewis, 1995a, 1995b, 1998).

Building on these, there are also stratified strategies that incorporate more complex rules by which items are selected. Specifically, items are subdivided into item groups, or strata, based on how similar they are to one another (either in content or difficulty). Then, items from each strata are chosen either randomly or based on some parameter for different segments of the test (Barrada, Olea, Ponsada, Abad, Ponsoda, & Abad, 2009; Chang & Ying, 1999; Parshall, Harmes, & Kromrey, 2000; Yi & Chang, 2003). Further, several methods combine two or more methods from different strategies together (Georgiadou et al., 2007). These can include strategies that combine stratified approaches with conditional selection approaches (Chang & Ying, 1999; Yi, 2002) or combining randomization approaches with conditional selection approaches (Barrada et al., 2009; Eggen, 2001) or other similar combinations (see Table 3 for the full list of methods).

Note that research has shown that the use of the majority of these exposure control methods does not impact the precision of the examinations, or their ability to predict performance (Hetter & Sympson, 1997). . They simply reduce the extent to which items are exposed to test-takers. There are some general trends regarding which methods are the best for reducing item exposure. Generally, the newer techniques outperform the older ones. For example, in conditional selection methods, DP and CSH both outperform SH (Parshall, Davey, & Nering, 1998). Randomization techniques generally perform at similar levels to conditional selection techniques, and are easier to implement (Davis & Dodd, 2005). Although stratified selection methods can reduce both item exposure and item overlap while increasing overall item pool utilization, conditional selection procedures such as SH and CSH control for item exposure more (Pastor et al., 2002). However, these same conditional selection procedures sacrifice a small degree of measurement precision, unlike the majority of other exposure control methods (Pastor et al., 2002).

Table 3. Control Methods for Minimizing Item Exposure in CAT

Strategies	Methods
Randomization	• 5-4-3-2-1 (McBride & Martin) • Randomesque (Kingsbury-Zara) • INFO4 procedure • Within .10 logits • Progressive strategy (Revuelta & Ponsada)
Conditional Selection	• Sympson & Hetter (SH) • Extended SH • Davey & Parshall (DP) • Stocking & Lewis Multinomial • Restricted Maximum Information Strategy • SH Conditional procedure (CSH)

Strategies	Methods
	• Stocking & Lewis Conditioning on Estimated Ability • TEC • Chen & Lei
Stratified Strategies	• a-str • a-str with freezing • a-str with b-blocking • a-str with unequal item exposure across strata • a-str design with content blocking (STR-C) • Multidimensional stratification • Stratification strategy
Combined Strategies	• Progressive restricted strategy • Nering, Davey, & Thompson Hybrid strategy • Eggen's strategy • Communication of the a-str with the SH strategy (STR-SH) • Incorporation of the SH into a-str with content blocking (STR-C-SH) • Content constrains in a-str CAT using a shadow test
Alternate Designs	• Computerized Adaptive Sequential Testing (CAST) • Adaptive multi-stage item bundles • Multiple forms structures (i.e. parallel test)

2.2 Situational Judgment Tests (SJT)

SJTs are an assessment method where examinees are presented with a variety of work related situations and a series of possible responses to those situations (Campion, Ployhart, & MacKenzie, 2014; Motowidlo, Dunnette, & Carter, 1990; Weekley & Ployhart, 2013). They measure job-related skills and abilities by asking individuals to indicate how each incident should be handled (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006). In an SJT, the job-related incident presented is generally a realistic, challenging work situation that the individual may be confronted with on the job. Examinees choose how to respond based on their problem solving skills and interaction styles. Responses may involve selecting the best option, worst option, or rating the effectiveness of each, depending on how the measure is constructed.

Because SJTs involve realistic situations that can reflect a variety of factors, the test content in SJTs are often multi-faceted. This is because the situations reflect a combination of many different constructs and competencies and measures multiple areas of knowledge, skill, and ability simultaneously (Campion et al., 2014). While this makes it more difficult for an SJT item to uniquely capture a given, distinct construct being measured, the multidimensional nature of the items makes them more predictive. This is due to their ability to incorporate numerous factors contributing to individuals' future job performance (Bergman et al., 2006; Guenole, Chernyshenko, & Weekly, 2017; Schmitt & Chan, 2006). However, this can also lead to issues with construct validity. Construct validity is an indicator of the relationship between what an assessment intends to measure theoretically and what it measures in practice (Carmines & Zeller, 1979; SIOP, 2003). With SJTs, construct-related validity can be problematic (McDaniel et al., 2003; Ployhart & Ehrhart, 2003). Because construct validity focuses on clarifying each individual dimension and how well it is being measured, the multidimensionality of SJTs makes this more difficult. This can have profound ramifications for promotion testing and evaluation of

candidates because when using SJTs, it becomes unclear which specific constructs are being assessed. Thus, when candidates are being evaluated, there is a lack of clarity regarding what specific aspects of the multidimensional test they did well in versus what specific aspects they did poorly in, leading to uncertainties regarding their overall performance on any given dimension of interest.

Many studies have supported SJTs as a valid predictor of overall future job performance (Campion et al., 2014; Chan & Schmitt, 2002; Clevenger, Pereira, Weichmann, Schmitt, & Harvey, 2001; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; Motowidlo et al., 1990; Weekley & Jones, 1999). They are also effective predictors of individual task performance, and context-dependent task performance (Chan & Schmitt, 2002; McDaniel et al., 2001; Motowidlo et al., 1990). Perhaps more importantly, in addition to their high predictive validity, SJTs also have high incremental validity over other commonly used predictors of job performance, such as cognitive ability tests, personality tests, and job experience (Campion et al., 2013; Chan & Schmitt, 2002; Clevenger et al., 2001; Weekley & Jones, 1999).

An issue with some commonly used selection tools such as cognitive ability tests and job knowledge tests is that they have been found to assess individuals from some subgroups differentially, unintentionally selecting individuals in one group at a higher rate than other groups (Bobko, Roth, & Buster, 2007; Murphy, 2002; Pulakos, 2005). For example, cognitive ability tests have been shown to differentially select for different ethnic groups such that white individuals are generally selected at higher rates than members of other ethnic categories when using this selection tool (Bobko et al., 2007). However, SJTs typically exhibit fewer subgroup differences (Campion et al., 2014; Clevenger et al., 2001), and more often select individuals across different groups at more similar rates. This decrease in differential selection is likely due to systematic variance in the non-cognitive components of SJTs (McDaniel & Nguyen, 2001; Weekley & Jones, 1999). Note that they exhibit more differences between individuals of various subgroups than personality measures (Clevenger et al., 2001), which is likely due to the extent that they capture general cognitive ability (McDaniel & Nguyen, 2001).

2.2.1 Developing and Implementing SJTs

SJTs have several characteristics that organizations must consider prior to implementing them. These include decisions made regarding the item stem, response options, response instructions, response effectiveness levels, and method(s) for scoring the responses (Campion et al., 2014; Legree, & Psotka, 2006; Oostrom, De Soete, & Lievens, 2015; Weekley, Ployhart, & Holtz, 2006).

The first step in developing an SJT is creating the item stems. The item stems are the situations that are presented to the examinees. There are two primary methods for developing the item stem; the critical incident method and the theory-based approach (Campion et al., 2014; Flanagan, 1954; Lievens & De Soete, 2015; Weekley et al., 2006).

- *Critical incident method*. In the critical incident method, individuals who are highly familiar with the job and job requirements (subject matter experts [SME]) are asked to give examples of situations that commonly occur on the job. Then, those situations are broken down into their constituent components (the antecedents of the situation, behavior during the situation, and consequence(s) of the situation) to create the item stem and possible response options.

- *Theory-based approach*. In a theory-based approach, items are written utilizing attributes surrounding the job based on either a job analysis or a review of the literature. When using a job analysis, a given job can be defined based in terms of the tasks it involves and the knowledge, skills, and abilities required for employees to accomplish them (Cascio & Aguinis, 2011). The two main purposes for a job analysis in this context are to inform the development of the SJT and to identify criteria to be used in the validation of the SJT (SIOP, 2003). When using a literature review, developers focus on the theory that exists around determinants of effective performance for a given field (Weekley et al., 2006). Specifically, they can engage in a comprehensive examination of the literature looking for which specific knowledge, skills, and abilities are necessary for performing a key job component (Stevens & Campion, 1999; Weekley et al., 2006).

After item stems have been developed, response options for each item stem must be generated (Lievens & De Soete, 2015). How individuals choose to respond to an item stem reflects different problem solving and interaction styles. As mentioned previously, SMEs frequently help develop aspects of SJTs, including response options. Their inputs are gathered through a variety of methods including interviews, focus groups, and survey responses to provide alternative courses of action for the situation stem (Motowidlo et al., 1997; Weekley et al., 2005). When it is not possible to obtain SME input in the writing process, SJT developers will often write the items themselves (Stevens & Campion, 1999; Weekley et al., 2006). However, even in this alternative development process, we recommended having SMEs review stems and response options after development to ensure realism.

Another critical component of SJT development is choosing the type of response instructions to provide examinees. Two commonly used formats are behavioral response instructions and knowledge-based response instructions (McDaniel, Hartman, & Grubb, 2003; Oostrom et al., 2015; Weekley et al., 2006).

- *Behavioral response instructions*. Behavioral response instructions evaluate the tendencies of examinees by asking them what their likely behavior and response *would* be in the provided situation (McDaniel et al., 2003).
- *Knowledge-based response instructions*. Knowledge-based response instructions look at the expertise of examinees by asking them what they *should* do given the provided situation (McDaniel et al., 2003).

Although behavioral instructions have higher correlations with personality measures, knowledge instructions tend to have higher correlations with cognitive ability measures (McDaniel & Nguyen, 2001; Lievens & De Soete, 2015; McDaniel et al., 2003; Weekley et al., 2006). However, there is a very low correlation between the two different instruction sets (Ployhart & Ehrhart, 2003). This suggests that they capture different aspects of the constructs being measured. Thus, different response instructions capture very different aspects of job candidates and should be chosen accordingly based on what test developers want to emphasize in their assessment of candidates (Oostrom et al., 2015; Whetzel & McDaniel, 2009). These choices can be made in several ways as discussed below.

Additionally, determinations must be made regarding the effectiveness of the response options. In practice, there are three commonly used methods for this: rational keying, empirical keying, and theoretical keying (Bergman et al., 2006; Campion et al., 2014; Lievens & De Soete, 2015; Oostrom et al., 2015; Weekley et al., 2006).

- *Rational keying*. In rational keying, SMEs are asked to evaluate the effectiveness of response options. Their judgments regarding the effectiveness of various responses are then pooled either using consensus methods or actuarial methods (McDaniel & Nguyen, 2001; Oostrom et al., 2015). With consensus methods, SMEs are asked to come to an agreement with one another about the relative effectiveness of each response option. In actuarial methods, a mean of SME ratings of effectiveness or some other normative approach is used.
- *Empirical keying*. In empirical keying, options are evaluated based on the extent to which the measure is related to a measured outcome variable (Bergman et al., 2006), usually job performance. There are two primary empirical keying methods, external and internal (Oostrom et al., 2015). In external methods, SJTs are generally administered to a large pilot sample and items are weighed based on their correlation with a criterion measure (Lievens et al., 2008; Oostrom et al., 2015). Internal methods, however, are much more commonly used and evaluate response options as scored on their interrelationships identified using factor analytic procedures.
- *Theoretical keying*. In a theory-based approach, response effectiveness is determined utilizing attributes of the job based on a job analysis or a review of the literature. This is similar to how the theory-based approach is used in other aspects of the SJT development process. It is the least frequently used method, and should only be utilized when the response options are constructed to reflect a theoretical model (Bergman et al., 2006; Oostrom et al., 2015). In this approach, developers should focus on how each of the response options relates back to the outcome of interest and then draw on the literature to see which options most strongly relate to the outcome (Weekley et al., 2006).

In addition to these three commonly used methods for determining response effectiveness, there are hybridized versions of these approaches that combine different methods (Bergman et al., 2006; Campion et al., 2014; Weekley et al., 2006).

The final determination in developing an SJT is the scoring method, or how responses are combined to create an overall score for the examinee. There are three primary methods used to score SJTs: forced-choice, Likert-type-scale (continuous), and combined methods (Bergman et al., 2006; Campion et al., 2014; Oostrom et al., 2015; Weekley et al., 2006).

- *Forced-choice methods*. A commonly used technique, forced-choice methods have one response option designated as the correct response, and all other options are considered incorrect. Scores are assigned based on correct and incorrect responses.
- *Likert-type-scale (continuous) methods*. In Likert-type-scale methods, instead of selecting a single response option, the examinees rate the effectiveness of all possible response options. Scores can be assigned in a few different ways. Scores are then assigned based on how similar the examinees effectiveness ratings are to SME ratings. Either they are scored on each response option in isolation, or they can be scored based on the difference between their rating and the SME rating (the more similar to the SME rating, the higher the score).
- *Combined methods*. Another way of scoring involves combining forced-choice and Likert-type-scale methods. Specifically, respondents can be asked to choose the most and least effective options. Then, they can be scored using either of the previous methods:

their score can be dependent on whether their responses were correct or they can be scores based on comparison to SME ratings.

Several considerations are involved in the implementation of SJTs. This starts with the scenario and response mediums and presentation (Campion et al., 2014). The scenario medium looks at how information is conveyed and presented to examinees. There are three primary methods for presenting a scenario and response options: written, interpersonal, and web-based formats (Campion et al., 2014). The written method of using paper and pencil is traditional and was one of the most commonly used formats (Campion et al., 2014), but has to some extent been replaced by web-based formats. Web-based formats include different ways to present the scenarios, such as written text blocks describing the scenario, 2D/3D graphics visually portraying the scenario, and videos (K. Horgen, personal communication, February 7, 2019). There are also different ways to present written response options. The two most common are static response options and interactive/branching response options. With static response options, examinees give the response options (using whichever response format is chosen in the development phase) and then are done with the item. With interactive/branching response options, the initial response option given by the examinee leads to subsequent questions based on the previous response(s) (Borlund & Ingwersen, 1997; Corstjens, Lievens, & Krumm, 2017). For example, after selecting on option, individuals can be asked how they would choose to implement it. Another example would be giving individuals a resulting consequence from their original choice and then give options for them to once again choose from regarding how to proceed based on that given consequence.

3.0 ITEM EXPOSURE ANALYSES

The purpose of this section is to describe analyses into the effects of item exposure on the NWAEs. As noted above, NWAE performance is one of several factors used in the NEAS to determine advancement. In particular, these analyses pertain to the Occupational Knowledge portion of the NWAE; this portion is comprised of 150 items specific to a Sailor's paygrade and rating. A few details regarding NWAEs:

- Examinees who do not advance may take a similar exam for a given rate in future administrations.
- E-4/5/6 NWAEs are administered twice per year, once in March and once in September. E-7 NWAEs are administered once per year in January.
- For E-4/5/6 rates, an item may not appear on the next two administrations once it has been administered. For E-7, an item may not appear on the next administration once it has been administered.
- A maximum of 33% of items appearing on an administration can appear on an eligible future administration (i.e., on the fourth administration for an E-4/5/6 NWAE, or on the third administration for an E-7 NWAE).

The Navy provided PDRI with data from eight E-4/5/6 administrations (March 2015 – September 2018) and four E-7 administrations (January 2015 – January 2018). A sample of rates was selected based on size, demographic composition (i.e., having sufficient representation from a variety of demographic groups), and centrality to the Navy's mission. Data were provided for the following rates:

- E-4: Aviation Boatswain's Mate Petty Officer 3rd Class (ABE3), Boatswain's Mate Petty Officer 3rd Class (BM3), Hospital Corpsman Petty Officer 3rd Class (HM3), Information Technician Petty Officer 3rd Class (IT3), Machine Accountant Petty Officer 3rd Class (MA3), Ship's Serviceman Petty Officer 3rd Class (SH3)
- E-5: Builder Petty Officer 2nd Class (BU2), Electrician's Mate Petty Officer 2nd Class (EM2), Electronics Technician Petty Officer 2nd Class (ET2), Hospital Corpsman Petty Officer 2nd Class (HM2), Information Technician Petty Officer 2nd Class (IT2), Machine Accountant Petty Officer 2nd Class (MA2), Machinist's Mate Nuclear Petty Officer 2nd Class (MMN2)
- E-6: Aviation Machinist's Mate Petty Officer 1st Class (AD1), Aviation Metalsmith 1st Class Petty Officer (AM1), Hospital Corpsman Petty Officer 1st Class (HM1), Information Technician Petty Officer 1st Class (IT1), Master-At-Arms Petty Officer First Class (MA1), Machinist's Mate 1st Class Petty Officer (MM1)
- E-7: Gunner's Mate Chief Petty Officer (GMC), Hospital Corpsman Chief Petty Officer (HMC),Information Technician Chief Petty Officer (ITC), Machine Accountant Chief Petty Officer (MAC), Operations Specialist Chief Petty Officer (OSC), Personnel Specialist Chief Petty Officer (PSC)

The provided data contained items that were deleted from scoring (i.e., no reference tied to the item, reference source had changed, etc.). Prior to analysis, these items were removed from the files.

Three sets of analyses were conducted:

Item parameter changes over time

- Item exposure may alter item parameters, even if the Sailors in the test-taking population for a rate change over time. Once an item has been exposed in a NWAE administration, it is possible that the item's content directly (e.g., by a Sailor remembering a specific item) or indirectly (e.g., by a theme influencing topics studied by future test-takers) could affect parameters among new and repeat test-takers. In this analysis, we examined whether item difficulties and item-total correlations change from an initial within-scope administration to the next administration.
- To analyze item difficulty changes over time, chi-square tests were conducted on the proportions of candidates who responded correctly at the initial within-scope administration and the next administration. To analyze item-total correlation changes over time, point-biserial correlations between correct/incorrect item responses and total scores for the Occupational Knowledge portion of the NWAE were calculated, and then differences were calculated using Fisher's procedure (Fisher, 1921). For both item difficulty and item-total correlation differences, items exhibiting significant ($p < .05$) differences were classified as having become "easier" or "harder" from time 1 to time 2, and the number of items within these categories was summed. Significant differences between the numbers of easier and harder items were assessed with a binomial test ($p < .05$), the null of which assumed an equal number of items in both categories.

Item parameter changes for repeat test-takers.

- Item parameters may change for examinees who see the same items multiple times. In this analysis, we examined whether or not item difficulties and item-total correlations changed for repeat test-takers on items that were viewed twice.
- To analyze item difficulty changes over time, McNemar's tests were conducted on the proportions of repeat candidates (i.e., those who saw the same item twice) who responded correctly at the initial within-scope administration and the next administration. To analyze item-total correlation changes over time, point-biserial correlations between correct/incorrect item responses and total scores for the Occupational Knowledge portion of the NWAE were calculated, and then differences were calculated using Fisher's procedure. For both item difficulty and item-total correlation differences, items that exhibited significant ($p < .05$) differences were classified as having become "easier" or "harder" from time 1 to time 2, and the number of items within these categories was summed. Significant differences between the numbers of easier and harder items were assessed with a binomial test ($p < .05$), the null of which assumed an equal number of items in both categories.

Candidate performance differences for initial vs. repeat exposures.

- It is possible that repeat test-takers perform better on items they see a second time versus items they see only once. In this analysis, we examined whether or not item difficulties and item-total correlations were different for items viewed a second time compared to items viewed only once.
- To analyze item difficulty changes over time, chi-square tests were conducted on the proportions of repeat items (i.e., subsequent viewings of an item that a candidate had seen before) and non-repeat items (i.e., items a candidate saw for the first time, regardless of whether or not the candidate saw those items in a later administration) to which a

candidate responded. Candidates exhibiting significant ($p < .05$) differences were classified as having performed "better" or "worse" on repeat items vs. non-repeat items, and the number of candidates within these categories was summed. Significant differences between the numbers of better and worse performing candidates were assessed with a binomial test ($p < .05$), the null of which assumed an equal number of candidates in both categories.

The first two analyses (i.e., item parameter changes over time and for repeat test-takers) were conducted by length of time between administrations. For E-4/5/6, results are presented for 2-series gaps, 3-series gaps, and 4-or-more series gaps between administrations. For E-7, results are presented for 1-series and 2-or-more series gaps between administrations.

All analyses were conducted by gender, race/ethnicity, and tenure-based group. Tenure groups were defined by time in paygrade, and for the purposes of data aggregation across paygrades, were classified as "low" and "high." Note that "low" and "high" tenure were broken down differently across paygrades, and subsequent tables reflect these splits. Specifically:

- E-4 was divided into low at less than or equal to 1 year and high at greater than 1 year.
- E-5 was divided into low at less than or equal to 1.5 years and high at greater than 1.5 years.
- E-6 was divided into low at less than or equal to 3.5 years and high at greater than 3.5 years.
- E-7 was divided into low at less than or equal to 3.5 years and high at greater than 3.5 years.

4.0 OVERALL RESULTS

First, results are presented across all within-scope rates. Tables 4 and 5 present total and demographic group sizes for each administration. The sample was predominately of high tenure, male, and Caucasian, though when non-Caucasian demographic groups were combined into one group, totals were similar to or greater than those of the Caucasian group.

Table 4. E-4/5/6 Sample Sizes by Administration

	Administration							
	227	**228**	**231**	**232**	**235**	**236**	**239**	**240**
Overall	27,444	27,963	26,449	27,336	27,712	25,986	27,022	24,984
Male	21,863	22,174	20,954	21,489	21,748	20,081	20,754	18,906
Female	5,581	5,789	5,495	5,847	5,964	5,905	6,268	6,078
Caucasian	11,786	12,208	11,600	12,136	12,352	11,738	12,087	11,264
African-American	4,338	4,512	4,226	4,394	4,430	4,046	4,207	3,917
Hispanic	5,179	5,151	4,974	5,130	5,193	4,821	5,119	4,730
Asian	2,166	2,187	2,112	2,221	2,269	2,185	2,293	2,111
Non-Caucasian	12,810	12,985	12,294	12,701	12,835	11,890	12,412	11,449
Low Tenure	10,730	11,328	10,009	10,666	9,927	9,521	10,014	8,992
High Tenure	16,714	16,635	16,440	16,670	17,785	16,465	17,008	15,992

Table 5. E-7 Sample Sizes by Administration

	Administration			
	226	**230**	**234**	**238**
Overall	5,611	6,233	6,412	6,133
Male	4,431	4,963	5,174	4,982
Female	1,180	1,270	1,238	1,151
Caucasian	2,227	2,416	2,421	2,273
African-American	1,347	1,455	1,452	1,284
Hispanic	981	1,152	1,176	1,156
Asian	698	757	753	694
Non-Caucasian	3,243	3,631	3,653	3,394
Low Tenure	1,914	1,840	1,618	1,415
High Tenure	3,697	4,393	4,794	4,718

4.1 Analysis 1: Item Parameter Changes Over Time

Tables 6 and 7 present the number of repeat items examined in Analysis 1. As was noted, results are between series with 1, 2, 3, or 4 or more administrations between the initial and subsequent administration.

Table 6. E-4/5/6 Repeat Items

Series	Administrations Between Series	# of Repeat Items
227-232	2	603
227-235	3	519
227-236	4	411
228-235	2	561
228-236	3	567
228-239	4	294
231-236	2	468
231-239	3	522
231-240	4	361
232-239	2	310
232-240	3	420
235-240	2	293
227-239	5	192
227-240	6	138
228-240	5	206
-	2 Combined	2,235
-	3 Combined	2,028
-	4+ Combined	1,602

Table 7. E-7 Repeat Items

Series	Administrations Between Series	# of Repeat Items
226-234	1	183
226-238	2	144
230-238	1	183
-	1 Combined	366

Table 8 presents the numbers of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-4/5/6 paygrades. For items with a 2-series gap between administrations, there was a significantly greater number of items that became easier than items that became harder. This was also true for Hispanic, Asian, Non-Caucasian, and High Tenure candidates. Though the number of items that became easier was also greater for Male, Female, Caucasian, and African-American candidates, these differences were non-significant. The number of items that became harder was significantly greater for Low Tenure candidates. Across 3-series and 4-series or more gaps, there was a significantly greater number of items that became harder than items that became easier. This effect holds for only some demographic groups in the 3-series gap results, but is true for all groups in the 4-series or more gap results. Looking across the sets of results, the change from a preponderance of easier items to a preponderance of harder items appears to be driven by an increase in the proportion of harder items. That is, there is little change in the proportion of easier items across each set of results, but the proportion of harder items increases as the gap between administrations increases.

Table 9 presents the numbers of items for which item-total correlation changed significantly from the initial within-scope administration to the next administration within the E-4/5/6 paygrades. For items with a 2-series gap between administrations, there was a significantly greater number of items with an increased item-total correlation than items with a decreased correlation. This was also true for Male, Female, Caucasian, Non-Caucasian, and High-Tenure candidates. Though the number of items with an increased item-total correlation was also greater for African-American, Hispanic, Asian, and Low-Tenure candidates, these differences were non-significant. For items with a 3-series gap between administrations, there was a significantly greater number of items with an increased item-total correlation than items with a decreased correlation overall and for all groups. For items with a 4-series or more gap between administrations, most results showed a non-significant difference between the number of items with increased and decreased correlations. However, there was a significantly greater number of items with a decreased item-total correlation than items with an increased correlation for Female, African-American, and Asian candidates.

Table 10 presents the number of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-7 paygrade. For items with a 1-series or 2-series gap between administrations, there was a significantly greater number of items that became easier than items that became harder overall and for each demographic group with the exception of Asian candidates, for which a non-significant difference existed for a 2-series gap.

Table 11 presents the numbers of items for which item-total correlation changed significantly from the initial within-scope administration to the next administration within the E-7 paygrade. For items with a 1-series or 2-series gap between administrations, differences in the numbers of correlation increases and decreases were non-significant overall and for most demographic groups. For 1-series gap results, a significantly greater number of items exhibited correlation increases for Female, Hispanic, Asian, and Non-Caucasian candidates. For 2-series gap results, a significantly greater number of items exhibited correlation decreases for Caucasian candidates, while a significantly greater number of items exhibited correlation increases for African-American, Asian, and Non-Caucasian candidates.

Table 8. E-4/5/6 Difficulty Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Easier		Harder			Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	444	19.87%	410	18.34%	0.03	393	19.38%	459	22.63%	0.00	326	20.35%	439	27.40%	0.00
Male	393	17.58%	371	16.60%	0.10	368	18.15%	414	20.41%	0.00	292	18.23%	398	24.84%	0.00
Female	224	10.02%	218	9.75%	0.32	230	11.34%	217	10.70%	0.17	171	10.67%	228	14.23%	0.00
Caucasian	304	13.60%	294	13.15%	0.25	307	15.14%	301	14.84%	0.34	254	15.86%	305	19.04%	0.00
African-American	199	8.90%	197	8.81%	0.42	192	9.47%	213	10.50%	0.05	133	8.30%	185	11.55%	0.00
Hispanic	218	9.75%	177	7.92%	0.00	199	9.81%	191	9.42%	0.26	158	9.86%	199	12.42%	0.00
Asian	154	6.89%	120	5.37%	0.00	143	7.05%	126	6.21%	0.06	98	6.12%	128	7.99%	0.00
Non-Caucasian	335	14.99%	298	13.33%	0.01	288	14.20%	316	15.58%	0.04	233	14.54%	329	20.54%	0.00
Low Tenure	289	12.93%	338	15.12%	0.00	290	14.30%	325	16.03%	0.01	227	14.17%	310	19.35%	0.00
High Tenure	397	17.76%	315	14.09%	0.00	347	17.11%	360	17.75%	0.21	271	16.92%	358	22.35%	0.00

Table 9. E-4/5/6 Correlation Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Increase		Decrease			Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	198	8.86%	158	7.07%	0.00	199	9.81%	146	7.20%	0.00	130	8.11%	135	8.43%	0.30
Male	210	9.40%	155	6.94%	0.00	199	9.81%	123	6.07%	0.00	125	7.80%	119	7.43%	0.27
Female	97	4.34%	74	3.31%	0.00	89	4.39%	71	3.50%	0.02	63	3.93%	83	5.18%	0.01
Caucasian	139	6.22%	121	5.41%	0.04	143	7.05%	97	4.78%	0.00	104	6.49%	108	6.74%	0.32
African-American	74	3.31%	68	3.04%	0.21	75	3.70%	56	2.76%	0.01	55	3.43%	72	4.49%	0.01
Hispanic	95	4.25%	85	3.80%	0.12	99	4.88%	83	4.09%	0.04	67	4.18%	63	3.93%	0.28
Asian	79	3.53%	78	3.49%	0.42	73	3.60%	57	2.81%	0.02	47	2.93%	66	4.12%	0.00
Non-Caucasian	166	7.43%	115	5.15%	0.00	142	7.00%	100	4.93%	0.00	89	5.56%	89	5.56%	0.47
Low Tenure	132	5.91%	127	5.68%	0.30	150	7.40%	100	4.93%	0.00	90	5.62%	96	5.99%	0.24
High Tenure	174	7.79%	124	5.55%	0.00	147	7.25%	102	5.03%	0.00	104	6.49%	105	6.55%	0.43

Table 10. E-7 Difficulty Changes

	1 Series Gap					2 Series Gap				
	Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*
Overall	85	23.22%	61	16.67%	0.00	38	26.39%	21	14.58%	0.00
Male	81	22.13%	56	15.30%	0.00	40	27.78%	18	12.50%	0.00
Female	40	10.93%	28	7.65%	0.01	22	15.28%	9	6.25%	0.00
Caucasian	57	15.57%	37	10.11%	0.00	33	22.92%	14	9.72%	0.00
African-American	49	13.39%	38	10.38%	0.03	22	15.28%	13	9.03%	0.01
Hispanic	48	13.11%	29	7.92%	0.00	19	13.19%	9	6.25%	0.00
Asian	32	8.74%	21	5.74%	0.01	15	10.42%	11	7.64%	0.08
Non-Caucasian	70	19.13%	53	14.48%	0.01	28	19.44%	18	12.50%	0.01
Low Tenure	57	15.57%	41	11.20%	0.00	26	18.06%	13	9.03%	0.00
High Tenure	73	19.95%	57	15.57%	0.01	38	26.39%	20	13.89%	0.00

Table 11. E-7 Correlation Changes

	1 Series Gap					2 Series Gap				
	Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*
Overall	22	6.01%	19	5.19%	0.20	12	8.33%	10	6.94%	0.20
Male	17	4.64%	19	5.19%	0.26	9	6.25%	10	6.94%	0.29
Female	17	4.64%	6	1.64%	0.00	7	4.86%	8	5.56%	0.27
Caucasian	11	3.01%	14	3.83%	0.14	3	2.08%	11	7.64%	0.00
African-American	11	3.01%	7	1.91%	0.05	6	4.17%	3	2.08%	0.03
Hispanic	17	4.64%	5	1.37%	0.00	6	4.17%	5	3.47%	0.24
Asian	11	3.01%	3	0.82%	0.00	10	6.94%	4	2.78%	0.00
Non-Caucasian	21	5.74%	12	3.28%	0.01	10	6.94%	6	4.17%	0.04
Low Tenure	15	4.10%	12	3.28%	0.15	6	4.17%	5	3.47%	0.24
High Tenure	19	5.19%	16	4.37%	0.18	9	6.25%	5	3.47%	0.03

In addition to analyses based on statistical significance, we also examined effect sizes. Tables 12-14 present effect sizes of difficulty changes for E-4/5/6 paygrades. Note that in these and similar tables, the sum of percentages of items that became easier and items that became harder do not always add up to 100% due to a small proportion of items that remained the same difficulty across administrations.

For administrations with a 2-series gap, a similar proportion of items became easier and harder overall for most groups, though Female candidates had a slightly higher proportion of easier items and low-tenure candidates had a slightly lower proportion of easier items relative to other groups. The vast majority of items exhibited negligible changes overall and for all groups. The proportion of items that had small or larger changes appeared to vary with group size. For example, the proportion of items exhibiting at least a small change (both easier and harder) was larger for smaller groups, such as Asian and Hispanic candidates, than overall or for Male candidates. As the length of time between administrations increased, the proportion of easier items decreased and the proportion of harder items increased. Though the vast majority of items exhibited negligible changes regardless of length of time between administrations, a greater

proportion of items exhibited small or larger differences as the length of time between administrations increased. Again, this may be in part due to decreased stability of effect size values as sample sizes decreased.

Table 12. E-4/5/6 Difficulty Change Effect Sizes- 2 Series Gap

	2 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	1142	51.10%	1052	86	4	0	1093	48.90%	984	105	4	0
Male	1124	50.29%	1026	94	4	0	1110	49.66%	996	110	4	0
Female	1190	53.24%	983	201	6	0	1044	46.71%	841	193	9	1
Caucasian	1153	51.59%	1023	125	5	0	1080	48.32%	949	127	3	1
African-American	1141	51.05%	935	202	3	1	1092	48.86%	839	240	13	0
Hispanic	1137	50.87%	916	215	5	1	1093	48.90%	867	222	4	0
Asian	1127	50.43%	764	345	15	3	1098	49.13%	743	324	27	4
Non-Caucasian	1165	52.13%	1043	119	2	1	1070	47.87%	940	126	4	0
Low Tenure	1068	47.79%	927	137	4	0	1167	52.21%	980	182	5	0
High Tenure	1168	52.26%	1022	142	4	0	1062	47.52%	936	122	4	0

Table 13. E-4/5/6 Difficulty Change Effect Sizes- 3 Series Gap

	3 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	956	47.14%	846	104	6	0	1072	52.86%	929	137	6	0
Male	955	47.09%	846	103	6	0	1071	52.81%	903	162	6	0
Female	989	48.77%	771	208	10	0	1038	51.18%	812	216	9	1
Caucasian	977	48.18%	818	153	6	0	1048	51.68%	867	173	7	1
African-American	985	48.57%	749	227	8	1	1034	50.99%	704	307	23	0
Hispanic	969	47.78%	745	213	11	0	1057	52.12%	801	242	14	0
Asian	1017	50.15%	661	320	30	6	987	48.67%	595	357	32	3
Non-Caucasian	970	47.83%	842	123	5	0	1057	52.12%	878	172	7	0
Low Tenure	951	46.89%	798	147	6	0	1077	53.11%	880	187	9	1
High Tenure	941	46.40%	820	115	6	0	1087	53.60%	900	179	8	0

Tables 15-17 present item-total correlation change effect sizes. Overall, there was a fairly similar proportion of items that exhibited increases and decreases regardless of length of time between administrations, though the proportion of items that exhibited increases was slightly higher after a 3-series gap than other lengths of time. Overall, the vast majority of items exhibited negligible changes, with a moderate proportion of small changes and few or zero medium or larger changes. As group sizes decreased, changes appeared to be less stable, and the proportion of medium and large changes increased.

Table 14. E-4/5/6 Difficulty Change Effect Sizes- 4+ Series Gap

	4+ Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	701	43.76%	617	82	2	0	901	56.24%	746	143	9	3
Male	693	43.26%	609	82	2	0	909	56.74%	749	148	9	3
Female	744	46.44%	578	164	2	0	853	53.25%	630	208	12	3
Caucasian	727	45.38%	612	111	4	0	875	54.62%	706	158	8	3
African-American	730	45.57%	562	161	7	0	868	54.18%	599	246	20	3
Hispanic	701	43.76%	536	160	5	0	901	56.24%	652	231	14	4
Asian	735	45.88%	471	246	16	2	860	53.68%	518	307	29	6
Non-Caucasian	697	43.51%	594	101	2	0	903	56.37%	702	190	8	3
Low Tenure	694	43.32%	582	108	4	0	908	56.68%	712	183	10	3
High Tenure	712	44.44%	599	111	2	0	890	55.56%	712	166	10	2

Table 15. E-4/5/6 Item-Total Correlation Change Effect Sizes- 2 Series Gap

	2 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	1130	50.56%	1000	129	1	0	1104	49.40%	968	136	0	0
Male	1150	51.45%	963	166	5	16	1090	48.77%	884	178	14	14
Female	1120	50.11%	729	370	20	1	1091	48.81%	718	371	1	1
Caucasian	1116	49.93%	844	247	7	18	1123	50.25%	835	256	16	16
African-American	1148	51.36%	718	399	27	4	1058	47.34%	667	383	4	4
Hispanic	1174	52.53%	742	393	37	2	1025	45.86%	644	369	6	6
Asian	1146	51.28%	495	499	129	23	1016	45.46%	476	458	41	41
Non-Caucasian	1142	51.10%	879	257	6	0	1084	48.50%	836	248	0	0
Low Tenure	1089	48.72%	801	257	16	15	1142	51.10%	835	275	16	16
High Tenure	1146	51.28%	912	228	6	0	1084	48.50%	862	222	0	0

Table 16. E-4/5/6 Item-Total Correlation Change Effect Sizes- 3 Series Gap

	3 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	1054	51.97%	880	170	4	0	973	47.98%	829	144	0	0
Male	1087	53.60%	874	199	5	9	939	46.30%	779	160	0	0
Female	1041	51.33%	651	355	32	3	960	47.34%	591	363	3	3
Caucasian	1054	51.97%	782	249	15	8	964	47.53%	730	234	0	0
African-American	1032	50.89%	588	385	52	7	952	46.94%	579	363	5	5
Hispanic	1012	49.90%	576	363	58	15	980	48.32%	608	350	11	11
Asian	1069	52.71%	453	443	138	35	856	42.21%	378	392	43	43
Non-Caucasian	1045	51.53%	781	249	13	2	977	48.18%	747	230	0	0
Low Tenure	1032	50.89%	723	285	16	8	984	48.52%	726	258	0	0
High Tenure	1064	52.47%	813	242	9	0	954	47.04%	737	217	0	0

Table 17. E-4/5/6 Item-Total Correlation Change Effect Sizes- 4+ Series Gap

	4+ Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	796	49.69%	683	113	0	0	804	50.19%	673	131	0	0
Male	790	49.31%	660	129	1	0	810	50.56%	676	134	0	0
Female	809	50.50%	500	284	23	2	759	47.38%	461	290	4	4
Caucasian	778	48.56%	549	223	6	0	815	50.87%	600	215	0	0
African-American	803	50.12%	451	309	39	4	772	48.19%	433	331	4	4
Hispanic	775	48.38%	442	299	31	3	800	49.94%	511	281	4	4
Asian	788	49.19%	365	320	77	26	716	44.69%	321	341	27	27
Non-Caucasian	776	48.44%	569	203	4	0	819	51.12%	616	203	0	0
Low Tenure	772	48.19%	561	208	3	0	820	51.19%	597	223	0	0
High Tenure	807	50.37%	611	190	6	0	788	49.19%	592	196	0	0

Tables 18 and 19 present effect sizes of difficulty changes for the E-7 paygrade. Overall and for most groups, there was a greater proportion of items that became easier than became harder regardless of length of time between administrations. As with the E-4/5/6 paygrades, the vast majority of items exhibited negligible changes, with very few medium or large differences. The proportion of small or larger changes tended to be greater as group size decreased.

Table 18. E-7 Difficulty Change Effect Sizes- 1 Series Gap

	1 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	199	54.37%	173	24	2	0	167	45.63%	144	20	2	1
Male	204	55.74%	175	27	2	0	162	44.26%	136	23	2	1
Female	204	55.74%	148	53	2	1	161	43.99%	118	38	5	0
Caucasian	202	55.19%	162	37	3	0	164	44.81%	137	22	4	1
African-American	205	56.01%	158	44	2	1	161	43.99%	127	30	3	1
Hispanic	191	52.19%	138	50	2	1	173	47.27%	138	32	2	1
Asian	206	56.28%	127	75	4	0	160	43.72%	115	41	3	1
Non-Caucasian	192	52.46%	160	29	3	0	174	47.54%	146	25	2	1
Low Tenure	188	51.37%	142	43	3	0	178	48.63%	140	34	3	1
High Tenure	206	56.28%	177	27	2	0	160	43.72%	134	23	2	1

Table 19. E-7 Difficulty Change Effect Sizes- 2 Series Gap

	2 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	82	56.94%	63	19	0	0	62	43.06%	50	12	0	0
Male	83	57.64%	64	19	0	0	61	42.36%	50	11	0	0
Female	84	58.33%	55	28	1	0	60	41.67%	47	13	0	0
Caucasian	83	57.64%	60	23	0	0	61	42.36%	50	10	1	0
African-American	79	54.86%	60	19	0	0	65	45.14%	51	14	0	0
Hispanic	79	54.86%	55	23	1	0	65	45.14%	48	17	0	0
Asian	72	50.00%	38	31	3	0	72	50.00%	47	25	0	0
Non-Caucasian	72	50.00%	55	17	0	0	72	50.00%	60	12	0	0
Low Tenure	76	52.78%	51	25	0	0	68	47.22%	57	11	0	0
High Tenure	78	54.17%	60	18	0	0	66	45.83%	54	12	0	0

Tables 20 and 21 present effect sizes of item-total correlation changes for the E-7 paygrade. Overall, there was a similar proportion of items that exhibited increases and decreases after a 1-series gap, but the proportion of items that exhibited increases was higher after a 2-series gap; these proportions varied by demographic group and did not exhibit a consistent pattern across 1- and 2-series gaps. As with the E-4/5/6 paygrades, the vast majority of items for most groups exhibited negligible changes, with very few medium or large differences. However, smaller groups exhibited more variability, and the proportion of small or larger changes tended to be greater as group size decreased.

Table 20. E-7 Item-Total Correlation Change Effect Sizes- 1 Series Gap

	1 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	186	50.82%	171	15	0	0	180	49.18%	163	17	0	0
Male	177	48.36%	157	20	0	0	189	51.64%	161	28	0	0
Female	206	56.28%	128	76	2	0	158	43.17%	91	65	1	1
Caucasian	178	48.63%	129	45	4	0	186	50.82%	130	56	0	0
African-American	195	53.28%	132	63	0	0	169	46.17%	111	58	0	0
Hispanic	190	51.91%	109	75	5	1	175	47.81%	109	66	0	0
Asian	179	48.91%	77	86	16	0	177	48.36%	85	86	3	3
Non-Caucasian	197	53.83%	173	23	1	0	169	46.17%	137	32	0	0
Low Tenure	196	53.55%	132	62	2	0	168	45.90%	108	60	0	0
High Tenure	171	46.72%	146	25	0	0	195	53.28%	158	37	0	0

Table 21. E-7 Item-Total Correlation Change Effect Sizes- 2 Series Gap

	2 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	78	54.17%	67	11	0	0	66	45.83%	55	11	0	0
Male	79	54.86%	68	11	0	0	65	45.14%	51	14	0	0
Female	72	50.00%	43	23	6	0	68	47.22%	37	31	0	0
Caucasian	61	42.36%	46	15	0	0	82	56.94%	58	24	0	0
African-American	78	54.17%	47	27	4	0	64	44.44%	46	18	0	0
Hispanic	81	56.25%	38	43	0	0	62	43.06%	38	24	0	0
Asian	75	52.08%	29	34	8	4	60	41.67%	25	31	2	2
Non-Caucasian	81	56.25%	61	20	0	0	63	43.75%	52	11	0	0
Low Tenure	78	54.17%	54	24	0	0	66	45.83%	45	21	0	0
High Tenure	74	51.39%	63	11	0	0	70	48.61%	51	19	0	0

4.1.1 Items Administered Prior to 2015

Though the current analyses do not encompass administrations prior to 2015, the Navy provided item history data indicating whether or not items had been administered prior to 2015. In this section, the results presented above in Analysis 1 are restricted to only those items that had been administered at least once prior to 2015.

Tables 22 and 23 present the number of repeat items examined in Analysis 1 that were administered prior to 2015.

Table 22. E-4/5/6 Repeat Items

Series	Administrations Between Series	# of Repeat Items
227-232	2	469
227-235	3	404
227-236	4	329
228-235	2	448
228-236	3	466
228-239	4	212
231-236	2	400
231-239	3	403
231-240	4	273
232-239	2	205
232-240	3	293
235-240	2	124
227-239	5	149
227-240	6	102
228-240	5	162
-	2 Combined	1,646
-	3 Combined	1,566
-	4+ Combined	1,277

Table 23. E-7 Repeat Items

Series	Administrations Between Series	# of Repeat Items
226-234	1	107
226-238	2	90
230-238	1	113
-	1 Combined	220

Table 24 presents the numbers of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-4/5/6 paygrades. For items with a 2-series gap between administrations, there was a non-significant difference overall. However, there was a significantly greater number of harder items among Female and Low-Tenure candidates. For items with a 3-series gap between administrations, there was a significantly greater number of items that became harder than items that became easier overall and for Male, African-American, Non-Caucasian, Low-Tenure, and High-Tenure candidates. For items with a 4-series gap between administrations, there was a significantly greater number of items that became harder than items that became easier overall and for each demographic group. Again, this appears to be driven by an increase in the proportion of harder items. That is, there is little change in the proportion of easier items across each set of results, but the proportion of harder items increased as the gap between administrations increased. Although it may seem

counterintuitive that items can become more difficult over time, there is some basis in the literature for this. Several studies indicate that when individuals retake an exam after a length of time, they consistently get the same answers wrong, suggesting that test takers are misinformed (i.e., their knowledge of the area being assessed is incorrect) regarding those items rather than uninformed (i.e., their knowledge of the area being assessed is incomplete; Feinberg, Raymond, Haist, 2015; Geving, Webb, & Davis, 2005).

Table 25 presents the numbers of items for which item-total correlation changed significantly from the initial within-scope administration to the next administration within the E-4/5/6 paygrades. For items with a 2-series gap between administrations, there was a significantly greater number of items with an increased item-total correlation than items with a decreased correlation. This was also true for Male, Non-Caucasian, and High-Tenure candidates. For items with a 3-series gap between administrations, there was a non-significant difference overall. However, there was a significantly greater number of items with an increased item-total correlation for Male, Caucasian, and Low-Tenure candidates. For items with a 4-series or more gap between administrations, most results showed a non-significant difference between the number of items with increased and decreased correlations. However, there was a significantly greater number of items with a decreased item-total correlation than items with an increased correlation for Female, African-American, and Asian candidates.

Table 26 presents the numbers of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-7 paygrade. For items with a 1-series gap between administrations, there was a non-significant difference overall between items that became easier and items that became harder. However, there was a significantly greater number of items that became easier for Hispanic and Asian candidates. For items with a 2-series gap between administrations, there was a significantly greater number of items that became easier than items that became harder overall and for each demographic group.

Table 27 presents the numbers of items for which item-total correlation changed significantly from the initial within-scope administration to the next administration within the E-7 paygrade. For items with a 1-series gap, an equal number of items exhibited increased and decreased correlations overall. There was a significantly greater number of items that exhibited increases for Female, Hispanic, and Asian candidates, while there was a significantly greater number of items that exhibited decreases for Caucasian candidates. For items with a 2-series gap between administrations, there was a significantly greater number of items that exhibited increased correlations overall, as well as for African-American, Asian and Low Tenure (<=3.5 years) candidates. There was a significantly greater number of items that exhibited decreased correlations for Caucasian candidates.

Table 24. E-4/5/6 Difficulty Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Easier		Harder			Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	288	17.50%	297	18.04%	0.27	273	17.43%	360	22.99%	0.00	226	18.42%	346	28.20%	0.00
Male	252	15.31%	267	16.22%	0.14	255	16.28%	322	20.56%	0.00	197	16.06%	310	25.26%	0.00
Female	131	7.96%	151	9.17%	0.03	159	10.15%	164	10.47%	0.32	113	9.21%	173	14.10%	0.00
Caucasian	202	12.27%	211	12.82%	0.24	216	13.79%	238	15.20%	0.05	175	14.26%	238	19.40%	0.00
African-American	118	7.17%	133	8.08%	0.07	124	7.92%	161	10.28%	0.00	89	7.25%	137	11.17%	0.00
Hispanic	134	8.14%	121	7.35%	0.10	135	8.62%	140	8.94%	0.31	114	9.29%	157	12.80%	0.00
Asian	84	5.10%	87	5.29%	0.34	89	5.68%	99	6.32%	0.13	61	4.97%	95	7.74%	0.00
Non-Caucasian	202	12.27%	209	12.70%	0.28	192	12.26%	241	15.39%	0.00	158	12.88%	256	20.86%	0.00
Low Tenure	198	12.03%	243	14.76%	0.00	199	12.71%	243	15.52%	0.00	165	13.45%	241	19.64%	0.00
High Tenure	246	14.95%	226	13.73%	0.07	234	14.94%	289	18.45%	0.00	182	14.83%	283	23.06%	0.00

Table 25. E-4/5/6 Correlation Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Increase		Decrease			Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	127	7.72%	108	6.56%	0.03	126	8.05%	114	7.28%	0.11	89	7.25%	104	8.48%	0.05
Male	134	8.14%	107	6.50%	0.00	126	8.05%	95	6.07%	0.00	85	6.93%	94	7.66%	0.14
Female	56	3.40%	56	3.40%	0.46	62	3.96%	51	3.26%	0.05	38	3.10%	62	5.05%	0.00
Caucasian	91	5.53%	93	5.65%	0.39	93	5.94%	76	4.85%	0.02	77	6.28%	86	7.01%	0.13
African-American	51	3.10%	44	2.67%	0.13	50	3.19%	41	2.62%	0.07	39	3.18%	59	4.81%	0.00
Hispanic	59	3.58%	55	3.34%	0.26	62	3.96%	66	4.21%	0.28	40	3.26%	46	3.75%	0.15
Asian	50	3.04%	56	3.40%	0.17	45	2.87%	45	2.87%	0.46	29	2.36%	49	3.99%	0.00
Non-Caucasian	105	6.38%	78	4.74%	0.00	74	4.73%	75	4.79%	0.42	60	4.89%	67	5.46%	0.16
Low Tenure	87	5.29%	97	5.89%	0.12	101	6.45%	76	4.85%	0.00	64	5.22%	75	6.11%	0.07
High Tenure	106	6.44%	83	5.04%	0.01	94	6.00%	81	5.17%	0.06	73	5.95%	78	6.36%	0.25

Table 26. E-7 Difficulty Changes

	1 Series Gap					2 Series Gap				
	Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*
Overall	46	20.91%	44	20.00%	0.33	29	32.22%	14	15.56%	0.00
Male	46	20.91%	41	18.64%	0.17	30	33.33%	11	12.22%	0.00
Female	21	9.55%	20	9.09%	0.35	18	20.00%	4	4.44%	0.00
Caucasian	33	15.00%	26	11.82%	0.06	23	25.56%	7	7.78%	0.00
African-American	28	12.73%	27	12.27%	0.37	18	20.00%	9	10.00%	0.00
Hispanic	29	13.18%	19	8.64%	0.01	17	18.89%	5	5.56%	0.00
Asian	20	9.09%	13	5.91%	0.02	14	15.56%	5	5.56%	0.00
Non-Caucasian	40	18.18%	34	15.45%	0.11	24	26.67%	9	10.00%	0.00
Low Tenure	34	15.45%	27	12.27%	0.07	17	18.89%	9	10.00%	0.00
High Tenure	40	18.18%	41	18.64%	0.39	31	34.44%	13	14.44%	0.00

Table 27. E-7 Correlation Changes

	1 Series Gap					2 Series Gap				
	Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*
Overall	15	6.82%	15	6.82%	0.43	11	12.22%	6	6.67%	0.02
Male	12	5.45%	14	6.36%	0.22	8	8.89%	5	5.56%	0.06
Female	9	4.09%	5	2.27%	0.03	5	5.56%	4	4.44%	0.21
Caucasian	6	2.73%	12	5.45%	0.01	3	3.33%	7	7.78%	0.01
African-American	4	1.82%	5	2.27%	0.21	6	6.67%	1	1.11%	0.00
Hispanic	13	5.91%	2	0.91%	0.00	6	6.67%	5	5.56%	0.23
Asian	6	2.73%	2	0.91%	0.00	5	5.56%	2	2.22%	0.02
Non-Caucasian	12	5.45%	9	4.09%	0.12	8	8.89%	5	5.56%	0.06
Low Tenure	9	4.09%	8	3.64%	0.28	6	6.67%	2	2.22%	0.00
High Tenure	11	5.00%	9	4.09%	0.19	7	7.78%	4	4.44%	0.05

Tables 28-30 present effect sizes of difficulty changes for E-4/5/6 paygrades. For administrations with a 2-series gap, a similar proportion of items became easier and harder overall and for most groups. As with the analysis using all items, the vast majority of items exhibited negligible changes overall and for all groups, and the proportion of items that had small or larger changes appeared to vary with group size. As the length of time between administrations increased, the proportion of easier items decreased and the proportion of harder items increased.

Table 28. E-4/5/6 Difficulty Change Effect Sizes- 2 Series Gap

	2 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	817	49.64%	760	56	1	0	829	50.36%	757	68	4	0
Male	803	48.78%	740	62	1	0	842	51.15%	765	73	4	0
Female	869	52.79%	724	141	4	0	776	47.14%	631	139	5	1
Caucasian	836	50.79%	743	91	2	0	809	49.15%	714	91	3	1
African-American	824	50.06%	685	136	3	0	820	49.82%	635	176	9	0
Hispanic	808	49.09%	647	159	2	0	834	50.67%	664	166	4	0
Asian	811	49.27%	555	247	9	0	826	50.18%	560	245	19	2
Non-Caucasian	838	50.91%	760	77	1	0	808	49.09%	718	86	4	0
Low Tenure	777	47.21%	678	98	1	0	869	52.79%	729	135	5	0
High Tenure	838	50.91%	746	91	1	0	808	49.09%	721	83	4	0

Table 29. E-4/5/6 Difficulty Change Effect Sizes- 3 Series Gap

	3 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	716	45.72%	639	73	4	0	850	54.28%	735	110	5	0
Male	708	45.21%	635	69	4	0	856	54.66%	725	126	5	0
Female	736	47.00%	576	153	7	0	829	52.94%	644	176	8	1
Caucasian	738	47.13%	625	109	4	0	825	52.68%	676	142	6	1
African-American	753	48.08%	573	174	6	0	808	51.60%	539	249	20	0
Hispanic	735	46.93%	576	151	8	0	829	52.94%	627	189	13	0
Asian	765	48.85%	494	243	24	4	780	49.81%	468	286	23	3
Non-Caucasian	725	46.30%	634	88	3	0	840	53.64%	697	137	6	0
Low Tenure	735	46.93%	627	105	3	0	831	53.07%	675	147	8	1
High Tenure	700	44.70%	619	77	4	0	866	55.30%	712	147	7	0

Table 30. E-4/5/6 Difficulty Change Effect Sizes- 4+ Series Gap

	4+ Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	511	41.65%	455	54	2	0	716	58.35%	592	115	7	2
Male	506	41.24%	451	53	2	0	721	58.76%	591	121	7	2
Female	554	45.15%	435	117	2	0	669	54.52%	495	164	8	2
Caucasian	536	43.68%	459	73	4	0	691	56.32%	555	128	6	2
African-American	547	44.58%	424	119	4	0	676	55.09%	474	187	13	2
Hispanic	516	42.05%	393	120	3	0	711	57.95%	504	194	10	3
Asian	546	44.50%	347	186	11	2	674	54.93%	404	242	24	4
Non-Caucasian	510	41.56%	437	71	2	0	715	58.27%	559	147	7	2
Low Tenure	507	41.32%	424	79	4	0	720	58.68%	559	150	9	2
High Tenure	524	42.71%	447	75	2	0	703	57.29%	568	127	6	2

Tables 31-33 present item-total correlation change effect sizes for E-4/5/6 paygrades. Overall, there was a similar proportion of items that exhibited increases and decreases regardless of length of time between administrations. Overall, the vast majority of items exhibited negligible changes, with a moderate proportion of small changes and few or zero medium or larger changes. As group sizes decreased, changes appeared to be less stable, and the proportion of medium and large changes increased.

Table 31. E-4/5/6 Item-Total Correlation Change Effect Sizes- 2 Series Gap

	2 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	832	50.55%	737	94	1	0	813	49.39%	712	101	0	0
Male	848	51.52%	713	125	3	7	805	48.91%	644	135	13	13
Female	817	49.64%	519	285	13	0	810	49.21%	526	284	0	0
Caucasian	818	49.70%	603	203	4	8	834	50.67%	603	203	14	14
African-American	839	50.97%	515	303	18	3	782	47.51%	478	300	2	2
Hispanic	858	52.13%	535	293	29	1	765	46.48%	464	289	6	6
Asian	834	50.67%	345	375	97	17	757	45.99%	351	352	27	27
Non-Caucasian	833	50.61%	636	194	3	0	807	49.03%	619	188	0	0
Low Tenure	820	49.82%	590	211	12	7	825	50.12%	588	209	14	14
High Tenure	819	49.76%	658	157	4	0	823	50.00%	652	171	0	0

Table 32. E-4/5/6 Item-Total Correlation Change Effect Sizes- 3 Series Gap

	3 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	804	51.34%	677	123	4	0	761	48.60%	642	119	0	0
Male	825	52.68%	665	149	5	6	739	47.19%	608	131	0	0
Female	794	50.70%	495	269	27	3	753	48.08%	459	288	3	3
Caucasian	793	50.64%	587	188	13	5	764	48.79%	581	183	0	0
African-American	779	49.74%	448	281	46	4	750	47.89%	451	289	5	5
Hispanic	764	48.79%	413	288	52	11	774	49.43%	471	285	9	9
Asian	813	51.92%	341	341	102	29	671	42.85%	287	316	34	34
Non-Caucasian	793	50.64%	597	183	12	1	768	49.04%	576	192	0	0
Low Tenure	784	50.06%	553	215	11	5	770	49.17%	566	204	0	0
High Tenure	820	52.36%	627	184	9	0	737	47.06%	562	175	0	0

Table 33. E-4/5/6 Item-Total Correlation Change Effect Sizes- 4+ Series Gap

	4+ Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	601	48.98%	516	85	0	0	624	50.86%	523	101	0	0
Male	600	48.90%	499	100	1	0	625	50.94%	524	101	0	0
Female	606	49.39%	372	217	16	1	592	48.25%	355	231	3	3
Caucasian	583	47.51%	408	170	5	0	637	51.92%	469	168	0	0
African-American	607	49.47%	348	231	25	3	596	48.57%	316	272	4	4
Hispanic	587	47.84%	335	223	26	3	618	50.37%	395	217	3	3
Asian	604	49.23%	271	255	58	20	555	45.23%	232	275	24	24
Non-Caucasian	582	47.43%	429	151	2	0	639	52.08%	483	156	0	0
Low Tenure	579	47.19%	421	155	3	0	639	52.08%	467	172	0	0
High Tenure	613	49.96%	471	136	6	0	609	49.63%	452	157	0	0

Tables 34 and 35 present effect sizes of difficulty changes for the E-7 paygrade. Overall and for most groups, there was a similar proportion of items that became easier and harder after a 1-series gap, but all groups saw a greater proportion that became easier after a 2-series gap. As with the E-4/5/6 paygrades, the vast majority of items exhibited negligible changes, with very few medium or large differences. The proportion of small or larger changes tended to be greater as group size decreased.

Table 34. E-7 Difficulty Change Effect Sizes- 1 Series Gap

	1 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	111	50.45%	99	11	1	0	109	49.55%	93	14	1	1
Male	112	50.91%	98	13	1	0	108	49.09%	91	15	1	1
Female	119	54.09%	90	27	1	1	100	45.45%	72	25	3	0
Caucasian	105	47.73%	85	19	1	0	115	52.27%	98	14	2	1
African-American	118	53.64%	95	21	1	1	102	46.36%	82	18	1	1
Hispanic	114	51.82%	84	28	1	1	105	47.73%	85	18	1	1
Asian	111	50.45%	68	40	3	0	109	49.55%	78	28	2	1
Non-Caucasian	106	48.18%	88	16	2	0	114	51.82%	96	16	1	1
Low Tenure	109	49.55%	85	22	2	0	111	50.45%	89	20	1	1
High Tenure	114	51.82%	99	14	1	0	106	48.18%	89	15	1	1

Table 35. E-7 Difficulty Change Effect Sizes- 2 Series Gap

	2 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	53	58.89%	37	16	0	0	37	41.11%	30	7	0	0
Male	53	58.89%	37	16	0	0	37	41.11%	31	6	0	0
Female	53	58.89%	31	22	0	0	37	41.11%	30	7	0	0
Caucasian	53	58.89%	38	15	0	0	37	41.11%	33	4	0	0
African-American	53	58.89%	38	15	0	0	37	41.11%	29	8	0	0
Hispanic	60	66.67%	39	20	1	0	30	33.33%	19	11	0	0
Asian	49	54.44%	26	21	2	0	41	45.56%	25	16	0	0
Non-Caucasian	50	55.56%	35	15	0	0	40	44.44%	34	6	0	0
Low Tenure	51	56.67%	36	15	0	0	39	43.33%	33	6	0	0
High Tenure	54	60.00%	38	16	0	0	36	40.00%	30	6	0	0

Tables 36 and 37 present effect sizes of item-total correlation changes for the E-7 paygrade. Overall, there was an equal proportion of items that exhibited increases and decreases after a 1-series gap, but the proportion of items that exhibited increases was higher after a 2-series gap. Most groups displayed a similar pattern of higher proportions of items exhibiting increases from 1- to 2-series gaps. The vast majority of items for most groups exhibited negligible changes, with very few medium or large differences. However, smaller groups exhibited more variability, and the proportion of small or larger changes tended to be greater as group size decreased.

Table 36. E-7 Item-Total Correlation Change Effect Sizes- 1 Series Gap

	1 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	110	50.00%	103	7	0	0	110	50.00%	98	12	0	0
Male	103	46.82%	94	9	0	0	117	53.18%	100	17	0	0
Female	122	55.45%	85	36	1	0	98	44.55%	57	39	1	1
Caucasian	105	47.73%	80	23	2	0	114	51.82%	80	34	0	0
African-American	110	50.00%	80	30	0	0	109	49.55%	71	38	0	0
Hispanic	112	50.91%	70	37	4	1	108	49.09%	70	38	0	0
Asian	114	51.82%	48	57	9	0	101	45.91%	47	50	2	2
Non-Caucasian	118	53.64%	106	11	1	0	102	46.36%	81	21	0	0
Low Tenure	112	50.91%	82	28	2	0	108	49.09%	69	39	0	0
High Tenure	101	45.91%	93	8	0	0	119	54.09%	98	21	0	0

Table 37. E-7 Item-Total Correlation Change Effect Sizes- 2 Series Gap

	2 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	55	61.11%	46	9	0	0	35	38.89%	29	6	0	0
Male	54	60.00%	47	7	0	0	36	40.00%	27	9	0	0
Female	50	55.56%	30	16	4	0	38	42.22%	20	18	0	0
Caucasian	43	47.78%	34	9	0	0	47	52.22%	33	14	0	0
African-American	53	58.89%	32	18	3	0	37	41.11%	23	14	0	0
Hispanic	52	57.78%	25	27	0	0	37	41.11%	23	14	0	0
Asian	47	52.22%	17	25	3	2	35	38.89%	16	19	0	0
Non-Caucasian	56	62.22%	41	15	0	0	34	37.78%	25	9	0	0
Low Tenure	53	58.89%	36	17	0	0	37	41.11%	27	10	0	0
High Tenure	48	53.33%	41	7	0	0	42	46.67%	33	9	0	0

4.1.2 Items Not Administered Prior to 2015

In this section, the results that were presented in the prior section of Analysis 1 are restricted to only those items that were not administered prior to 2015.

Tables 38 and 39 present the number of repeat items examined in Analysis 1 that were not administered prior to 2015.

Table 38. E-4/5/6 Repeat Items

Series	Administrations Between Series	# of Repeat Items
227-232	2	134
227-235	3	115
227-236	4	82
228-235	2	113
228-236	3	101
228-239	4	82
231-236	2	68
231-239	3	119
231-240	4	88
232-239	2	105
232-240	3	127
235-240	2	169
227-239	5	43
227-240	6	36
228-240	5	44
-	2 Combined	589
-	3 Combined	462
-	4+ Combined	375

Table 39. E-7 Repeat Items

Series	Administrations Between Series	# of Repeat Items
226-234	1	76
226-238	2	54
230-238	1	70
-	1 Combined	146

Table 40 presents the numbers of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-4/5/6 paygrades. For items with a 2-series or 3-series gap between administrations, there was a significantly greater number of items that became easier overall and for all demographic groups with the exception of Low-Tenure candidates. For items with a 4-series gap between administrations, there was a non-significant difference in the number of items that became easier or harder overall and for all groups with the exception of High-Tenure candidates.

Table 41 presents the number of items for which item-total correlation changed significantly from the initial within-scope administration to the next administration within the E-4/5/6 paygrades. For items with a 2-series gap between administrations, there was a significantly greater number of items with an increased item-total correlation than items with a decreased correlation. This was also true for Male, Female, Caucasian, Non-Caucasian, Low-Tenure, and High-Tenure candidates. For items with a 3-series gap between administrations, there was a significantly greater number of items with an increased item-total correlation than items with a decreased correlation overall and for all demographic groups with the exception of Females. For items with a 4-series or more gap between administrations, there was a significantly greater number of items with an increased item-total correlation than items with a decreased correlation. This was also true for Male and Hispanic candidates.

Table 42 presents the numbers of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-7 paygrade. For items with a 1-series gap between administrations, there was a significantly greater number of easier items overall and for all groups except Asians. For items with a 2-series gap between administrations, there was a non-significant difference between the numbers of easier and harder items overall. However, there was a significantly greater number of harder items for Asian and Non-Caucasian candidates, as well as a significantly greater number of easier items for Low-Tenure candidates.

Table 43 presents the numbers of items for which item-total correlation changed significantly from the initial within-scope administration to the next administration within the E-7 paygrade. For items with a 1-series gap, there was a non-significant difference between the number of items exhibiting increases and decreases. However, there was a significantly greater number of items that exhibited increases for Female, Caucasian, African-American, Asian, and Non-Caucasian candidates. For items with a 2-series gap between administrations, there was a significantly greater number of items that exhibited decreased correlations overall, as well as for

Male, Caucasian, African-American, and Low Tenure (<= 3.5 years) candidates. There was a significantly greater number of items that exhibited increased correlations for Asians.

Table 40. E-4/5/6 Difficulty Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Easier		Harder			Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	156	26.49%	113	19.19%	0.00	120	25.97%	99	21.43%	0.01	100	26.67%	93	24.80%	0.18
Male	141	23.94%	104	17.66%	0.00	113	24.46%	92	19.91%	0.01	95	25.33%	88	23.47%	0.18
Female	93	15.79%	67	11.38%	0.00	71	15.37%	53	11.47%	0.00	58	15.47%	55	14.67%	0.30
Caucasian	102	17.32%	83	14.09%	0.01	91	19.70%	63	13.64%	0.00	79	21.07%	67	17.87%	0.05
African-American	81	13.75%	64	10.87%	0.01	68	14.72%	52	11.26%	0.01	44	11.73%	48	12.80%	0.23
Hispanic	84	14.26%	56	9.51%	0.00	64	13.85%	51	11.04%	0.03	44	11.73%	42	11.20%	0.33
Asian	70	11.88%	33	5.60%	0.00	54	11.69%	27	5.84%	0.00	37	9.87%	33	8.80%	0.20
Non-Caucasian	133	22.58%	89	15.11%	0.00	96	20.78%	75	16.23%	0.00	75	20.00%	73	19.47%	0.37
Low Tenure	91	15.45%	95	16.13%	0.30	91	19.70%	82	17.75%	0.12	62	16.53%	69	18.40%	0.15
High Tenure	151	25.64%	89	15.11%	0.00	113	24.46%	71	15.37%	0.00	89	23.73%	75	20.00%	0.03

Table 41. E-4/5/6 Correlation Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Increase		Decrease			Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	71	12.05%	50	8.49%	0.00	73	15.80%	32	6.93%	0.00	41	10.93%	31	8.27%	0.03
Male	76	12.90%	48	8.15%	0.00	73	15.80%	28	6.06%	0.00	40	10.67%	25	6.67%	0.00
Female	41	6.96%	18	3.06%	0.00	27	5.84%	20	4.33%	0.05	25	6.67%	21	5.60%	0.16
Caucasian	48	8.15%	28	4.75%	0.00	50	10.82%	21	4.55%	0.00	27	7.20%	22	5.87%	0.12
African-American	23	3.90%	24	4.07%	0.36	25	5.41%	15	3.25%	0.01	16	4.27%	13	3.47%	0.16
Hispanic	36	6.11%	30	5.09%	0.11	37	8.01%	17	3.68%	0.00	27	7.20%	17	4.53%	0.01
Asian	29	4.92%	22	3.74%	0.06	28	6.06%	12	2.60%	0.00	18	4.80%	17	4.53%	0.34
Non-Caucasian	61	10.36%	37	6.28%	0.00	68	14.72%	25	5.41%	0.00	29	7.73%	22	5.87%	0.05
Low Tenure	45	7.64%	30	5.09%	0.00	49	10.61%	24	5.19%	0.00	26	6.93%	21	5.60%	0.11
High Tenure	68	11.54%	41	6.96%	0.00	53	11.47%	21	4.55%	0.00	31	8.27%	27	7.20%	0.18

Table 42. E-7 Difficulty Changes

	1 Series Gap					2 Series Gap				
	Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*
Overall	39	26.71%	17	11.64%	0.00	9	16.67%	7	12.96%	0.15
Male	35	23.97%	15	10.27%	0.00	10	18.52%	7	12.96%	0.08
Female	19	13.01%	8	5.48%	0.00	4	7.41%	5	9.26%	0.21
Caucasian	24	16.44%	11	7.53%	0.00	10	18.52%	7	12.96%	0.08
African-American	21	14.38%	11	7.53%	0.00	4	7.41%	4	7.41%	0.37
Hispanic	19	13.01%	10	6.85%	0.00	2	3.70%	4	7.41%	0.05
Asian	12	8.22%	8	5.48%	0.06	1	1.85%	6	11.11%	0.00
Non-Caucasian	30	20.55%	19	13.01%	0.00	4	7.41%	9	16.67%	0.01
Low Tenure	23	15.75%	14	9.59%	0.01	9	16.67%	4	7.41%	0.01
High Tenure	33	22.60%	16	10.96%	0.00	7	12.96%	7	12.96%	0.40

Table 43. E-7 Correlation Changes

	1 Series Gap					2 Series Gap				
	Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*
Overall	7	4.79%	4	2.74%	0.05	1	1.85%	4	7.41%	0.00
Male	5	3.42%	5	3.42%	0.38	1	1.85%	5	9.26%	0.00
Female	8	5.48%	1	0.68%	0.00	2	3.70%	4	7.41%	0.05
Caucasian	5	3.42%	2	1.37%	0.02	0	0.00%	4	7.41%	0.00
African-American	7	4.79%	2	1.37%	0.00	0	0.00%	2	3.70%	0.00
Hispanic	4	2.74%	3	2.05%	0.18	0	0.00%	0	0.00%	N/A
Asian	5	3.42%	1	0.68%	0.00	5	9.26%	2	3.70%	0.01
Non-Caucasian	9	6.16%	3	2.05%	0.00	2	3.70%	1	1.85%	0.08
Low Tenure	6	4.11%	4	2.74%	0.11	0	0.00%	3	5.56%	0.00
High Tenure	8	5.48%	7	4.79%	0.27	2	3.70%	1	1.85%	0.08

Tables 44-46 present effect sizes of difficulty changes for E-4/5/6 paygrades. For administrations with a 2-series gap, a greater proportion of items became easier than harder overall and for most groups. As with the analysis utilizing all items, the vast majority of items exhibited negligible changes overall and for all groups, and the proportion of items that had small or larger changes appeared to vary with group size. As the length of time between administrations increased, the proportion of easier items decreased and the proportion of harder items increased.

Table 44. E-4/5/6 Difficulty Change Effect Sizes - 2 Series Gap

	2 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	325	55.18%	292	30	3	0	264	44.82%	227	37	0	0
Male	321	54.50%	286	32	3	0	268	45.50%	231	37	0	0
Female	321	54.50%	259	60	2	0	268	45.50%	210	54	4	0
Caucasian	317	53.82%	280	34	3	0	271	46.01%	235	36	0	0
African-American	317	53.82%	250	66	0	1	272	46.18%	204	64	4	0
Hispanic	329	55.86%	269	56	3	1	259	43.97%	203	56	0	0
Asian	316	53.65%	209	98	6	3	272	46.18%	183	79	8	2
Non-Caucasian	327	55.52%	283	42	1	1	262	44.48%	222	40	0	0
Low Tenure	291	49.41%	249	39	3	0	298	50.59%	251	47	0	0
High Tenure	330	56.03%	276	51	3	0	254	43.12%	215	39	0	0

Table 45. E-4/5/6 Difficulty Change Effect Sizes - 3 Series Gap

	3 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	240	51.95%	207	31	2	0	222	48.05%	194	27	1	0
Male	247	53.46%	211	34	2	0	215	46.54%	178	36	1	0
Female	253	54.76%	195	55	3	0	209	45.24%	168	40	1	0
Caucasian	239	51.73%	193	44	2	0	223	48.27%	191	31	1	0
African-American	232	50.22%	176	53	2	1	226	48.92%	165	58	3	0
Hispanic	234	50.65%	169	62	3	0	228	49.35%	174	53	1	0
Asian	252	54.55%	167	77	6	2	207	44.81%	127	71	9	0
Non-Caucasian	245	53.03%	208	35	2	0	217	46.97%	181	35	1	0
Low Tenure	216	46.75%	171	42	3	0	246	53.25%	205	40	1	0
High Tenure	241	52.16%	201	38	2	0	221	47.84%	188	32	1	0

Table 46. E-4/5/6 Difficulty Change Effect Sizes - 4+ Series Gap

	4+ Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	190	50.67%	162	28	0	0	185	49.33%	154	28	2	1
Male	187	49.87%	158	29	0	0	188	50.13%	158	27	2	1
Female	190	50.67%	143	47	0	0	184	49.07%	135	44	4	1
Caucasian	191	50.93%	153	38	0	0	184	49.07%	151	30	2	1
African-American	183	48.80%	138	42	3	0	192	51.20%	125	59	7	1
Hispanic	185	49.33%	143	40	2	0	190	50.67%	148	37	4	1
Asian	189	50.40%	124	60	5	0	186	49.60%	114	65	5	2
Non-Caucasian	187	49.87%	157	30	0	0	188	50.13%	143	43	1	1
Low Tenure	187	49.87%	158	29	0	0	188	50.13%	153	33	1	1
High Tenure	188	50.13%	152	36	0	0	187	49.87%	144	39	4	0

Tables 47-49 present item-total correlation change effect sizes. Overall, there was a similar proportion of items that exhibited increases and decreases after a 2-series gap and 4-series gap. The proportion of items that exhibited increases was slightly higher after a 3-series gap than for 2- or 4-series gaps. Overall, the vast majority of items exhibited negligible changes, with a moderate proportion of small changes and no medium or larger changes. As group sizes decreased, changes appeared to be less stable, and the proportion of medium and large changes increased.

Table 47. E-4/5/6 Item-Total Correlation Change Effect Sizes - 2 Series Gap

	2 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	298	50.59%	263	35	0	0	291	49.41%	256	35	0	0
Male	302	51.27%	250	41	2	9	285	48.39%	240	43	1	1
Female	303	51.44%	210	85	7	1	281	47.71%	192	87	1	1
Caucasian	298	50.59%	241	44	3	10	289	49.07%	232	53	2	2
African-American	309	52.46%	203	96	9	1	276	46.86%	189	83	2	2
Hispanic	316	53.65%	207	100	8	1	260	44.14%	180	80	0	0
Asian	312	52.97%	150	124	32	6	259	43.97%	125	106	14	14
Non-Caucasian	309	52.46%	243	63	3	0	277	47.03%	217	60	0	0
Low Tenure	269	45.67%	211	46	4	8	317	53.82%	247	66	2	2
High Tenure	327	55.52%	254	71	2	0	261	44.31%	210	51	0	0

Table 48. E-4/5/6 Item-Total Correlation Change Effect Sizes - 3 Series Gap

	3 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	250	54.11%	203	47	0	0	212	45.89%	187	25	0	0
Male	262	56.71%	209	50	0	3	200	43.29%	171	29	0	0
Female	247	53.46%	156	86	5	0	207	44.81%	132	75	0	0
Caucasian	261	56.49%	195	61	2	3	200	43.29%	149	51	0	0
African-American	253	54.76%	140	104	6	3	202	43.72%	128	74	0	0
Hispanic	248	53.68%	163	75	6	4	206	44.59%	137	65	2	2
Asian	256	55.41%	112	102	36	6	185	40.04%	91	76	9	9
Non-Caucasian	252	54.55%	184	66	1	1	209	45.24%	171	38	0	0
Low Tenure	248	53.68%	170	70	5	3	214	46.32%	160	54	0	0
High Tenure	244	52.81%	186	58	0	0	217	46.97%	175	42	0	0

Table 49. E-4/5/6 Item-Total Correlation Change Effect Sizes - 4+ Series Gap

	4+ Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	195	52.00%	167	28	0	0	180	48.00%	150	30	0	0
Male	190	50.67%	161	29	0	0	185	49.33%	152	33	0	0
Female	203	54.13%	128	67	7	1	167	44.53%	106	59	1	1
Caucasian	195	52.00%	141	53	1	0	178	47.47%	131	47	0	0
African-American	196	52.27%	103	78	14	1	176	46.93%	117	59	0	0
Hispanic	188	50.13%	107	76	5	0	182	48.53%	116	64	1	1
Asian	184	49.07%	94	65	19	6	161	42.93%	89	66	3	3
Non-Caucasian	194	51.73%	140	52	2	0	180	48.00%	133	47	0	0
Low Tenure	193	51.47%	140	53	0	0	181	48.27%	130	51	0	0
High Tenure	194	51.73%	140	54	0	0	179	47.73%	140	39	0	0

Tables 50 and 51 present effect sizes of difficulty changes for the E-7 paygrade. Overall, and for most groups, there was a greater proportion of items that became easier than became harder regardless of length of time between administrations. This difference was greater after a 1-series gap than a 2-series gap. As with the E-4/5/6 paygrades, the vast majority of items exhibited negligible changes, with very few medium or large differences.

Table 50. E-7 Difficulty Change Effect Sizes - 1 Series Gap

	1 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	88	60.27%	74	13	1	0	58	39.73%	51	6	1	0
Male	92	63.01%	77	14	1	0	54	36.99%	45	8	1	0
Female	85	58.22%	58	26	1	0	61	41.78%	46	13	2	0
Caucasian	97	66.44%	77	18	2	0	49	33.56%	39	8	2	0
African-American	87	59.59%	63	23	1	0	59	40.41%	45	12	2	0
Hispanic	77	52.74%	54	22	1	0	68	46.58%	53	14	1	0
Asian	95	65.07%	59	35	1	0	51	34.93%	37	13	1	0
Non-Caucasian	86	58.90%	72	13	1	0	60	41.10%	50	9	1	0
Low Tenure	79	54.11%	57	21	1	0	67	45.89%	51	14	2	0
High Tenure	92	63.01%	78	13	1	0	54	36.99%	45	8	1	0

Table 51. E-7 Difficulty Change Effect Sizes - 2 Series Gap

	2 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	29	53.70%	26	3	0	0	25	46.30%	20	5	0	0
Male	30	55.56%	27	3	0	0	24	44.44%	19	5	0	0
Female	31	57.41%	24	6	1	0	23	42.59%	17	6	0	0
Caucasian	30	55.56%	22	8	0	0	24	44.44%	17	6	1	0
African-American	26	48.15%	22	4	0	0	28	51.85%	22	6	0	0
Hispanic	19	35.19%	16	3	0	0	35	64.81%	29	6	0	0
Asian	23	42.59%	12	10	1	0	31	57.41%	22	9	0	0
Non-Caucasian	22	40.74%	20	2	0	0	32	59.26%	26	6	0	0
Low Tenure	25	46.30%	15	10	0	0	29	53.70%	24	5	0	0
High Tenure	24	44.44%	22	2	0	0	30	55.56%	24	6	0	0

Tables 52 and 53 present effect sizes of item-total correlation changes for the E-7 paygrade. Overall, there was a similar proportion of items that exhibited increases and decreases after a 1-series gap, but the proportion of items that exhibited decreases was higher after a 2-series gap. These proportions varied by demographic group and did not exhibit a consistent pattern across 1- and 2-series gaps. As with the E-4/5/6 paygrades, the vast majority of items for most groups exhibited negligible changes, with very few medium or large differences. However, smaller groups exhibited more variability, and the proportion of small or larger changes tended to be greater as group size decreased.

Table 52. E-7 Item-Total Correlation Change Effect Sizes - 1 Series Gap

	1 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	76	52.05%	68	8	0	0	70	47.95%	65	5	0	0
Male	74	50.68%	63	11	0	0	72	49.32%	61	11	0	0
Female	84	57.53%	43	40	1	0	60	41.10%	34	26	0	0
Caucasian	73	50.00%	49	22	2	0	72	49.32%	50	22	0	0
African-American	85	58.22%	52	33	0	0	60	41.10%	40	20	0	0
Hispanic	78	53.42%	39	38	1	0	67	45.89%	39	28	0	0
Asian	65	44.52%	29	29	7	0	76	52.05%	38	36	1	1
Non-Caucasian	79	54.11%	67	12	0	0	67	45.89%	56	11	0	0
Low Tenure	84	57.53%	50	34	0	0	60	41.10%	39	21	0	0
High Tenure	70	47.95%	53	17	0	0	76	52.05%	60	16	0	0

Table 53. E-7 Item-Total Correlation Change Effect Sizes - 2 Series Gap

	2 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	23	42.59%	21	2	0	0	31	57.41%	26	5	0	0
Male	25	46.30%	21	4	0	0	29	53.70%	24	5	0	0
Female	22	40.74%	13	7	2	0	30	55.56%	17	13	0	0
Caucasian	18	33.33%	12	6	0	0	35	64.81%	25	10	0	0
African-American	25	46.30%	15	9	1	0	27	50.00%	23	4	0	0
Hispanic	29	53.70%	13	16	0	0	25	46.30%	15	10	0	0
Asian	28	51.85%	12	9	5	2	25	46.30%	9	12	2	2
Non-Caucasian	25	46.30%	20	5	0	0	29	53.70%	27	2	0	0
Low Tenure	25	46.30%	18	7	0	0	29	53.70%	18	11	0	0
High Tenure	26	48.15%	22	4	0	0	28	51.85%	18	10	0	0

Summary

E-4/5/6

- Most items did not change in difficulty over administrations. The percentage of items that become easier was relatively stable as the time between administrations increased. After a 2-series gap, the percentage of items that became easier was higher than the percentage of items that became harder. However, the percentage of items that became harder increased as the length of time between administrations increased, and the percentage of items that become harder was significantly greater after a 3-series or 4-series or more gap. This pattern (stable percentages of easier items and increasing percentages of harder items) was present in both items administered prior to 2015 and items not administered prior to 2015, though a greater percentage of items not administered prior to 2015 became easier than items administered prior to 2015.
- Most items did not exhibit significant item-total correlation changes over administrations. It was more common for items to increase in item-total correlation than to decrease, though the difference between numbers of items that increased or decreased was smaller as the time between administrations increased. A greater percentage of items not administered prior to 2015 exhibited correlation increases than items administered prior to 2015.

E-7

- Most items did not change in difficulty over administrations. A higher percentage of items became easier than became harder regardless of the length of time between administrations. However, results indicated a non-significant difference for a 1-series gap for items administered prior to 2015 and a non-significant difference for a 2-series gap for items not administered prior to 2015.
- Item-total correlations were fairly stable across administrations.

4.2 Analysis 2: Item Parameter Changes for Repeat Test-Takers

Tables 54 and 55 present the number of repeat test-takers for each possible series pair for the E-4/5/6 and E-7 paygrades. Note that within these tables, candidates can be counted in multiple

series pairs if they took more than two exams within the rating in the scope of the study. However, only their initial and second viewing of an item is included in the analyses in this section.

Table 54. E-4/5/6 Repeat Test-Taker Sample Sizes

Series	Admins. Between Series	Overall	Male	Female	Caucasian	African-American	Hispanic	Asian	Non-Caucasian	Low Tenure	High Tenure
227-232	2	9,325	7,494	1,831	3,868	1,592	1,827	818	4,606	3,665	5,660
227-235	3	6,604	5,353	1,251	2,701	1,160	1,301	603	3,337	2,826	3,778
227-236	4	4,237	3,465	772	1,728	715	853	409	2,155	1,930	2,307
228-235	2	10,306	8,300	2,006	4,303	1,777	1,985	928	5,110	6,017	4,289
228-236	3	6,638	5,392	1,246	2,752	1,114	1,310	623	3,328	4,007	2,631
228-239	4	4,964	4,064	900	2,008	845	1,001	504	2,556	3,069	1,895
231-236	2	9,024	7,284	1,740	3,755	1,511	1,799	834	4,481	6,270	2,754
231-239	3	6,788	5,533	1,255	2,771	1,158	1,379	649	3,441	4,811	1,977
231-240	4	4,533	3,729	804	1,789	790	962	431	2,361	3,258	1,275
232-239	2	10,029	8,003	2,026	4,220	1,656	2,000	904	4,920	7,523	2,506
232-240	3	6,621	5,308	1,313	2,679	1,149	1,351	615	3,370	5,024	1,597
235-240	2	9,337	7,379	1,958	3,855	1,582	1,904	821	4,649	7,317	2,020
227-239	5	3,145	2,597	548	1,245	554	651	320	1,654	1,468	1,677
227-240	6	2,122	1,759	363	808	397	464	230	1,169	1,033	1,089
228-240	5	3,293	2,703	590	1,275	600	692	343	1,766	2,058	1,235
-	2 Combined	48,021	38,460	9,561	20,001	8,118	9,515	4,305	23,766	30,792	17,229
-	3 Combined	26,651	21,586	5,065	10,903	4,581	5,341	2,490	13,476	16,668	9,983
-	4+ Combined	22,294	18,317	3,977	8,853	3,901	4,623	2,237	11,661	12,816	9,478

Table 55. E-7 Repeat Test-Taker Sample Sizes

Series	Admins. Between Series	Overall	Male	Female	Caucasian	African-American	Hispanic	Asian	Non-Caucasian	Low Tenure	High Tenure
226-234	1	3,075	2,453	622	1,188	738	553	403	1,813	1,208	1,867
226-238	2	2,153	1,739	414	824	499	412	280	1,273	916	1,237
230-238	1	3,332	2,682	650	1,260	774	645	418	1,976	1,951	1,381
-	1 Combined	6,407	5,135	1,272	2,448	1,512	1,198	821	3,789	3,159	3,248

Table 56 presents the number of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-4/5/6 paygrades. For items with a 2-series, 3-series, or 4-series or more gap between administrations, there was a significantly greater number of items that became easier overall and for all demographic groups. The percentage of items that became harder was relatively consistent across the different lengths of time between administrations, but the percentage of items that became easier decreased as the length of time between administrations increased overall and for all demographic groups. Note that only repeat items are included in these analyses. Although in this case many of the items became easier after repeated administrations, the literature suggests that this does not happen across all testing contexts. Answers that test takers change in between a first and second testing

are almost equally likely to be changed from wrong to right as they are to be changed from right to wrong (O'Neill, Sun, Peabody, & Royal, 2015).

Table 57 presents the numbers of items for which item-total correlations changed significantly from the initial within-scope administration to the next administration within the E-4/5/6 paygrades. For items with a 2-series, 3-series, or 4-series or more gap between administrations, there was a significantly greater number of items for which the item-total correlation increased overall and for all demographic groups with the exception of Females (for which there was a non-significant difference in the number of items after a 4-series gap). The percentage of items for which item-total correlations decreased was relatively consistent across the different lengths of time between administrations, but the percentage of items for which item-total correlations increased declined as the length of time between administrations increased overall and for most demographic groups.

Table 56. E-4/5/6 Difficulty Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Easier		Harder			Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	779	34.76%	113	5.04%	0.00	609	30.01%	126	6.21%	0.00	400	24.95%	81	5.05%	0.00
Male	711	31.73%	107	4.77%	0.00	573	28.24%	113	5.57%	0.00	372	23.21%	69	4.30%	0.00
Female	382	17.05%	84	3.75%	0.00	277	13.65%	52	2.56%	0.00	138	8.61%	31	1.93%	0.00
Caucasian	516	23.03%	89	3.97%	0.00	419	20.65%	80	3.94%	0.00	268	16.72%	61	3.81%	0.00
African-American	393	17.54%	50	2.23%	0.00	271	13.36%	47	2.32%	0.00	136	8.48%	35	2.18%	0.00
Hispanic	344	15.35%	57	2.54%	0.00	275	13.55%	61	3.01%	0.00	150	9.36%	28	1.75%	0.00
Asian	220	9.82%	40	1.78%	0.00	170	8.38%	24	1.18%	0.00	107	6.67%	25	1.56%	0.00
Non-Caucasian	595	26.55%	91	4.06%	0.00	454	22.38%	75	3.70%	0.00	279	17.40%	53	3.31%	0.00
Low Tenure	685	30.57%	87	3.88%	0.00	550	27.11%	61	3.01%	0.00	334	20.84%	64	3.99%	0.00
High Tenure	426	19.01%	93	4.15%	0.00	316	15.57%	24	1.18%	0.00	190	11.85%	53	3.31%	0.00

Table 57. E-4/5/6 Item-Total Correlation Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Increase		Decrease			Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	363	16.20%	38	1.70%	0.00	280	13.80%	21	1.03%	0.00	205	12.79%	42	2.62%	0.00
Male	279	12.45%	30	1.34%	0.00	202	9.96%	23	1.13%	0.00	145	9.05%	33	2.06%	0.00
Female	128	5.71%	25	1.12%	0.00	66	3.25%	20	0.99%	0.00	34	2.12%	27	1.68%	0.08
Caucasian	188	8.39%	24	1.07%	0.00	137	6.75%	20	0.99%	0.00	71	4.43%	28	1.75%	0.00
African-American	102	4.55%	21	0.94%	0.00	79	3.89%	25	1.23%	0.00	49	3.06%	29	1.81%	0.00
Hispanic	132	5.89%	22	0.98%	0.00	87	4.29%	21	1.03%	0.00	76	4.74%	17	1.06%	0.00
Asian	70	3.12%	17	0.76%	0.00	54	2.66%	15	0.74%	0.00	46	2.87%	33	2.06%	0.01
Non-Caucasian	201	8.97%	29	1.29%	0.00	36	1.77%	18	0.89%	0.00	30	1.87%	12	0.75%	0.00
Low Tenure	174	7.76%	28	1.25%	0.00	150	7.39%	13	0.64%	0.00	80	4.99%	20	1.25%	0.00
High Tenure	156	6.96%	17	0.76%	0.00	92	4.53%	11	0.54%	0.00	54	3.37%	12	0.75%	0.00

Table 58 presents the numbers of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-7 paygrade. For items with a 1-series or 2-series gap between administrations, there was a significantly greater number of items that became easier overall and for all demographic groups. The percentage of items that became harder decreased as the length of time between administrations increased, but the percentage of items that became easier increased as the length of time between administrations increased overall and for most demographic groups.

Table 59 presents the numbers of items for which item-total correlations changed significantly from the initial within-scope administration to the next administration within the E-7 paygrade. For items with a 1-series gap between administrations, there was a significantly greater number of items for which the item-total correlation increased overall and for all demographic groups with the exception of African-Americans (for which there was a non-significant difference in the number of items for the 1-series gap). For items with a 2-series gap between administrations, there was a significantly greater number of items for which the item-total correlation increased overall and for most demographic groups, the exceptions being Caucasian, African-American, and Asian candidates.

Table 58. E-7 Difficulty Changes

	1 Series Gap					2 Series Gap				
	Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*
Overall	138	33.33%	38	9.18%	0.00	62	36.69%	7	4.14%	0.00
Male	132	31.88%	40	9.66%	0.00	57	33.73%	5	2.96%	0.00
Female	72	17.39%	20	4.83%	0.00	39	23.08%	5	2.96%	0.00
Caucasian	94	22.71%	22	5.31%	0.00	46	27.22%	4	2.37%	0.00
African-American	75	18.12%	21	5.07%	0.00	30	17.75%	5	2.96%	0.00
Hispanic	71	17.15%	20	4.83%	0.00	35	20.71%	4	2.37%	0.00
Asian	57	13.77%	19	4.59%	0.00	30	17.75%	0	0.00%	0.00
Non-Caucasian	116	28.02%	34	8.21%	0.00	51	30.18%	4	2.37%	0.00
Low Tenure	113	27.29%	22	5.31%	0.00	58	34.32%	3	1.78%	0.00
High Tenure	128	30.92%	35	8.45%	0.00	46	27.22%	5	2.96%	0.00

Table 59. E-7 Item-Total Correlation Changes

	1 Series Gap					2 Series Gap				
	Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*
Overall	33	7.97%	6	1.45%	0.00	21	12.43%	6	3.55%	0.00
Male	22	5.31%	7	1.69%	0.00	18	10.65%	7	4.14%	0.00
Female	22	5.31%	3	0.72%	0.00	9	5.33%	3	1.78%	0.00
Caucasian	10	2.42%	4	0.97%	0.00	7	4.14%	4	2.37%	0.05
African-American	11	2.66%	7	1.69%	0.05	7	4.14%	7	4.14%	0.40
Hispanic	21	5.07%	5	1.21%	0.00	15	8.88%	5	2.96%	0.00
Asian	23	5.56%	3	0.72%	0.00	2	1.18%	3	1.78%	0.14
Non-Caucasian	32	7.73%	6	1.45%	0.00	24	14.20%	3	1.78%	0.00
Low Tenure	17	4.11%	4	0.97%	0.00	9	5.33%	5	2.96%	0.03
High Tenure	18	4.35%	2	0.48%	0.00	13	7.69%	6	3.55%	0.00

Summary

E-4/5/6

- The percentage of items that became easier for repeat test-takers was significantly greater than the percentage of items that became harder. Overall, over a third of items became easier after a 2-series gap, but this decreased to approximately one quarter of items after a 4-series gap.
- The percentage of items that exhibited an increased item-total correlation for repeat test-takers was significantly greater than the percentage of items that exhibited a decreased item-total correlation.

E-7

- The percentage of items that became easier for repeat test-takers was significantly greater than the percentage of items that became harder. The percentage of items that became easier increased as the length of time between administrations increased, while the percentage of items that became harder decreased as the length of time between administrations increased.
- The percentage of items that exhibited an increased item-total correlation for repeat test-takers was significantly greater than the percentage of items that exhibited a decreased item-total correlation.

4.3 Analysis 3: Candidate Performance Differences for Initial vs. Repeat Exposures

Tables 60 and 61 present the number of candidates who performed better or worse on repeat items (i.e., items they had seen in prior administrations) vs. non-repeat items (i.e., items they had not seen in prior administrations). As noted in the introduction, chi-square tests were conducted on the proportions of repeat items (i.e., subsequent viewings of an item that a candidate had seen before) and non-repeat items (i.e., items a candidate saw for the first time, regardless of whether or not the candidate saw those items in a later administration) to which a candidate responded. Candidates exhibiting significant ($p < .05$) differences were classified as having performed "better" or "worse" on repeat items vs. non-repeat items, and the number of candidates within these categories was summed. Significant differences between the numbers of better and worse

performing candidates were assessed with a binomial test ($p < .05$), the null of which assumed an equal number of candidates in both categories. In the tables that follow, candidates with non-significant differences are included in the "Repeat N" column, but not in the "Sig. Better [N/%]" or "Sig. Worse [N/%]" columns because they performed neither significantly better nor significantly worse.

For candidates at the E-4/5/6 and E-7 paygrades, a significantly greater number of candidates performed better on repeat items than on non-repeat items overall and in all demographic groups.

Table 60. E-4/5/6 Test-Taker Repeat Performance

	Repeat N	Sig. Better N	Sig. Better %	Sig. Worse N	Sig. Worse %	*p*
Overall	23,831	4,815	20.20%	617	2.59%	0.00
Male	18,882	3,876	20.53%	503	2.66%	0.00
Female	4,949	939	18.97%	114	2.30%	0.00
Caucasian	10,185	2,014	19.77%	275	2.70%	0.00
African-American	3,931	796	20.25%	99	2.52%	0.00
Hispanic	4,653	977	21.00%	125	2.69%	0.00
Asian	1,979	438	22.13%	35	1.77%	0.00
Non-Caucasian	11,445	2,397	20.94%	285	2.49%	0.00
Low Tenure	17,080	3,584	20.98%	394	2.31%	0.00
High Tenure	6,751	1,231	18.23%	223	3.30%	0.00

Table 61. E-7 Test-Taker Repeat Performance

	Repeat N	Sig. Better N	Sig. Better %	Sig. Worse N	Sig. Worse %	*p*
Overall	4,531	514	11.34%	35	0.77%	0.00
Male	3,625	382	10.54%	32	0.88%	0.00
Female	906	132	14.57%	3	0.33%	0.00
Caucasian	1,735	159	9.16%	17	0.98%	0.00
African-American	1,079	127	11.77%	8	0.74%	0.00
Hispanic	838	94	11.22%	5	0.60%	0.00
Asian	569	108	18.98%	2	0.35%	0.00
Non-Caucasian	2,675	344	12.86%	18	0.67%	0.00
Low Tenure	2,350	276	11.74%	10	0.43%	0.00
High Tenure	2,181	238	10.91%	25	1.15%	0.00

Summary

Most candidates at the E-4/5/6 and E-7 paygrades do not perform significantly better or worse on repeat items compared to non-repeat items. Of those that do perform better or worse, a significantly greater number of candidates performed better on repeat items than on non-repeat items overall and in all demographic groups.

5.0 E-4 OVERALL RESULTS

The results in this section are for the E-4 paygrade. Table 62 presents total and demographic group sizes for each administration. The sample was predominately of high tenure, male, and Caucasian, though when non-Caucasian demographic groups were combined into one group, totals were greater than those of the Caucasian group.

Table 62. E-4 Sample Sizes by Administration

	Administration							
	227	**228**	**231**	**232**	**235**	**236**	**239**	**240**
Overall	10,478	10,100	8,885	8,682	8,150	7,718	7,634	7,170
Male	8,051	7,679	6,717	6,468	5,989	5,480	5,330	4,764
Female	2,427	2,421	2,168	2,214	2,161	2,238	2,304	2,406
Caucasian	4,335	4,272	3,769	3,782	3,548	3,448	3,422	3,236
African-American	1,950	1,957	1,734	1,681	1,579	1,443	1,349	1,288
Hispanic	1,856	1,746	1,565	1,569	1,491	1,377	1,399	1,317
Asian	816	775	710	723	679	652	655	617
Non-Caucasian	4,957	4,779	4,248	4,174	3,924	3,638	3,532	3,331
Low Tenure	2,632	2,519	1,938	2,146	1,800	1,907	1,988	2,036
High Tenure	7,846	7,581	6,947	6,536	6,350	5,811	5,646	5,134

5.1 Analysis 1: Item Parameter Changes Over Time

Table 63 presents the number of repeat items examined in Analysis 1 for the E-4 paygrade. As was noted, results are presented for series with 2, 3, or 4 or more administrations between the initial and subsequent administration.

Table 64 presents the numbers of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-4 paygrade. For items with a 2-series gap between administrations, there was a non-significant difference between the number of items that became easier and the number of items that became harder overall. This was also true for all demographic groups with the exception of Low Tenure candidates, for which the number of items that became harder was significantly greater than the number of items that became easier. Across 3-series and 4-series or more gaps, there was a significantly greater number of items that became harder than items that became easier. This effect holds for all demographic groups in the 3-series gap results with the exception of Asians, and holds for all demographic groups in the 4-series or more gap results. Looking across the sets of results, the change from a non-significant difference in the 2-series gap to a preponderance of harder items as the gap increases appears to be primarily driven by an increase in the proportion of harder items. That is, there is a much smaller change in the proportion of easier items across each set of results compared to the increase in the proportion of harder items as the gap between administrations increases.

Table 63. E-4 Repeat Items

Series	Administrations Between Series	# of Repeat Items
227-232	2	261
227-235	3	156
227-236	4	112
228-235	2	185
228-236	3	203
228-239	4	101
231-236	2	170
231-239	3	161
231-240	4	117
232-239	2	120
232-240	3	124
235-240	2	127
227-239	5	45
227-240	6	33
228-240	5	46
-	2 Combined	863
-	3 Combined	644
-	4+ Combined	454

Table 65 presents the numbers of items for which item-total correlation changed significantly from the initial within-scope administration to the next administration within the E-4 paygrade. For items with a 2-series or 3-series gap between administrations, there was a non-significant difference overall between the number of items with an increased item-total correlation and items with a decreased correlation. This was also true for all demographic groups after a 2-series gap with the exceptions of Male, Female, and High-Tenure candidates, as well as for all demographic subgroups after a 3-series gap with the exceptions of Male and African-American candidates. For these demographic groups, there was a significantly greater number of items with an increased item-total correlation than items with a decreased correlation. For items with a 4-series or more gap between administrations, there was a significantly greater number of items with a decreased item-total correlation than items with an increased correlation overall as well as for Female, Caucasian, Hispanic, Non-Caucasian, Low-Tenure, and High-Tenure candidates.

Table 64. E-4 Difficulty Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Easier		Harder			Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	142	16.45%	152	17.61%	0.17	88	13.66%	170	26.40%	0.00	70	15.42%	134	29.52%	0.00
Male	125	14.48%	131	15.18%	0.26	81	12.58%	153	23.76%	0.00	55	12.11%	120	26.43%	0.00
Female	90	10.43%	93	10.78%	0.34	65	10.09%	85	13.20%	0.00	43	9.47%	82	18.06%	0.00
Caucasian	97	11.24%	107	12.40%	0.13	70	10.87%	99	15.37%	0.00	54	11.89%	86	18.94%	0.00
African-American	78	9.04%	83	9.62%	0.25	43	6.68%	106	16.46%	0.00	34	7.49%	79	17.40%	0.00
Hispanic	73	8.46%	62	7.18%	0.07	48	7.45%	70	10.87%	0.00	29	6.39%	59	13.00%	0.00
Asian	44	5.10%	44	5.10%	0.46	41	6.37%	46	7.14%	0.19	20	4.41%	42	9.25%	0.00
Non-Caucasian	104	12.05%	119	13.79%	0.05	72	11.18%	121	18.79%	0.00	51	11.23%	107	23.57%	0.00
Low Tenure	87	10.08%	110	12.75%	0.00	63	9.78%	97	15.06%	0.00	52	11.45%	83	18.28%	0.00
High Tenure	124	14.37%	124	14.37%	0.48	86	13.35%	145	22.52%	0.00	63	13.88%	120	26.43%	0.00

Table 65. E-4 Correlation Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Increase		Decrease			Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	66	7.65%	65	7.53%	0.42	54	8.39%	58	9.01%	0.26	31	6.83%	48	10.57%	0.00
Male	69	8.00%	50	5.79%	0.00	59	9.16%	41	6.37%	0.00	37	8.15%	39	8.59%	0.33
Female	39	4.52%	30	3.48%	0.04	30	4.66%	28	4.35%	0.31	21	4.63%	31	6.83%	0.01
Caucasian	37	4.29%	38	4.40%	0.39	41	6.37%	34	5.28%	0.10	23	5.07%	32	7.05%	0.03
African-American	30	3.48%	29	3.36%	0.38	35	5.43%	20	3.11%	0.00	23	5.07%	28	6.17%	0.12
Hispanic	36	4.17%	33	3.82%	0.26	29	4.50%	36	5.59%	0.08	12	2.64%	26	5.73%	0.00
Asian	22	2.55%	27	3.13%	0.12	25	3.88%	20	3.11%	0.11	17	3.74%	19	4.19%	0.26
Non-Caucasian	54	6.26%	51	5.91%	0.30	41	6.37%	37	5.75%	0.22	25	5.51%	33	7.27%	0.04
Low Tenure	36	4.17%	39	4.52%	0.27	38	5.90%	30	4.66%	0.06	17	3.74%	31	6.83%	0.00
High Tenure	60	6.95%	48	5.56%	0.04	47	7.30%	46	7.14%	0.40	27	5.95%	39	8.59%	0.01

Tables 66-68 present effect sizes of difficulty changes for the E-4 paygrade. For administrations with a 2-series gap, the vast majority of items exhibited negligible changes overall and for all groups. The proportion of items that had small or larger changes appeared to vary with group size. For example, the proportion of items exhibiting at least a small change (both easier and harder) was larger for smaller groups, such as Asians and Hispanics, than for larger groups (e.g., Overall or Male candidates). As the length of time between administrations increased, the proportion of easier items decreased and the proportion of harder items increased. Though the vast majority of items exhibited negligible changes regardless of length of time between administrations, a greater proportion of items exhibited small or larger differences after a 3-series or 4-series gap compared to after a 2-series gap.

Table 66. E-4 Difficulty Change Effect Sizes- 2 Series Gap

	2 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	424	49.13%	390	33	1	0	439	50.87%	390	47	2	0
Male	401	46.47%	366	34	1	0	461	53.42%	410	49	2	0
Female	459	53.19%	377	80	2	0	403	46.70%	318	79	5	1
Caucasian	436	50.52%	382	53	1	0	425	49.25%	366	57	1	1
African-American	419	48.55%	347	72	0	0	444	51.45%	350	91	3	0
Hispanic	417	48.32%	325	91	0	1	444	51.45%	321	122	1	0
Asian	411	47.62%	255	148	7	1	447	51.80%	273	161	13	0
Non-Caucasian	426	49.36%	373	52	1	0	437	50.64%	370	65	2	0
Low Tenure	400	46.35%	336	63	1	0	463	53.65%	372	88	3	0
High Tenure	432	50.06%	381	50	1	0	426	49.36%	367	57	2	0

Table 67. E-4 Difficulty Change Effect Sizes- 3 Series Gap

	3 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	279	43.32%	243	35	1	0	365	56.68%	302	60	3	0
Male	273	42.39%	241	31	1	0	371	57.61%	295	73	3	0
Female	281	43.63%	212	67	2	0	362	56.21%	270	86	6	0
Caucasian	290	45.03%	239	50	1	0	351	54.50%	282	63	6	0
African-American	277	43.01%	222	54	1	0	364	56.52%	241	119	4	0
Hispanic	280	43.48%	197	80	3	0	362	56.21%	253	103	6	0
Asian	310	48.14%	191	110	8	1	332	51.55%	176	144	12	0
Non-Caucasian	276	42.86%	236	39	1	0	367	56.99%	289	74	4	0
Low Tenure	301	46.74%	245	53	3	0	343	53.26%	260	77	6	0
High Tenure	269	41.77%	228	40	1	0	375	58.23%	289	82	4	0

Table 68. E-4 Difficulty Change Effect Sizes- 4+ Series Gap

	4+ Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	191	42.07%	174	17	0	0	263	57.93%	210	49	4	0
Male	184	40.53%	168	16	0	0	270	59.47%	210	55	5	0
Female	194	42.73%	151	43	0	0	258	56.83%	177	76	5	0
Caucasian	196	43.17%	171	24	1	0	258	56.83%	202	53	3	0
African-American	186	40.97%	146	39	1	0	268	59.03%	176	88	4	0
Hispanic	193	42.51%	147	45	1	0	261	57.49%	172	83	5	1
Asian	205	45.15%	126	72	7	0	249	54.85%	130	108	10	1
Non-Caucasian	177	38.99%	152	25	0	0	276	60.79%	200	73	3	0
Low Tenure	195	42.95%	161	33	1	0	259	57.05%	192	63	4	0
High Tenure	173	38.11%	145	28	0	0	281	61.89%	216	61	4	0

Tables 69-71 present item-total correlation change effect sizes. Overall, the vast majority of items exhibited negligible changes, with a moderate proportion of small changes and few or no medium or larger changes. As group sizes decreased, changes appeared to be less stable, and the proportion of medium and large changes increased. There was a fairly similar proportion of items that exhibited increases and decreases regardless of length of time between administrations, though the proportion of items that exhibited decreases was higher after a 4-series gap than other lengths of time.

Table 69. E-4 Item-Total Correlation Change Effect Sizes- 2 Series Gap

	2 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	419	48.55%	358	61	0	0	444	51.45%	367	77	0	0
Male	424	49.13%	348	75	1	0	439	50.87%	330	109	0	0
Female	435	50.41%	280	143	11	1	413	47.86%	276	135	1	1
Caucasian	436	50.52%	312	120	4	0	422	48.90%	307	115	0	0
African-American	430	49.83%	272	153	5	0	423	49.02%	266	155	1	1
Hispanic	437	50.64%	241	169	26	1	401	46.47%	225	174	1	1
Asian	430	49.83%	175	178	63	14	392	45.42%	179	177	18	18
Non-Caucasian	413	47.86%	300	112	1	0	449	52.03%	324	125	0	0
Low Tenure	411	47.62%	270	131	10	0	441	51.10%	295	146	0	0
High Tenure	417	48.32%	331	84	2	0	445	51.56%	347	98	0	0

Table 70. E-4 Item-Total Correlation Change Effect Sizes- 3 Series Gap

	3 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	320	49.69%	250	70	0	0	323	50.16%	258	65	0	0
Male	320	49.69%	234	84	2	0	322	50.00%	245	77	0	0
Female	319	49.53%	203	106	10	0	316	49.07%	182	132	1	1
Caucasian	320	49.69%	212	100	8	0	317	49.22%	216	101	0	0
African-American	335	52.02%	187	133	15	0	301	46.74%	191	110	0	0
Hispanic	306	47.52%	146	129	28	3	311	48.29%	173	132	3	3
Asian	307	47.67%	93	132	73	9	280	43.48%	116	136	14	14
Non-Caucasian	320	49.69%	219	100	1	0	322	50.00%	224	98	0	0
Low Tenure	316	49.07%	196	113	7	0	317	49.22%	213	104	0	0
High Tenure	309	47.98%	232	74	3	0	330	51.24%	242	88	0	0

Table 71. E-4 Item-Total Correlation Change Effect Sizes- 4+ Series Gap

	4+ Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	205	45.15%	160	45	0	0	248	54.63%	191	57	0	0
Male	210	46.26%	160	50	0	0	243	53.52%	187	56	0	0
Female	217	47.80%	141	65	9	2	219	48.24%	130	87	1	1
Caucasian	201	44.27%	125	72	4	0	248	54.63%	176	72	0	0
African-American	221	48.68%	128	82	11	0	224	49.34%	129	95	0	0
Hispanic	211	46.48%	103	90	18	0	225	49.56%	134	87	2	2
Asian	212	46.70%	87	87	29	9	208	45.81%	83	99	13	13
Non-Caucasian	218	48.02%	150	68	0	0	233	51.32%	150	83	0	0
Low Tenure	211	46.48%	147	62	2	0	237	52.20%	153	84	0	0
High Tenure	215	47.36%	151	63	1	0	234	51.54%	172	62	0	0

5.1.1 Items Administered Prior to 2015

Table 72 presents the number of repeat items examined in Analysis 1 that were administered prior to 2015 for the E-4 paygrade.

Table 72. E-4 Repeat Items

Series	Administrations Between Series	# of Repeat Items
227-232	2	221
227-235	3	129
227-236	4	97
228-235	2	151
228-236	3	170
228-239	4	72
231-236	2	150
231-239	3	115
231-240	4	79
232-239	2	76
232-240	3	82
235-240	2	47
227-239	5	30
227-240	6	21
228-240	5	38
-	2 Combined	645
-	3 Combined	496
-	4+ Combined	337

Table 73 presents the numbers of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-4 paygrade. For items with a 2-series gap between administrations, there was a non-significant difference overall. However, there was a significantly greater number of harder items among Non-Caucasian and Low-Tenure candidates. For items with a 3-series gap between administrations, there was a significantly greater number of items that became harder than items that became easier overall and for all demographic groups with the exception of Asian candidates. For items with a 4-series gap between administrations, there was a significantly greater number of items that became harder than items that became easier overall and for all groups. Again, this appears to be primarily driven by an increase in the proportion of harder items. That is, there is little change in the proportion of easier items across each set of results, but the proportion of harder items increased as the gap between administrations increased.

Table 74 presents the numbers of items for which item-total correlation changed significantly from the initial within-scope administration to the next administration within the E-4 paygrade. For items with a 2-series or 3-series gap between administrations, there was a non-significant difference between the number of items with an increased item-total correlation and the number of items with a decreased correlation. This was true for all demographic groups except for Males with a 2-series gap and African-Americans and Hispanics with a 3-series gap. For items with a 4-series or larger gap between administrations, there was a significantly greater number of items with a decreased item-total correlation than items with an increased correlation overall as well as for Female, Caucasian, African-American, Hispanic, Low Tenure, and High Tenure candidates.

Table 73. E-4 Difficulty Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Easier		Harder			Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	96	14.88%	107	16.59%	0.10	64	12.90%	129	26.01%	0.00	43	12.76%	109	32.34%	0.00
Male	85	13.18%	93	14.42%	0.16	60	12.10%	118	23.79%	0.00	31	9.20%	95	28.19%	0.00
Female	58	8.99%	65	10.08%	0.15	48	9.68%	64	12.90%	0.01	23	6.82%	67	19.88%	0.00
Caucasian	64	9.92%	73	11.32%	0.11	54	10.89%	79	15.93%	0.00	32	9.50%	71	21.07%	0.00
African-American	55	8.53%	60	9.30%	0.22	33	6.65%	78	15.73%	0.00	24	7.12%	64	18.99%	0.00
Hispanic	47	7.29%	41	6.36%	0.15	31	6.25%	51	10.28%	0.00	19	5.64%	49	14.54%	0.00
Asian	26	4.03%	34	5.27%	0.05	26	5.24%	31	6.25%	0.13	11	3.26%	32	9.50%	0.00
Non-Caucasian	68	10.54%	83	12.87%	0.03	48	9.68%	91	18.35%	0.00	32	9.50%	87	25.82%	0.00
Low Tenure	61	9.46%	76	11.78%	0.02	42	8.47%	68	13.71%	0.00	32	9.50%	66	19.58%	0.00
High Tenure	85	13.18%	87	13.49%	0.38	64	12.90%	113	22.78%	0.00	40	11.87%	96	28.49%	0.00

Table 74. E-4 Correlation Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Increase		Decrease			Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	45	6.98%	50	7.75%	0.20	35	7.06%	44	8.87%	0.05	22	6.53%	38	11.28%	0.00
Male	51	7.91%	40	6.20%	0.03	40	8.06%	32	6.45%	0.06	26	7.72%	32	9.50%	0.10
Female	23	3.57%	22	3.41%	0.36	18	3.63%	19	3.83%	0.35	12	3.56%	23	6.82%	0.00
Caucasian	27	4.19%	31	4.81%	0.19	25	5.04%	26	5.24%	0.37	15	4.45%	24	7.12%	0.01
African-American	22	3.41%	18	2.79%	0.14	25	5.04%	13	2.62%	0.00	17	5.04%	24	7.12%	0.04
Hispanic	26	4.03%	22	3.41%	0.16	19	3.83%	28	5.65%	0.02	7	2.08%	17	5.04%	0.00
Asian	14	2.17%	20	3.10%	0.05	13	2.62%	14	2.82%	0.32	11	3.26%	14	4.15%	0.14
Non-Caucasian	36	5.58%	37	5.74%	0.39	24	4.84%	28	5.65%	0.17	19	5.64%	25	7.42%	0.07
Low Tenure	25	3.88%	29	4.50%	0.18	26	5.24%	26	5.24%	0.45	11	3.26%	24	7.12%	0.00
High Tenure	37	5.74%	36	5.58%	0.39	33	6.65%	34	6.85%	0.38	19	5.64%	31	9.20%	0.00

Tables 75-77 present effect sizes of difficulty changes for the E-4 paygrade. As with the analysis utilizing all items, the vast majority of items exhibited negligible changes overall as well as for all demographic groups, and the proportion of items that had small or larger changes appeared to vary with group size. As the length of time between administrations increased, the proportion of easier items decreased and the proportion of harder items increased.

Table 75. E-4 Difficulty Change Effect Sizes- 2 Series Gap

	2 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	313	48.53%	291	22	0	0	332	51.47%	301	29	2	0
Male	295	45.74%	273	22	0	0	349	54.11%	316	31	2	0
Female	347	53.80%	287	59	1	0	297	46.05%	238	55	3	1
Caucasian	328	50.85%	286	42	0	0	316	48.99%	273	41	1	1
African-American	319	49.46%	267	52	0	0	326	50.54%	265	59	2	0
Hispanic	301	46.67%	236	65	0	0	343	53.18%	251	91	1	0
Asian	297	46.05%	188	106	3	0	344	53.33%	213	120	11	0
Non-Caucasian	319	49.46%	282	37	0	0	326	50.54%	279	45	2	0
Low Tenure	305	47.29%	259	46	0	0	340	52.71%	271	66	3	0
High Tenure	317	49.15%	284	33	0	0	328	50.85%	290	36	2	0

Table 76. E-4 Difficulty Change Effect Sizes- 3 Series Gap

	3 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	210	42.34%	182	27	1	0	286	57.66%	237	47	2	0
Male	204	41.13%	181	22	1	0	292	58.87%	234	56	2	0
Female	210	42.34%	155	53	2	0	285	57.46%	213	67	5	0
Caucasian	220	44.35%	181	38	1	0	273	55.04%	215	53	5	0
African-American	221	44.56%	177	43	1	0	274	55.24%	181	91	2	0
Hispanic	212	42.74%	152	57	3	0	282	56.85%	198	79	5	0
Asian	241	48.59%	150	84	7	0	253	51.01%	131	115	7	0
Non-Caucasian	213	42.94%	183	29	1	0	282	56.85%	221	58	3	0
Low Tenure	233	46.98%	192	39	2	0	263	53.02%	199	59	5	0
High Tenure	206	41.53%	176	29	1	0	290	58.47%	222	65	3	0

Table 77. E-4 Difficulty Change Effect Sizes- 4+ Series Gap

	4+ Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	130	38.58%	121	9	0	0	207	61.42%	163	40	4	0
Male	127	37.69%	117	10	0	0	210	62.31%	160	45	5	0
Female	137	40.65%	110	27	0	0	199	59.05%	133	62	4	0
Caucasian	138	40.95%	124	13	1	0	199	59.05%	153	43	3	0
African-American	134	39.76%	108	25	1	0	203	60.24%	130	70	3	0
Hispanic	135	40.06%	102	32	1	0	202	59.94%	127	70	4	1
Asian	142	42.14%	83	55	4	0	195	57.86%	101	83	10	1
Non-Caucasian	124	36.80%	107	17	0	0	212	62.91%	150	59	3	0
Low Tenure	129	38.28%	105	23	1	0	208	61.72%	153	51	4	0
High Tenure	120	35.61%	103	17	0	0	217	64.39%	165	49	3	0

Tables 78-80 present item-total correlation change effect sizes. Overall, the vast majority of items exhibited negligible changes, with a moderate proportion of small changes and few or no medium or larger changes. As group sizes decreased, changes appeared to be less stable, and the proportion of medium and large changes increased.

Table 78. E-4 Item-Total Correlation Change Effect Sizes- 2 Series Gap

	2 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	311	48.22%	262	49	0	0	334	51.78%	277	57	0	0
Male	316	48.99%	254	61	1	0	329	51.01%	246	83	0	0
Female	327	50.70%	210	110	7	0	306	47.44%	202	104	0	0
Caucasian	324	50.23%	219	102	3	0	318	49.30%	228	90	0	0
African-American	323	50.08%	201	119	3	0	314	48.68%	198	116	0	0
Hispanic	324	50.23%	169	136	19	0	304	47.13%	167	135	1	1
Asian	310	48.06%	123	130	47	10	302	46.82%	141	139	11	11
Non-Caucasian	312	48.37%	229	83	0	0	333	51.63%	239	94	0	0
Low Tenure	322	49.92%	197	115	10	0	314	48.68%	203	111	0	0
High Tenure	300	46.51%	241	58	1	0	345	53.49%	268	77	0	0

Table 79. E-4 Item-Total Correlation Change Effect Sizes- 3 Series Gap

	3 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	251	50.60%	196	55	0	0	244	49.19%	192	52	0	0
Male	247	49.80%	176	69	2	0	247	49.80%	183	64	0	0
Female	248	50.00%	164	76	8	0	245	49.40%	144	99	1	1
Caucasian	246	49.60%	159	80	7	0	244	49.19%	167	77	0	0
African-American	253	51.01%	145	95	13	0	237	47.78%	153	84	0	0
Hispanic	230	46.37%	100	105	23	2	245	49.40%	138	101	3	3
Asian	228	45.97%	68	102	52	6	222	44.76%	91	113	9	9
Non-Caucasian	247	49.80%	168	78	1	0	247	49.80%	167	80	0	0
Low Tenure	243	48.99%	146	93	4	0	242	48.79%	158	84	0	0
High Tenure	241	48.59%	179	59	3	0	251	50.60%	185	66	0	0

Table 80. E-4 Item-Total Correlation Change Effect Sizes- 4+ Series Gap

	4+ Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	150	44.51%	117	33	0	0	186	55.19%	141	45	0	0
Male	158	46.88%	120	38	0	0	178	52.82%	135	43	0	0
Female	161	47.77%	107	49	4	1	162	48.07%	94	68	0	0
Caucasian	148	43.92%	91	54	3	0	186	55.19%	129	57	0	0
African-American	164	48.66%	93	64	7	0	164	48.66%	94	70	0	0
Hispanic	156	46.29%	72	69	15	0	169	50.15%	97	68	2	2
Asian	155	45.99%	59	63	24	9	158	46.88%	55	81	11	11
Non-Caucasian	161	47.77%	109	52	0	0	174	51.63%	113	61	0	0
Low Tenure	157	46.59%	110	45	2	0	174	51.63%	112	62	0	0
High Tenure	157	46.59%	110	46	1	0	177	52.52%	131	46	0	0

5.1.2 Items Not Administered Prior to 2015

Table 81 presents the number of repeat items examined in Analysis 1 that were not administered prior to 2015 for the E-4 paygrade.

Table 81. E-4 Repeat Items

Series	Administrations Between Series	# of Repeat Items
227-232	2	40
227-235	3	27
227-236	4	15
228-235	2	34
228-236	3	33
228-239	4	29
231-236	2	20
231-239	3	46
231-240	4	38
232-239	2	44
232-240	3	42
235-240	2	80
227-239	5	15
227-240	6	12
228-240	5	8
-	2 Combined	218
-	3 Combined	148
-	4+ Combined	117

Table 82 presents the numbers of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-4 paygrade. For items with a 2-series gap between administrations, there was a non-significant difference in the number of items that became easier or harder overall and for all groups with the exception of Asian and Low-Tenure candidates. For items with a 3-series gap between administrations, there was a significantly greater number of items that became harder overall and for Male, African-American, Low-Tenure, and High-Tenure candidates. For items with a 4-series gap between administrations, there was a non-significant difference in the number of items that became easier or harder overall and for all groups with the exception of Caucasian candidates.

Table 83 presents the numbers of items for which item-total correlation changed significantly from the initial within-scope administration to the next administration within the E-4 paygrade. For items with a 2-series gap between administrations, there was a non-significant difference in the number of items that exhibited increased or decreased item-total correlations overall and for all groups with the exception of Male, Female, and High-Tenure candidates, with these groups exhibiting significantly more items that increased in item-total correlation. For items with a 3-series gap between administrations, there was a non-significant difference in the number of items that exhibited increased or decreased item-total correlations overall, though Male, Caucasian, Asian, Non-Caucasian, and Low-Tenure candidates exhibited significantly more items that increased in item-total correlation. For items with a 4-series or more gap between administrations, there was a non-significant difference in the number of items that exhibited increased or decreased item-total correlations overall and for all groups with the exception of Hispanics.

Table 82. E-4 Difficulty Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Easier		Harder			Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	46	21.10%	45	20.64%	0.40	24	16.22%	41	27.70%	0.00	27	23.08%	25	21.37%	0.28
Male	40	18.35%	38	17.43%	0.32	21	14.19%	35	23.65%	0.00	24	20.51%	25	21.37%	0.36
Female	32	14.68%	28	12.84%	0.18	17	11.49%	21	14.19%	0.12	20	17.09%	15	12.82%	0.07
Caucasian	33	15.14%	34	15.60%	0.38	16	10.81%	20	13.51%	0.12	22	18.80%	15	12.82%	0.02
African-American	23	10.55%	23	10.55%	0.44	10	6.76%	28	18.92%	0.00	10	8.55%	15	12.82%	0.04
Hispanic	26	11.93%	21	9.63%	0.11	17	11.49%	19	12.84%	0.25	10	8.55%	10	8.55%	0.42
Asian	18	8.26%	10	4.59%	0.01	15	10.14%	15	10.14%	0.43	9	7.69%	10	8.55%	0.29
Non-Caucasian	36	16.51%	36	16.51%	0.46	24	16.22%	30	20.27%	0.08	19	16.24%	20	17.09%	0.34
Low Tenure	26	11.93%	34	15.60%	0.04	21	14.19%	29	19.59%	0.03	20	17.09%	17	14.53%	0.18
High Tenure	39	17.89%	37	16.97%	0.32	22	14.86%	32	21.62%	0.01	23	19.66%	24	20.51%	0.36

Table 83. E-4 Correlation Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Increase		Decrease			Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	21	9.63%	15	6.88%	0.05	19	12.84%	14	9.46%	0.07	9	7.69%	10	8.55%	0.29
Male	18	8.26%	10	4.59%	0.01	19	12.84%	9	6.08%	0.00	11	9.40%	7	5.98%	0.05
Female	16	7.34%	8	3.67%	0.00	12	8.11%	9	6.08%	0.12	9	7.69%	8	6.84%	0.28
Caucasian	10	4.59%	7	3.21%	0.10	16	10.81%	8	5.41%	0.00	8	6.84%	8	6.84%	0.41
African-American	8	3.67%	11	5.05%	0.11	10	6.76%	7	4.73%	0.09	6	5.13%	4	3.42%	0.11
Hispanic	10	4.59%	11	5.05%	0.30	10	6.76%	8	5.41%	0.18	5	4.27%	9	7.69%	0.03
Asian	8	3.67%	7	3.21%	0.27	12	8.11%	6	4.05%	0.01	6	5.13%	5	4.27%	0.23
Non-Caucasian	18	8.26%	14	6.42%	0.11	17	11.49%	9	6.08%	0.00	6	5.13%	8	6.84%	0.15
Low Tenure	11	5.05%	10	4.59%	0.30	12	8.11%	4	2.70%	0.00	6	5.13%	7	5.98%	0.25
High Tenure	23	10.55%	12	5.50%	0.00	14	9.46%	12	8.11%	0.22	8	6.84%	8	6.84%	0.41

Tables 84-86 present effect sizes of difficulty changes for the E-4 paygrade. As with the analysis using all items, the vast majority of items exhibited negligible changes overall and for all groups. The proportion of items that had small or larger changes appeared to vary with group size. The proportion of easier items decreased from a 2-series gap to a 3-series gap, but was highest for a 4 or more series gap.

Table 84. E-4 Difficulty Change Effect Sizes- 2 Series Gap

	2 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	111	50.92%	99	11	1	0	107	49.08%	89	18	0	0
Male	106	48.62%	93	12	1	0	112	51.38%	94	18	0	0
Female	112	51.38%	90	21	1	0	106	48.62%	80	24	2	0
Caucasian	108	49.54%	96	11	1	0	109	50.00%	93	16	0	0
African-American	100	45.87%	80	20	0	0	118	54.13%	85	32	1	0
Hispanic	116	53.21%	89	26	0	1	101	46.33%	70	31	0	0
Asian	114	52.29%	67	42	4	1	103	47.25%	60	41	2	0
Non-Caucasian	107	49.08%	91	15	1	0	111	50.92%	91	20	0	0
Low Tenure	95	43.58%	77	17	1	0	123	56.42%	101	22	0	0
High Tenure	115	52.75%	97	17	1	0	98	44.95%	77	21	0	0

Table 85. E-4 Difficulty Change Effect Sizes- 3 Series Gap

	3 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	69	46.62%	61	8	0	0	79	53.38%	65	13	1	0
Male	69	46.62%	60	9	0	0	79	53.38%	61	17	1	0
Female	71	47.97%	57	14	0	0	77	52.03%	57	19	1	0
Caucasian	70	47.30%	58	12	0	0	78	52.70%	67	10	1	0
African-American	56	37.84%	45	11	0	0	90	60.81%	60	28	2	0
Hispanic	68	45.95%	45	23	0	0	80	54.05%	55	24	1	0
Asian	69	46.62%	41	26	1	1	79	53.38%	45	29	5	0
Non-Caucasian	63	42.57%	53	10	0	0	85	57.43%	68	16	1	0
Low Tenure	68	45.95%	53	14	1	0	80	54.05%	61	18	1	0
High Tenure	63	42.57%	52	11	0	0	85	57.43%	67	17	1	0

Table 86. E-4 Difficulty Change Effect Sizes- 4+ Series Gap

	4+ Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	61	52.14%	53	8	0	0	56	47.86%	47	9	0	0
Male	57	48.72%	51	6	0	0	60	51.28%	50	10	0	0
Female	57	48.72%	41	16	0	0	59	50.43%	44	14	1	0
Caucasian	58	49.57%	47	11	0	0	59	50.43%	49	10	0	0
African-American	52	44.44%	38	14	0	0	65	55.56%	46	18	1	0
Hispanic	58	49.57%	45	13	0	0	59	50.43%	45	13	1	0
Asian	63	53.85%	43	17	3	0	54	46.15%	29	25	0	0
Non-Caucasian	53	45.30%	45	8	0	0	64	54.70%	50	14	0	0
Low Tenure	66	56.41%	56	10	0	0	51	43.59%	39	12	0	0
High Tenure	53	45.30%	42	11	0	0	64	54.70%	51	12	1	0

Tables 87-89 present item-total correlation change effect sizes. Overall, the vast majority of items exhibited negligible changes, with a moderate proportion of small changes and no medium or larger changes. As group sizes decreased, changes appeared to be less stable, and the proportion of medium and large changes increased.

Table 87. E-4 Item-Total Correlation Change Effect Sizes- 2 Series Gap

	2 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	108	49.54%	96	12	0	0	110	50.46%	90	20	0	0
Male	108	49.54%	94	14	0	0	110	50.46%	84	26	0	0
Female	108	49.54%	70	33	4	1	107	49.08%	74	31	1	1
Caucasian	112	51.38%	93	18	1	0	104	47.71%	79	25	0	0
African-American	107	49.08%	71	34	2	0	109	50.00%	68	39	1	1
Hispanic	113	51.83%	72	33	7	1	97	44.50%	58	39	0	0
Asian	120	55.05%	52	48	16	4	90	41.28%	38	38	7	7
Non-Caucasian	101	46.33%	71	29	1	0	116	53.21%	85	31	0	0
Low Tenure	89	40.83%	73	16	0	0	127	58.26%	92	35	0	0
High Tenure	117	53.67%	90	26	1	0	100	45.87%	79	21	0	0

Table 88. E-4 Item-Total Correlation Change Effect Sizes- 3 Series Gap

	3 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	69	46.62%	54	15	0	0	79	53.38%	66	13	0	0
Male	73	49.32%	58	15	0	0	75	50.68%	62	13	0	0
Female	71	47.97%	39	30	2	0	71	47.97%	38	33	0	0
Caucasian	74	50.00%	53	20	1	0	73	49.32%	49	24	0	0
African-American	82	55.41%	42	38	2	0	64	43.24%	38	26	0	0
Hispanic	76	51.35%	46	24	5	1	66	44.59%	35	31	0	0
Asian	79	53.38%	25	30	21	3	58	39.19%	25	23	5	5
Non-Caucasian	73	49.32%	51	22	0	0	75	50.68%	57	18	0	0
Low Tenure	73	49.32%	50	20	3	0	75	50.68%	55	20	0	0
High Tenure	68	45.95%	53	15	0	0	79	53.38%	57	22	0	0

Table 89. E-4 Item-Total Correlation Change Effect Sizes- 4+ Series Gap

	4+ Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	55	47.01%	43	12	0	0	62	52.99%	50	12	0	0
Male	52	44.44%	40	12	0	0	65	55.56%	52	13	0	0
Female	56	47.86%	34	16	5	1	57	48.72%	36	19	1	1
Caucasian	53	45.30%	34	18	1	0	62	52.99%	47	15	0	0
African-American	57	48.72%	35	18	4	0	60	51.28%	35	25	0	0
Hispanic	55	47.01%	31	21	3	0	56	47.86%	37	19	0	0
Asian	57	48.72%	28	24	5	0	50	42.74%	28	18	2	2
Non-Caucasian	57	48.72%	41	16	0	0	59	50.43%	37	22	0	0
Low Tenure	54	46.15%	37	17	0	0	63	53.85%	41	22	0	0
High Tenure	58	49.57%	41	17	0	0	57	48.72%	41	16	0	0

Summary

- Most items did not exhibit significant changes in difficulty regardless of the length of time between administrations. As the length of time between administrations increased, the percentage of items that became significantly harder increased, with significantly greater numbers of harder items after 3-series and 4 or more series gaps. This pattern reflected items administered prior to 2015, though for items not administered prior to 2015, there was a non-significant difference in the number of items that became easier or harder after a 4 or more series gap.
- Most items did not exhibit significant item-total correlation changes regardless of the length of time between administrations. For items administered prior to 2015, the proportion of items with decreased item-total correlations increased as the length of time between administrations increased, but the pattern was less clear for items administered after 2015.

5.2 Analysis 2: Item Parameter Changes for Repeat Test-Takers

Table 90 presents the number of repeat test-takers for each possible series pair for the E-4 paygrade. Note that within these tables, candidates can be counted in multiple series pairs if they took the exam more than two exams within the rating in the scope of the study; however, only their initial and second viewing of an item is included in the analyses in this section.

Table 91 presents the number of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-4 paygrade. For items with a 2-series, 3-series, or 4-series or more gap between administrations, there was a significantly greater number of items that became easier overall and for all demographic groups. The percentage of items that became harder was relatively consistent across the different lengths of time between administrations, but the percentage of items that became easier decreased as the length of time between administrations increased overall and for all demographic groups.

Table 92 presents the numbers of items for which item-total correlations changed significantly from the initial within-scope administration to the next administration within the E-4 paygrade. For items with a 2-series, 3-series, or 4-series or more gap between administrations, there was a significantly greater number of items for which the item-total correlation increased overall and for all demographic groups. The percentages of items for which item-total correlations increased and decreased were relatively consistent across the different lengths of time between administrations overall and within demographic groups.

Table 90. E-4 Repeat Test-Taker Sample Sizes

Series	Admins. Between Series	Overall	Male	Female	Caucasian	African-American	Hispanic	Asian	Non-Caucasian	Low Tenure	High Tenure
227-232	2	3,047	2,355	692	1,251	584	545	274	1,483	1,018	2,029
227-235	3	1,769	1,372	397	712	352	324	156	878	676	1,093
227-236	4	977	777	200	408	200	169	83	478	448	529
228-235	2	2,820	2,189	631	1,149	576	508	245	1,404	1,525	1,295
228-236	3	1,693	1,341	352	725	335	294	143	818	1,031	662
228-239	4	1,053	848	205	458	188	184	101	498	687	366
231-236	2	2,389	1,872	517	982	498	433	208	1,204	1,579	810
231-239	3	1,524	1,217	307	641	309	271	135	750	1,062	462
231-240	4	801	625	176	315	191	136	63	411	560	241
232-239	2	2,364	1,793	571	1,015	450	448	210	1,156	1,691	673
232-240	3	1,283	955	328	529	283	233	107	654	927	356
235-240	2	1,940	1,389	551	827	401	363	159	959	1,384	556
227-239	5	557	454	103	228	111	101	53	277	265	292
227-240	6	266	206	60	104	68	42	24	139	124	142
228-240	5	524	410	114	215	116	86	41	255	333	191
-	2 Combined	12,560	9,598	2,962	5,224	2,509	2,297	1,096	6,206	7,197	5,363
-	3 Combined	6,269	4,885	1,384	2,607	1,279	1,122	541	3,100	3,696	2,573
-	4+ Combined	4,178	3,320	858	1,728	874	718	365	2,058	2,417	1,761

Table 91. E-4 Difficulty Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Easier		Harder			Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	263	30.44%	21	2.43%	0.00	129	20.03%	30	4.66%	0.00	56	12.33%	20	4.41%	0.00
Male	227	26.27%	22	2.55%	0.00	116	18.01%	24	3.73%	0.00	53	11.67%	15	3.30%	0.00
Female	143	16.55%	21	2.43%	0.00	80	12.42%	9	1.40%	0.00	23	5.07%	7	1.54%	0.00
Caucasian	161	18.63%	18	2.08%	0.00	70	10.87%	21	3.26%	0.00	36	7.93%	16	3.52%	0.00
African-American	171	19.79%	15	1.74%	0.00	78	12.11%	12	1.86%	0.00	35	7.71%	7	1.54%	0.00
Hispanic	106	12.27%	9	1.04%	0.00	48	7.45%	10	1.55%	0.00	27	5.95%	3	0.66%	0.00
Asian	73	8.45%	5	0.58%	0.00	35	5.43%	4	0.62%	0.00	12	2.64%	4	0.88%	0.00
Non-Caucasian	204	23.61%	19	2.20%	0.00	107	16.61%	15	2.33%	0.00	51	11.23%	8	1.76%	0.00
Low Tenure	223	25.81%	22	2.55%	0.00	108	16.77%	10	1.55%	0.00	48	10.57%	14	3.08%	0.00
High Tenure	157	18.17%	21	2.43%	0.00	74	11.49%	4	0.62%	0.00	31	6.83%	6	1.32%	0.00

Table 92. E-4 Item-Total Correlation Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Increase		Decrease			Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	144	16.67%	10	1.16%	0.00	104	16.15%	5	0.78%	0.00	75	16.52%	10	2.20%	0.00
Male	121	14.00%	9	1.04%	0.00	72	11.18%	4	0.62%	0.00	63	13.88%	4	0.88%	0.00
Female	42	4.86%	5	0.58%	0.00	28	4.35%	7	1.09%	0.00	9	1.98%	5	1.10%	0.03
Caucasian	76	8.80%	7	0.81%	0.00	52	8.07%	4	0.62%	0.00	26	5.73%	3	0.66%	0.00
African-American	49	5.67%	7	0.81%	0.00	31	4.81%	1	0.16%	0.00	13	2.86%	8	1.76%	0.03
Hispanic	57	6.60%	5	0.58%	0.00	30	4.66%	8	1.24%	0.00	24	5.29%	0	0.00%	0.00
Asian	16	1.85%	5	0.58%	0.00	23	3.57%	2	0.31%	0.00	14	3.08%	3	0.66%	0.00
Non-Caucasian	83	9.61%	7	0.81%	0.00	5	0.78%	1	0.16%	0.00	6	1.32%	0	0.00%	0.00
Low Tenure	64	7.41%	10	1.16%	0.00	48	7.45%	3	0.47%	0.00	29	6.39%	1	0.22%	0.00
High Tenure	65	7.52%	4	0.46%	0.00	31	4.81%	1	0.16%	0.00	20	4.41%	0	0.00%	0.00

Summary

- The percentage of items that became easier for repeat test-takers was significantly greater than the percentage of items that became harder. The percentage of items that became harder was relatively consistent across the different lengths of time between administrations, but the percentage of items that became easier decreased as the length of time between administrations increased.
- The percentage of items that exhibited an increased item-total correlation for repeat test-takers was significantly greater than the percentage of items that exhibited a decreased item-total correlation.

5.3 Analysis 3: Candidate Performance Differences for Initial vs. Repeat Exposures

Table 93 presents the number of candidates who performed better or worse on repeat items (i.e., items they had seen in prior administrations) vs. non-repeat items (i.e., items they had not seen in prior administrations) for candidates at the E-4 paygrade. A significantly greater number of candidates performed better on repeat items than on non-repeat items overall and in all demographic groups.

Table 93. E-4 Test-Taker Repeat Performance

	Repeat N	Sig. Better N	Sig. Better %	Sig. Worse N	Sig. Worse %	*p*
Overall	6,895	1,209	17.53%	252	3.65%	0.00
Male	5,161	915	17.73%	212	4.11%	0.00
Female	1,734	294	16.96%	40	2.31%	0.00
Caucasian	2,889	471	16.30%	105	3.63%	0.00
African-American	1,358	268	19.73%	51	3.76%	0.00
Hispanic	1,266	243	19.19%	45	3.55%	0.00
Asian	588	117	19.90%	20	3.40%	0.00
Non-Caucasian	3,380	658	19.47%	123	3.64%	0.00
Low Tenure	3,861	717	18.57%	121	3.13%	0.00
High Tenure	3,034	492	16.22%	131	4.32%	0.00

Summary

Most candidates at the E-4 paygrade do not perform significantly better or worse on repeat items compared to non-repeat items. Of those that do perform better or worse, a significantly greater number of candidates performed better on repeat items than on non-repeat items overall and in all demographic groups.

6.0 E-5 OVERALL RESULTS

The results in this section are for the E-5 paygrade. Table 94 presents total and demographic group sizes for each administration. The sample was predominately of High- Tenure, Male, and Caucasian.

Table 94. E-5 Sample Sizes by Administration

	Administration							
	227	**228**	**231**	**232**	**235**	**236**	**239**	**240**
Overall	11,339	11,241	11,079	11,716	12,069	11,110	11,818	10,473
Male	9,215	9,101	8,933	9,353	9,616	8,771	9,263	8,210
Female	2,124	2,140	2,146	2,363	2,453	2,339	2,555	2,263
Caucasian	5,255	5,350	5,352	5,738	5,940	5,535	5,764	5,179
African-American	1,388	1,418	1,411	1,540	1,627	1,474	1,611	1,422
Hispanic	2,043	1,925	1,896	1,985	2,014	1,858	2,058	1,816
Asian	756	743	763	835	886	856	920	815
Non-Caucasian	4,689	4,571	4,484	4,764	4,895	4,491	4,890	4,304
Low Tenure	5,142	5,246	5,202	5,645	5,582	5,075	5,334	4,449
High Tenure	6,197	5,995	5,877	6,071	6,487	6,035	6,484	6,024

6.1 Analysis 1: Item Parameter Changes Over Time

Table 95 presents the number of repeat items examined in Analysis 1 for the E-5 paygrade. As was noted, results are presented for series with 2, 3, or 4 or more administrations between the initial and subsequent administration.

Table 95. E-5 Repeat Items

Series	Administrations Between Series	# of Repeat Items
227-232	2	193
227-235	3	175
227-236	4	146
228-235	2	210
228-236	3	160
228-239	4	103
231-236	2	192
231-239	3	179
231-240	4	131
232-239	2	127
232-240	3	169
235-240	2	92
227-239	5	70
227-240	6	49
228-240	5	74
-	2 Combined	814
-	3 Combined	683
-	4+ Combined	573

Table 96 presents the number of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-5 paygrade. For items with a 2-series or 3-series gap between administrations, there was a non-significant difference between the number of items that became easier and the number of items that became harder overall. However, the number of items that became easier was significantly greater than the number of items that became harder for Hispanic, Asian, Non-Caucasian, and High-Tenure candidates in the 2-series gap results, as well as for Hispanic, Asian, and High-Tenure candidates in the 3-series gap results. Additionally, for Low-Tenure candidates, the number of items that became harder was significantly greater than the number of items that became easier after a 2-series or 3-series gap. For items with a 4-series or larger gap, there was a significantly greater number of items that became harder than items that became easier. This effect holds for all demographic groups except Caucasians and Asians. Looking across the sets of results, the change from a non-significant difference to a preponderance of harder items appears to be primarily driven by an increase in the proportion of harder items.

Table 97 presents the numbers of items for which item-total correlation changed significantly from the initial within-scope administration to the next administration within the E-5 paygrade. For items with a 2-series or 3-series gap between administrations, there was a significantly greater number of items with an increased item-total correlation than items with a decreased correlation. This was also true for all demographic groups after a 2-series gap except African-American, Asian, and Low-Tenure candidates, as well as African-American and Asian candidates after a 3-series gap. For items with a 4-series or larger gap between administrations, there was a non-significant difference between the number of items with an increased or decreased item-total correlation overall, though Female, Asian, and Low-Tenure candidates exhibited a significantly greater number of items with a decreased item-total correlation.

Table 96. E-5 Difficulty Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Easier		Harder			Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	180	22.11%	171	21.01%	0.21	142	20.79%	153	22.40%	0.14	129	22.51%	169	29.49%	0.00
Male	164	20.15%	152	18.67%	0.13	133	19.47%	139	20.35%	0.26	122	21.29%	145	25.31%	0.01
Female	77	9.46%	89	10.93%	0.07	71	10.40%	76	11.13%	0.24	63	10.99%	96	16.75%	0.00
Caucasian	135	16.58%	121	14.86%	0.08	126	18.45%	125	18.30%	0.44	112	19.55%	126	21.99%	0.06
African-American	68	8.35%	70	8.60%	0.37	63	9.22%	54	7.91%	0.09	45	7.85%	64	11.17%	0.00
Hispanic	88	10.81%	74	9.09%	0.04	77	11.27%	52	7.61%	0.00	62	10.82%	79	13.79%	0.01
Asian	61	7.49%	48	5.90%	0.03	53	7.76%	32	4.69%	0.00	43	7.50%	46	8.03%	0.28
Non-Caucasian	127	15.60%	106	13.02%	0.01	98	14.35%	89	13.03%	0.14	91	15.88%	126	21.99%	0.00
Low Tenure	129	15.85%	148	18.18%	0.03	111	16.25%	132	19.33%	0.01	89	15.53%	135	23.56%	0.00
High Tenure	169	20.76%	124	15.23%	0.00	123	18.01%	105	15.37%	0.03	108	18.85%	125	21.82%	0.03

Table 97. E-5 Correlation Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Increase		Decrease			Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	92	11.30%	56	6.88%	0.00	85	12.45%	53	7.76%	0.00	56	9.77%	61	10.65%	0.22
Male	103	12.65%	68	8.35%	0.00	84	12.30%	49	7.17%	0.00	54	9.42%	54	9.42%	0.46
Female	37	4.55%	28	3.44%	0.04	33	4.83%	21	3.07%	0.00	17	2.97%	37	6.46%	0.00
Caucasian	78	9.58%	65	7.99%	0.04	64	9.37%	40	5.86%	0.00	46	8.03%	53	9.25%	0.13
African-American	27	3.32%	22	2.70%	0.12	19	2.78%	21	3.07%	0.27	16	2.79%	21	3.66%	0.09
Hispanic	42	5.16%	29	3.56%	0.01	44	6.44%	26	3.81%	0.00	27	4.71%	23	4.01%	0.17
Asian	33	4.05%	36	4.42%	0.26	29	4.25%	22	3.22%	0.06	19	3.32%	27	4.71%	0.03
Non-Caucasian	69	8.48%	39	4.79%	0.00	55	8.05%	33	4.83%	0.00	30	5.24%	33	5.76%	0.25
Low Tenure	71	8.72%	68	8.35%	0.32	68	9.96%	41	6.00%	0.00	39	6.81%	50	8.73%	0.03
High Tenure	74	9.09%	42	5.16%	0.00	59	8.64%	35	5.12%	0.00	39	6.81%	41	7.16%	0.33

Tables 98-100 present effect sizes of difficulty changes for the E-5 paygrade. For administrations with a 2-series gap, the vast majority of items exhibited negligible changes overall and for all groups. The proportion of items that had small or larger changes appeared to vary with group size. For example, the proportion of items exhibiting at least a small change (both easier and harder) was larger for smaller groups, such as Asians and Hispanics, than for larger groups (e.g., Overall or for Male candidates). As the length of time between administrations increased, the proportion of easier items decreased and the proportion of harder items increased. Though the vast majority of items exhibited negligible changes regardless of length of time between administrations, the proportion of items exhibiting small or larger differences generally increased as the length of time between administrations increased.

Table 98. E-5 Difficulty Change Effect Sizes - 2 Series Gap

	2 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	424	52.09%	394	28	2	0	390	47.91%	351	37	2	0
Male	425	52.21%	387	36	2	0	389	47.79%	348	39	2	0
Female	425	52.21%	361	61	3	0	389	47.79%	317	68	4	0
Caucasian	430	52.83%	384	43	3	0	384	47.17%	341	41	2	0
African-American	403	49.51%	326	74	3	0	409	50.25%	298	102	9	0
Hispanic	425	52.21%	332	89	4	0	389	47.79%	316	70	3	0
Asian	421	51.72%	291	123	6	1	388	47.67%	275	97	12	4
Non-Caucasian	416	51.11%	372	43	1	0	398	48.89%	358	38	2	0
Low Tenure	388	47.67%	334	52	2	0	426	52.33%	367	57	2	0
High Tenure	430	52.83%	374	54	2	0	384	47.17%	339	43	2	0

Table 99. E-5 Difficulty Change Effect Sizes - 3 Series Gap

	3 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	325	47.58%	291	31	3	0	358	52.42%	315	41	2	0
Male	329	48.17%	289	37	3	0	352	51.54%	300	50	2	0
Female	327	47.88%	266	58	3	0	356	52.12%	289	64	2	1
Caucasian	326	47.73%	275	48	3	0	357	52.27%	296	59	1	1
African-American	326	47.73%	235	87	4	0	354	51.83%	231	107	16	0
Hispanic	338	49.49%	257	75	6	0	345	50.51%	261	78	6	0
Asian	341	49.93%	199	118	20	4	324	47.44%	203	103	15	3
Non-Caucasian	332	48.61%	286	44	2	0	351	51.39%	295	54	2	0
Low Tenure	305	44.66%	261	43	1	0	378	55.34%	314	60	3	1
High Tenure	329	48.17%	290	36	3	0	354	51.83%	301	51	2	0

Table 100. E-5 Difficulty Change Effect Sizes - 4+ Series Gap

	4+ Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	245	42.76%	217	27	1	0	328	57.24%	268	57	2	1
Male	244	42.58%	216	27	1	0	329	57.42%	270	57	1	1
Female	250	43.63%	203	45	2	0	321	56.02%	245	72	3	1
Caucasian	260	45.38%	219	39	2	0	313	54.62%	251	59	2	1
African-American	254	44.33%	196	55	3	0	316	55.15%	205	99	11	1
Hispanic	252	43.98%	193	58	1	0	321	56.02%	229	86	5	1
Asian	272	47.47%	168	95	7	2	295	51.48%	178	101	13	3
Non-Caucasian	252	43.98%	217	34	1	0	320	55.85%	241	75	3	1
Low Tenure	235	41.01%	201	32	2	0	338	58.99%	259	74	4	1
High Tenure	267	46.60%	228	38	1	0	306	53.40%	246	58	1	1

Tables 101-103 present item-total correlation change effect sizes. Overall, the proportion of items that exhibited decreases was higher after a 4-series gap than other lengths of time. Overall, the vast majority of items exhibited negligible changes, with a moderate proportion of small changes and few or no medium or larger changes. As group sizes decreased, changes appeared to be less stable, and the proportion of medium and large changes increased.

Table 101. E-5 Item-Total Correlation Change Effect Sizes - 2 Series Gap

	2 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	418	51.35%	367	50	1	0	395	48.53%	358	37	0	0
Male	426	52.33%	342	64	4	16	393	48.28%	320	45	14	14
Female	408	50.12%	270	135	3	0	399	49.02%	278	121	0	0
Caucasian	415	50.98%	324	70	3	18	408	50.12%	294	82	16	16
African-American	430	52.83%	268	141	17	4	369	45.33%	241	122	3	3
Hispanic	439	53.93%	297	131	10	1	368	45.21%	241	117	5	5
Asian	430	52.83%	185	202	34	9	366	44.96%	156	164	23	23
Non-Caucasian	424	52.09%	328	91	5	0	382	46.93%	304	78	0	0
Low Tenure	411	50.49%	326	64	6	15	410	50.37%	295	83	16	16
High Tenure	439	53.93%	349	86	4	0	374	45.95%	310	64	0	0

Table 102. E-5 Item-Total Correlation Change Effect Sizes - 3 Series Gap

	3 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	368	53.88%	303	61	4	0	315	46.12%	265	50	0	0
Male	384	56.22%	305	67	3	9	299	43.78%	249	50	0	0
Female	363	53.15%	235	114	12	2	307	44.95%	194	111	1	1
Caucasian	350	51.24%	273	63	6	8	330	48.32%	258	72	0	0
African-American	348	50.95%	187	123	31	7	308	45.10%	181	117	5	5
Hispanic	366	53.59%	199	127	28	12	310	45.39%	186	108	8	8
Asian	380	55.64%	152	170	32	26	289	42.31%	120	113	28	28
Non-Caucasian	366	53.59%	268	84	12	2	313	45.83%	241	72	0	0
Low Tenure	343	50.22%	257	70	8	8	339	49.63%	264	75	0	0
High Tenure	382	55.93%	290	88	4	0	297	43.48%	229	68	0	0

Table 103. E-5 Item-Total Correlation Change Effect Sizes - 4+ Series Gap

	4+ Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	283	49.39%	242	41	0	0	289	50.44%	241	48	0	0
Male	281	49.04%	233	47	1	0	291	50.79%	244	47	0	0
Female	279	48.69%	194	82	3	0	278	48.52%	174	104	0	0
Caucasian	291	50.79%	228	61	2	0	279	48.69%	203	76	0	0
African-American	284	49.56%	147	115	18	4	282	49.21%	156	118	4	4
Hispanic	275	47.99%	155	107	10	3	292	50.96%	198	90	2	2
Asian	275	47.99%	138	102	19	16	271	47.29%	118	129	12	12
Non-Caucasian	263	45.90%	190	69	4	0	306	53.40%	238	68	0	0
Low Tenure	281	49.04%	217	64	0	0	289	50.44%	220	69	0	0
High Tenure	283	49.39%	228	53	2	0	290	50.61%	223	67	0	0

6.1.1 Items Administered Prior to 2015

Table 104 presents the number of repeat items examined in Analysis 1 that were administered prior to 2015 for the E-5 paygrade.

Table 104. E-5 Repeat Items

Series	Administrations Between Series	# of Repeat Items
227-232	2	139
227-235	3	149
227-236	4	107
228-235	2	181
228-236	3	132
228-239	4	79
231-236	2	174
231-239	3	155
231-240	4	103
232-239	2	94
232-240	3	126
235-240	2	44
227-239	5	58
227-240	6	40
228-240	5	57
-	2 Combined	632
-	3 Combined	562
-	4+ Combined	444

Table 105 presents the numbers of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-5 paygrade. For items with a 2-series gap between administrations, there was a non-significant difference overall. However, there was a significantly greater number of items that became harder among Females and a significantly greater number of items that became easier among High-Tenure candidates. For items with a 3-series gap between administrations, there was a significantly greater number of items that became harder than items that became easier overall and for Low-Tenure candidates, whereas Hispanic and Asian candidates exhibited a significantly greater number of items that became easier. For items with a 4-series gap between administrations, there was a significantly greater number of items that became harder than items that became easier overall and for all demographic groups except Caucasians and Asians. Again, this appears to be primarily driven by an increase in the proportion of harder items as the gap between administrations increases.

Table 106 presents the numbers of items for which item-total correlation changed significantly from the initial within-scope administration to the next administration within the E-5 paygrade. For items with a 2-series or 3-series gap between administrations, there was a significantly greater number of items with an increased item-total correlation than the number of items with a decreased correlation. This was also true for Male, Non-Caucasian, and High-Tenure candidates for the 2-series gap and for all demographic groups for the 3-series gap except African-

Table 105. E-5 Difficulty Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Easier		Harder			Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	130	20.57%	133	21.04%	0.36	108	19.22%	125	22.24%	0.03	93	20.95%	130	29.28%	0.00
Male	114	18.04%	114	18.04%	0.47	102	18.15%	111	19.75%	0.15	87	19.59%	113	25.45%	0.00
Female	49	7.75%	64	10.13%	0.01	54	9.61%	60	10.68%	0.18	48	10.81%	69	15.54%	0.00
Caucasian	101	15.98%	94	14.87%	0.20	98	17.44%	98	17.44%	0.47	83	18.69%	95	21.40%	0.07
African-American	42	6.65%	48	7.59%	0.15	42	7.47%	44	7.83%	0.34	30	6.76%	45	10.14%	0.00
Hispanic	57	9.02%	53	8.39%	0.26	59	10.50%	40	7.12%	0.00	49	11.04%	62	13.96%	0.02
Asian	37	5.85%	34	5.38%	0.26	38	6.76%	29	5.16%	0.04	31	6.98%	33	7.43%	0.31
Non-Caucasian	84	13.29%	78	12.34%	0.21	72	12.81%	70	12.46%	0.37	64	14.41%	94	21.17%	0.00
Low Tenure	97	15.35%	111	17.56%	0.06	84	14.95%	104	18.51%	0.01	70	15.77%	105	23.65%	0.00
High Tenure	113	17.88%	94	14.87%	0.02	92	16.37%	88	15.66%	0.30	75	16.89%	94	21.17%	0.01

Table 106. E-5 Correlation Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Increase		Decrease			Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	62	9.81%	38	6.01%	0.00	56	9.96%	41	7.30%	0.01	37	8.33%	48	10.81%	0.03
Male	66	10.44%	50	7.91%	0.01	55	9.79%	37	6.58%	0.00	36	8.11%	42	9.46%	0.13
Female	22	3.48%	22	3.48%	0.44	27	4.80%	18	3.20%	0.02	10	2.25%	30	6.76%	0.00
Caucasian	52	8.23%	51	8.07%	0.41	44	7.83%	32	5.69%	0.01	34	7.66%	44	9.91%	0.03
African-American	19	3.01%	16	2.53%	0.19	11	1.96%	16	2.85%	0.05	9	2.03%	17	3.83%	0.00
Hispanic	24	3.80%	22	3.48%	0.29	30	5.34%	21	3.74%	0.02	16	3.60%	17	3.83%	0.34
Asian	24	3.80%	27	4.27%	0.23	22	3.91%	18	3.20%	0.14	13	2.93%	23	5.18%	0.00
Non-Caucasian	46	7.28%	29	4.59%	0.00	31	5.52%	25	4.45%	0.10	15	3.38%	25	5.63%	0.01
Low Tenure	47	7.44%	53	8.39%	0.16	46	8.19%	31	5.52%	0.00	29	6.53%	38	8.56%	0.04
High Tenure	49	7.75%	30	4.75%	0.00	41	7.30%	29	5.16%	0.01	26	5.86%	31	6.98%	0.13

American, Asian, and Non-Caucasian candidates. For items with a 4-series or more gap between administrations, there was a significantly greater number of items with a decreased item-total correlation than items with an increased correlation overall and for all demographic groups except Males, Hispanics, and High-Tenure candidates

Tables 107-109 present effect sizes of difficulty changes for the E-5 paygrade. As with the analysis using all items, the vast majority of items exhibited negligible changes overall and for all groups, and the proportion of items that had small or larger changes appeared to vary with group size. As the length of time between administrations increased, the proportion of easier items decreased, and the proportion of harder items increased.

Table 107. E-5 Difficulty Change Effect Sizes - 2 Series Gap

	2 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	321	50.79%	298	22	1	0	311	49.21%	282	27	2	0
Male	321	50.79%	293	27	1	0	311	49.21%	280	29	2	0
Female	324	51.27%	276	45	3	0	308	48.73%	253	53	2	0
Caucasian	329	52.06%	291	36	2	0	303	47.94%	271	30	2	0
African-American	304	48.10%	250	51	3	0	326	51.58%	238	81	7	0
Hispanic	323	51.11%	251	70	2	0	309	48.89%	250	56	3	0
Asian	330	52.22%	225	100	5	0	297	46.99%	211	77	7	2
Non-Caucasian	312	49.37%	285	26	1	0	320	50.63%	290	28	2	0
Low Tenure	303	47.94%	262	40	1	0	329	52.06%	283	44	2	0
High Tenure	325	51.42%	285	39	1	0	307	48.58%	272	33	2	0

Table 108. E-5 Difficulty Change Effect Sizes - 3 Series Gap

	3 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	268	47.69%	245	20	3	0	294	52.31%	254	38	2	0
Male	270	48.04%	243	24	3	0	290	51.60%	244	44	2	0
Female	261	46.44%	214	44	3	0	301	53.56%	240	58	2	1
Caucasian	268	47.69%	231	34	3	0	294	52.31%	242	50	1	1
African-American	261	46.44%	192	65	4	0	299	53.20%	189	94	16	0
Hispanic	282	50.18%	223	54	5	0	280	49.82%	209	65	6	0
Asian	274	48.75%	162	92	16	4	272	48.40%	168	89	12	3
Non-Caucasian	267	47.51%	235	30	2	0	295	52.49%	246	47	2	0
Low Tenure	257	45.73%	223	33	1	0	305	54.27%	250	51	3	1
High Tenure	267	47.51%	240	24	3	0	295	52.49%	246	47	2	0

Table 109. E-5 Difficulty Change Effect Sizes - 4+ Series Gap

	4+ Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	188	42.34%	166	21	1	0	256	57.66%	209	44	2	1
Male	186	41.89%	167	18	1	0	258	58.11%	211	45	1	1
Female	192	43.24%	151	39	2	0	250	56.31%	196	50	3	1
Caucasian	198	44.59%	169	27	2	0	246	55.41%	197	46	2	1
African-American	192	43.24%	146	44	2	0	249	56.08%	166	75	7	1
Hispanic	189	42.57%	140	48	1	0	255	57.43%	179	71	4	1
Asian	213	47.97%	131	74	6	2	225	50.68%	138	74	11	2
Non-Caucasian	188	42.34%	161	26	1	0	255	57.43%	197	54	3	1
Low Tenure	182	40.99%	155	25	2	0	262	59.01%	197	60	4	1
High Tenure	204	45.95%	174	29	1	0	240	54.05%	198	40	1	1

Tables 110-112 present item-total correlation change effect sizes. Overall, the vast majority of items exhibited negligible changes, with a moderate proportion of small changes and few or no medium or larger changes. As group sizes decreased, changes appeared to be less stable, and the proportion of medium and large changes increased.

Table 110. E-5 Item-Total Correlation Change Effect Sizes - 2 Series Gap

	2 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	324	51.27%	287	36	1	0	307	48.58%	280	27	0	0
Male	332	52.53%	273	50	2	7	307	48.58%	246	35	13	13
Female	310	49.05%	197	110	3	0	317	50.16%	218	99	0	0
Caucasian	320	50.63%	250	61	1	8	321	50.79%	227	66	14	14
African-American	331	52.37%	203	113	12	3	288	45.57%	180	104	2	2
Hispanic	328	51.90%	225	93	9	1	301	47.63%	188	103	5	5
Asian	333	52.69%	135	164	27	7	286	45.25%	121	133	16	16
Non-Caucasian	320	50.63%	243	74	3	0	306	48.42%	239	67	0	0
Low Tenure	318	50.32%	255	54	2	7	322	50.95%	227	67	14	14
High Tenure	331	52.37%	263	65	3	0	300	47.47%	251	49	0	0

Table 111. E-5 Item-Total Correlation Change Effect Sizes - 3 Series Gap

	3 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	295	52.49%	247	44	4	0	267	47.51%	225	42	0	0
Male	311	55.34%	253	49	3	6	251	44.66%	211	40	0	0
Female	286	50.89%	177	97	10	2	264	46.98%	166	96	1	1
Caucasian	275	48.93%	217	47	6	5	284	50.53%	228	56	0	0
African-American	277	49.29%	150	95	28	4	262	46.62%	152	100	5	5
Hispanic	296	52.67%	156	104	27	9	261	46.44%	151	98	6	6
Asian	313	55.69%	126	138	26	23	240	42.70%	98	94	24	24
Non-Caucasian	294	52.31%	218	64	11	1	265	47.15%	206	59	0	0
Low Tenure	270	48.04%	208	50	7	5	291	51.78%	232	59	0	0
High Tenure	314	55.87%	241	69	4	0	244	43.42%	186	58	0	0

Table 112. E-5 Item-Total Correlation Change Effect Sizes - 4+ Series Gap

	4+ Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	212	47.75%	181	31	0	0	231	52.03%	193	38	0	0
Male	211	47.52%	173	37	1	0	232	52.25%	196	36	0	0
Female	206	46.40%	144	59	3	0	223	50.23%	135	88	0	0
Caucasian	215	48.42%	169	44	2	0	226	50.90%	167	59	0	0
African-American	218	49.10%	123	81	11	3	220	49.55%	113	99	4	4
Hispanic	207	46.62%	115	81	8	3	230	51.80%	157	71	1	1
Asian	211	47.52%	105	85	10	11	215	48.42%	83	110	11	11
Non-Caucasian	196	44.14%	144	50	2	0	244	54.95%	188	56	0	0
Low Tenure	214	48.20%	164	50	0	0	228	51.35%	173	55	0	0
High Tenure	210	47.30%	172	36	2	0	234	52.70%	179	55	0	0

6.1.2 Items Not Administered Prior to 2015

Table 113 presents the number of repeat items examined in Analysis 1 that were not administered prior to 2015 for the E-5 paygrade.

Table 113. E-5 Repeat Items

Series	Administrations Between Series	# of Repeat Items
227-232	2	54
227-235	3	26
227-236	4	39
228-235	2	29
228-236	3	28
228-239	4	24
231-236	2	18
231-239	3	24
231-240	4	28
232-239	2	33
232-240	3	43
235-240	2	48
227-239	5	12
227-240	6	9
228-240	5	17
-	2 Combined	182
-	3 Combined	121
-	4+ Combined	129

Table 114 presents the number of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-5 paygrade. For items with a 2-series gap between administrations, there was a significantly greater number of items that became easier than items that became harder overall as well as for all demographic groups with the exception of Female, Caucasian, African-American, and Low-Tenure candidates. For items with a 3-series gap between administrations, there was a non-significant difference in the number of items that became easier and harder overall, but a significantly greater number of items became easier for African-American, Hispanic, Asian, Non-Caucasian, and High-Tenure candidates. For items with a 4-series gap between administrations, there was a non-significant difference in the number of items that became easier or harder overall and for all demographic groups with the exception of Female and Low-Tenure candidates, with both groups exhibiting a significantly greater number of items that became harder.

Table 115 presents the number of items for which item-total correlation changed significantly from the initial within-scope administration to the next administration within the E-5 paygrade. For items with a 2-series gap between administrations, a significantly greater number of items exhibited increased item-total correlations than exhibited decreased item-total correlations overall as well as for all demographic groups with the exception of African-Americans and Asians. For items with a 3-series gap between administrations, a significantly greater number of items exhibited increased item-total correlations than decreased item-total correlations overall and for all groups with the exception of African-Americans and Asians. For items with a 4-series or more gap between administrations, a significantly greater number of items exhibited increased item-total correlations than decreased item-total correlations overall and for Male, Hispanic, and Non-Caucasian candidates.

Table 114. E-5 Difficulty Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Easier		Harder			Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	50	27.47%	38	20.88%	0.01	34	28.10%	28	23.14%	0.08	36	27.91%	39	30.23%	0.24
Male	50	27.47%	38	20.88%	0.01	31	25.62%	28	23.14%	0.22	35	27.13%	32	24.81%	0.24
Female	28	15.38%	25	13.74%	0.22	17	14.05%	16	13.22%	0.33	15	11.63%	27	20.93%	0.00
Caucasian	34	18.68%	27	14.84%	0.06	28	23.14%	27	22.31%	0.36	29	22.48%	31	24.03%	0.29
African-American	26	14.29%	22	12.09%	0.15	21	17.36%	10	8.26%	0.00	15	11.63%	19	14.73%	0.11
Hispanic	31	17.03%	21	11.54%	0.01	18	14.88%	12	9.92%	0.03	13	10.08%	17	13.18%	0.10
Asian	24	13.19%	14	7.69%	0.00	15	12.40%	3	2.48%	0.00	12	9.30%	13	10.08%	0.31
Non-Caucasian	43	23.63%	28	15.38%	0.00	26	21.49%	19	15.70%	0.03	27	20.93%	32	24.81%	0.12
Low Tenure	32	17.58%	37	20.33%	0.14	27	22.31%	28	23.14%	0.36	19	14.73%	30	23.26%	0.00
High Tenure	56	30.77%	30	16.48%	0.00	31	25.62%	17	14.05%	0.00	33	25.58%	31	24.03%	0.30

Table 115. E-5 Correlation Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Increase		Decrease			Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	30	16.48%	18	9.89%	0.00	29	23.97%	12	9.92%	0.00	19	14.73%	13	10.08%	0.03
Male	37	20.33%	18	9.89%	0.00	29	23.97%	12	9.92%	0.00	18	13.95%	12	9.30%	0.03
Female	15	8.24%	6	3.30%	0.00	6	4.96%	3	2.48%	0.03	7	5.43%	7	5.43%	0.40
Caucasian	26	14.29%	14	7.69%	0.00	20	16.53%	8	6.61%	0.00	12	9.30%	9	6.98%	0.12
African-American	8	4.40%	6	3.30%	0.15	8	6.61%	5	4.13%	0.06	7	5.43%	4	3.10%	0.05
Hispanic	18	9.89%	7	3.85%	0.00	14	11.57%	5	4.13%	0.00	11	8.53%	6	4.65%	0.02
Asian	9	4.95%	9	4.95%	0.41	7	5.79%	4	3.31%	0.05	6	4.65%	4	3.10%	0.11
Non-Caucasian	23	12.64%	10	5.49%	0.00	24	19.83%	8	6.61%	0.00	15	11.63%	8	6.20%	0.01
Low Tenure	24	13.19%	15	8.24%	0.01	22	18.18%	10	8.26%	0.00	10	7.75%	12	9.30%	0.20
High Tenure	25	13.74%	12	6.59%	0.00	18	14.88%	6	4.96%	0.00	13	10.08%	10	7.75%	0.13

Tables 116-118 present effect sizes of difficulty changes for the E-5 paygrade. As with the analysis utilizing all items, the vast majority of items exhibited negligible changes overall and for all groups, and the proportion of items that had small or larger changes appeared to vary with group size. The proportion of easier items decreased as the length of time between administrations increased.

Table 116. E-5 Difficulty Change Effect Sizes - 2 Series Gap

	2 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	103	56.59%	96	6	1	0	79	43.41%	69	10	0	0
Male	104	57.14%	94	9	1	0	78	42.86%	68	10	0	0
Female	101	55.49%	85	16	0	0	81	44.51%	64	15	2	0
Caucasian	101	55.49%	93	7	1	0	81	44.51%	70	11	0	0
African-American	99	54.40%	76	23	0	0	83	45.60%	60	21	2	0
Hispanic	102	56.04%	81	19	2	0	80	43.96%	66	14	0	0
Asian	91	50.00%	66	23	1	1	91	50.00%	64	20	5	2
Non-Caucasian	104	57.14%	87	17	0	0	78	42.86%	68	10	0	0
Low Tenure	85	46.70%	72	12	1	0	97	53.30%	84	13	0	0
High Tenure	105	57.69%	89	15	1	0	77	42.31%	67	10	0	0

Table 117. E-5 Difficulty Change Effect Sizes - 3 Series Gap

	3 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	57	47.11%	46	11	0	0	64	52.89%	61	3	0	0
Male	59	48.76%	46	13	0	0	62	51.24%	56	6	0	0
Female	66	54.55%	52	14	0	0	55	45.45%	49	6	0	0
Caucasian	58	47.93%	44	14	0	0	63	52.07%	54	9	0	0
African-American	65	53.72%	43	22	0	0	55	45.45%	42	13	0	0
Hispanic	56	46.28%	34	21	1	0	65	53.72%	52	13	0	0
Asian	67	55.37%	37	26	4	0	52	42.98%	35	14	3	0
Non-Caucasian	65	53.72%	51	14	0	0	56	46.28%	49	7	0	0
Low Tenure	48	39.67%	38	10	0	0	73	60.33%	64	9	0	0
High Tenure	62	51.24%	50	12	0	0	59	48.76%	55	4	0	0

Table 118. E-5 Difficulty Change Effect Sizes - 4+ Series Gap

	4+ Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	57	44.19%	51	6	0	0	72	55.81%	59	13	0	0
Male	58	44.96%	49	9	0	0	71	55.04%	59	12	0	0
Female	58	44.96%	52	6	0	0	71	55.04%	49	22	0	0
Caucasian	62	48.06%	50	12	0	0	67	51.94%	54	13	0	0
African-American	62	48.06%	50	11	1	0	67	51.94%	39	24	4	0
Hispanic	63	48.84%	53	10	0	0	66	51.16%	50	15	1	0
Asian	59	45.74%	37	21	1	0	70	54.26%	40	27	2	1
Non-Caucasian	64	49.61%	56	8	0	0	65	50.39%	44	21	0	0
Low Tenure	53	41.09%	46	7	0	0	76	58.91%	62	14	0	0
High Tenure	63	48.84%	54	9	0	0	66	51.16%	48	18	0	0

Tables 119-121 present item-total correlation change effect sizes. Overall, the vast majority of items showed negligible changes, with a moderate proportion of small changes and no medium or larger changes. Generally, as group sizes decreased, changes appeared to be less stable, and the proportion of medium and large changes increased. However, some larger groups exhibited larger numbers of medium and larger changes than small groups in some instances (e.g., Males and Caucasians increased item-total correlations after a 2-series gap).

Table 119. E-5 Item-Total Correlation Change Effect Sizes - 2 Series Gap

	2 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	94	51.65%	80	14	0	0	88	48.35%	78	10	0	0
Male	94	51.65%	69	14	2	9	86	47.25%	74	10	1	1
Female	98	53.85%	73	25	0	0	82	45.05%	60	22	0	0
Caucasian	95	52.20%	74	9	2	10	87	47.80%	67	16	2	2
African-American	99	54.40%	65	28	5	1	81	44.51%	61	18	1	1
Hispanic	111	60.99%	72	38	1	0	67	36.81%	53	14	0	0
Asian	97	53.30%	50	38	7	2	80	43.96%	35	31	7	7
Non-Caucasian	104	57.14%	85	17	2	0	76	41.76%	65	11	0	0
Low Tenure	93	51.10%	71	10	4	8	88	48.35%	68	16	2	2
High Tenure	108	59.34%	86	21	1	0	74	40.66%	59	15	0	0

Table 120. E-5 Item-Total Correlation Change Effect Sizes - 3 Series Gap

	3 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	73	60.33%	56	17	0	0	48	39.67%	40	8	0	0
Male	73	60.33%	52	18	0	3	48	39.67%	38	10	0	0
Female	77	63.64%	58	17	2	0	43	35.54%	28	15	0	0
Caucasian	75	61.98%	56	16	0	3	46	38.02%	30	16	0	0
African-American	71	58.68%	37	28	3	3	46	38.02%	29	17	0	0
Hispanic	70	57.85%	43	23	1	3	49	40.50%	35	10	2	2
Asian	67	55.37%	26	32	6	3	49	40.50%	22	19	4	4
Non-Caucasian	72	59.50%	50	20	1	1	48	39.67%	35	13	0	0
Low Tenure	73	60.33%	49	20	1	3	48	39.67%	32	16	0	0
High Tenure	68	56.20%	49	19	0	0	53	43.80%	43	10	0	0

Table 121. E-5 Item-Total Correlation Change Effect Sizes - 4+ Series Gap

	4+ Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	71	55.04%	61	10	0	0	58	44.96%	48	10	0	0
Male	70	54.26%	60	10	0	0	59	45.74%	48	11	0	0
Female	73	56.59%	50	23	0	0	55	42.64%	39	16	0	0
Caucasian	76	58.91%	59	17	0	0	53	41.09%	36	17	0	0
African-American	66	51.16%	24	34	7	1	62	48.06%	43	19	0	0
Hispanic	68	52.71%	40	26	2	0	62	48.06%	41	19	1	1
Asian	64	49.61%	33	17	9	5	56	43.41%	35	19	1	1
Non-Caucasian	67	51.94%	46	19	2	0	62	48.06%	50	12	0	0
Low Tenure	67	51.94%	53	14	0	0	61	47.29%	47	14	0	0
High Tenure	73	56.59%	56	17	0	0	56	43.41%	44	12	0	0

Summary

- After a 2-series gap, the percentage of items that became easier was higher than the percentage of items that became harder. However, the percentage of items that became harder increased as the length of time between administrations increased, and the percentage of items that became harder was significantly greater after a 4-series or more gap. This pattern (stable percentages of easier items and increasing percentages of harder items) was present in both items administered prior to 2015 and items not administered prior to 2015, though a greater percentage of items not administered prior to 2015 became easier than items administered prior to 2015.
- Most items did not show significant item-total correlation changes regardless of the length of time between administrations. For items administered prior to 2015, there was a significantly greater number of items exhibiting increased rather than decreased item-total correlations after a 2-series or 3-series gap. However, there was a non-significant difference after a 4 or more series gap. For items not administered prior to 2015, there

was a significantly greater number of items exhibiting increased rather than decreased item-total correlations regardless of the length of time between administrations.

6.2 Analysis 2: Item Parameter Changes for Repeat Test-Takers

Table 122 presents the number of repeat test-takers for each possible series pair for the E-5 paygrade. Note that within these tables, candidates can be counted in multiple series pairs if they took more than two exams within the rating in the scope of the study. However, only their initial and second viewing of an item were included in the analyses in this section.

Table 123 presents the number of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-5 paygrade. For items with a 2-series, 3-series, or 4-series or more gap between administrations, there was a significantly greater number of items that became easier overall and for all demographic groups. The percentage of items that became harder was relatively consistent or decreased across the different lengths of time between administrations. The percentage of items that became easier decreased as the length of time between administrations increased overall and for all demographic groups.

Table 124 presents the numbers of items for which item-total correlations changed significantly from the initial within-scope administration to the next administration within the E-5 paygrade. For items with a 2-series, 3-series, or 4-series or more gap between administrations, there was a significantly greater number of items for which the item-total correlation increased overall and for all demographic groups with the exception of Females and Asians within the 4-series gap results. The percentages of items for which item-total correlations increased were generally smaller as the length of time between administrations increased. However, the percentages of items for which correlations decreased were relatively consistent across the different lengths of time between administrations overall and within demographic groups.

Table 122. E-5 Repeat Test-Taker Sample Sizes

Series	Admins. Between Series	Overall	Male	Female	Caucasian	African-American	Hispanic	Asian	Non-Caucasian	Low Tenure	High Tenure
227-232	2	3673	3016	657	1666	515	666	259	1596	1153	2520
227-235	3	2485	2056	429	1097	368	445	188	1110	780	1705
227-236	4	1567	1305	262	695	215	288	116	687	494	1073
228-235	2	3843	3134	709	1765	534	664	299	1652	1978	1865
228-236	3	2314	1899	415	1051	307	413	177	991	1162	1152
228-239	4	1686	1400	286	739	245	310	135	749	854	832
231-236	2	3392	2744	648	1564	470	600	270	1451	2227	1165
231-239	3	2529	2056	473	1137	367	461	202	1112	1678	851
231-240	4	1591	1324	267	692	233	311	126	725	1078	513
232-239	2	4069	3241	828	1884	580	719	320	1751	2962	1107
232-240	3	2528	2030	498	1130	378	454	211	1128	1852	676
235-240	2	3720	2948	772	1701	562	665	284	1623	2907	813
227-239	5	1150	965	185	484	177	221	89	528	366	784
227-240	6	707	604	103	290	118	146	53	340	231	476
228-240	5	1038	870	168	436	164	195	91	486	537	501
-	2 Combined	18697	15083	3614	8580	2661	3314	1432	8073	11227	7470
-	3 Combined	9856	8041	1815	4415	1420	1773	778	4341	5472	4384
-	4+ Combined	7739	6468	1271	3336	1152	1471	610	3515	3560	4179

Table 123. E-5 Difficulty Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Easier		Harder			Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	307	37.67%	56	6.87%	0.00	227	33.24%	40	5.86%	0.00	162	28.22%	28	4.88%	0.00
Male	287	35.21%	50	6.13%	0.00	214	31.33%	39	5.71%	0.00	147	25.61%	26	4.53%	0.00
Female	150	18.40%	35	4.29%	0.00	93	13.62%	18	2.64%	0.00	56	9.76%	10	1.74%	0.00
Caucasian	228	27.98%	42	5.15%	0.00	179	26.21%	26	3.81%	0.00	120	20.91%	23	4.01%	0.00
African-American	117	14.36%	17	2.09%	0.00	74	10.83%	10	1.46%	0.00	36	6.27%	11	1.92%	0.00
Hispanic	131	16.07%	25	3.07%	0.00	93	13.62%	13	1.90%	0.00	55	9.58%	11	1.92%	0.00
Asian	79	9.69%	16	1.96%	0.00	57	8.35%	6	0.88%	0.00	45	7.84%	8	1.39%	0.00
Non-Caucasian	217	26.63%	41	5.03%	0.00	153	22.40%	17	2.49%	0.00	95	16.55%	20	3.48%	0.00
Low Tenure	265	32.52%	40	4.91%	0.00	212	31.04%	13	1.90%	0.00	124	21.60%	25	4.36%	0.00
High Tenure	175	21.47%	45	5.52%	0.00	117	17.13%	6	0.88%	0.00	85	14.81%	25	4.36%	0.00

Table 124. E-5 Item-Total Correlation Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Increase		Decrease			Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	151	18.53%	15	1.84%	0.00	99	14.49%	8	1.17%	0.00	78	13.59%	12	2.09%	0.00
Male	111	13.62%	13	1.60%	0.00	71	10.40%	12	1.76%	0.00	50	8.71%	10	1.74%	0.00
Female	56	6.87%	8	0.98%	0.00	24	3.51%	10	1.46%	0.00	15	2.61%	12	2.09%	0.15
Caucasian	83	10.18%	10	1.23%	0.00	55	8.05%	8	1.17%	0.00	27	4.70%	10	1.74%	0.00
African-American	30	3.68%	7	0.86%	0.00	22	3.22%	12	1.76%	0.00	22	3.83%	7	1.22%	0.00
Hispanic	52	6.38%	8	0.98%	0.00	30	4.39%	8	1.17%	0.00	28	4.88%	9	1.57%	0.00
Asian	31	3.80%	7	0.86%	0.00	13	1.90%	2	0.29%	0.00	14	2.44%	12	2.09%	0.23
Non-Caucasian	77	9.45%	8	0.98%	0.00	14	2.05%	9	1.32%	0.04	7	1.22%	3	0.52%	0.01
Low Tenure	75	9.20%	10	1.23%	0.00	58	8.49%	7	1.02%	0.00	25	4.36%	12	2.09%	0.00
High Tenure	63	7.73%	5	0.61%	0.00	40	5.86%	5	0.73%	0.00	21	3.66%	2	0.35%	0.00

Summary

- The percentage of items that became easier for repeat test-takers was significantly greater than the percentage of items that became harder. The percentage of items that became harder and the percentage of items that became easier decreased as lengths of time between administrations increased.
- The percentage of items that showed an increased item-total correlation for repeat test-takers was significantly greater than the percentage of items that exhibited a decreased item-total correlation regardless of the length of time between administrations. The percentages of items for which correlations increased declined as the length of time between administrations increased.

6.3 Analysis 3: Candidate Performance Differences for Initial vs. Repeat Exposures

Table 125 shows the numbers of candidates who performed better or worse on repeat items (i.e., items they had seen in prior administrations) vs. non-repeat items (i.e., items they had not seen in prior administrations) for candidates at the E-5 paygrade. A significantly greater number of candidates performed better on repeat items than on non-repeat items overall and in all demographic groups.

Table 125. E-5 Test-Taker Repeat Performance

	Repeat N	Sig. Better N	Sig. Better %	Sig. Worse N	Sig. Worse %	*p*
Overall	10,013	2,268	22.65%	266	2.66%	0.00
Male	8,036	1,829	22.76%	213	2.65%	0.00
Female	1,977	439	22.21%	53	2.68%	0.00
Caucasian	4,733	1,054	22.27%	133	2.81%	0.00
African-American	1,392	299	21.48%	35	2.51%	0.00
Hispanic	1,751	396	22.62%	53	3.03%	0.00
Asian	719	182	25.31%	11	1.53%	0.00
Non-Caucasian	4,205	966	22.97%	109	2.59%	0.00
Low Tenure	7,868	1,868	23.74%	201	2.55%	0.00
High Tenure	2,145	400	18.65%	65	3.03%	0.00

Summary

Most candidates at the E-5 paygrade did not perform significantly better or worse on repeat items compared to non-repeat items. Of those that did perform better or worse, a significantly greater number of candidates performed better on repeat items than on non-repeat items overall and in all demographic groups.

7.0 E-6 OVERALL RESULTS

The results in this section are for the E-6 paygrade. Table 126 presents total and demographic group sizes for each administration. The sample was predominately male and Caucasian, though when racial/ethnic groups were combined into one Non-Caucasian category, it was consistently larger than the Caucasian category. Administrations 227 and 228 had mostly Low-Tenure candidates, but the other administrations had mostly High-Tenure candidates.

Table 126. E-6 Sample Sizes by Administration

	Administration							
	227	**228**	**231**	**232**	**235**	**236**	**239**	**240**
Overall	5,627	6,622	6,485	6,938	7,493	7,158	7,570	7,341
Male	4,597	5,394	5,304	5,668	6,143	5,830	6,161	5,932
Female	1,030	1,228	1,181	1,270	1,350	1,328	1,409	1,409
Caucasian	2,196	2,586	2,479	2,616	2,864	2,755	2,901	2,849
African-American	1,000	1,137	1,081	1,173	1,224	1,129	1,247	1,207
Hispanic	1,280	1,480	1,513	1,576	1,688	1,586	1,662	1,597
Asian	594	669	639	663	704	677	718	679
Non-Caucasian	3,164	3,635	3,562	3,763	4,016	3,761	3,990	3,814
Low Tenure	2,956	3,563	2,869	2,875	2,545	2,539	2,692	2,507
High Tenure	2,671	3,059	3,616	4,063	4,948	4,619	4,878	4,834

7.1 Analysis 1: Item Parameter Changes Over Time

Table 127 presents the number of repeat items examined in Analysis 1 for the E-6 paygrade. As was noted, results are presented for series with 2, 3, or 4 or more administrations between the initial and subsequent administration.

Table 128 shows the numbers of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-6 paygrade. For items with a 2-series or 3-series gap between administrations, there was a significantly greater number of items that became easier than items that became harder. This effect holds for all groups for a 2-series gap except Caucasian, African-American, and Low-Tenure candidates, as well as for all groups for a 3-series gap except Hispanic and Asian candidates. For items with a 4-series or more gap, there was a non-significant difference in the number of items that became harder and items that became easier overall. This was also true for all demographic groups except Males, who saw a significantly greater number of harder items, and African-Americans, who saw a significantly greater number of easier items. Looking across these results, the change from a significantly greater number of easier items to a non-significant difference appears to be primarily driven by an increase in the proportion of harder items.

Table 127. E-6 Repeat Items

Series	Administrations Between Series	# of Repeat Items
227-232	2	149
227-235	3	188
227-236	4	153
228-235	2	166
228-236	3	204
228-239	4	90
231-236	2	106
231-239	3	182
231-240	4	113
232-239	2	63
232-240	3	127
235-240	2	74
227-239	5	77
227-240	6	56
228-240	5	86
-	2 Combined	558
-	3 Combined	701
-	4+ Combined	575

Table 129 presents the numbers of items for which item-total correlation changed significantly from the initial within-scope administration to the next administration within the E-6 paygrade. For items with a 2-series gap between administrations, there was a non-significant difference between the number of items with increased and decreased item-total correlations overall and for all demographic groups except Asian and Non-Caucasian candidates, who exhibited a greater number of items with an increased item-total correlation than a decreased correlation. For items with a 3-series gap between administrations, there was a significantly greater number of items with an increased item-total correlation than items with a decreased correlation overall. This was also true for all demographic groups except Females, African-Americans, Hispanics, and Asians. For items with a 4-series or more gap between administrations, there was a significantly greater number of items with an increased item-total correlation than a decreased correlation. This was true for Female, Caucasian, Hispanic, Non-Caucasian, Low-Tenure, and High-Tenure candidates. African-Americans and Asians exhibited a significantly greater number of items with a decreased item-total correlation than an increased correlation, whereas Males exhibited no significant difference.

Table 128. E-6 Difficulty Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Easier		Harder			Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	122	21.86%	87	15.59%	0.00	163	23.25%	136	19.40%	0.01	127	22.09%	136	23.65%	0.17
Male	104	18.64%	88	15.77%	0.03	154	21.97%	122	17.40%	0.00	115	20.00%	133	23.13%	0.03
Female	57	10.22%	36	6.45%	0.00	94	13.41%	56	7.99%	0.00	65	11.30%	50	8.70%	0.01
Caucasian	72	12.90%	66	11.83%	0.20	111	15.83%	77	10.98%	0.00	88	15.30%	93	16.17%	0.26
African-American	53	9.50%	44	7.89%	0.07	86	12.27%	53	7.56%	0.00	54	9.39%	42	7.30%	0.03
Hispanic	57	10.22%	41	7.35%	0.01	74	10.56%	69	9.84%	0.24	67	11.65%	61	10.61%	0.19
Asian	49	8.78%	28	5.02%	0.00	49	6.99%	48	6.85%	0.40	35	6.09%	40	6.96%	0.17
Non-Caucasian	104	18.64%	73	13.08%	0.00	118	16.83%	106	15.12%	0.10	91	15.83%	96	16.70%	0.26
Low Tenure	73	13.08%	80	14.34%	0.17	116	16.55%	96	13.69%	0.01	86	14.96%	92	16.00%	0.22
High Tenure	104	18.64%	67	12.01%	0.00	138	19.69%	110	15.69%	0.00	100	17.39%	113	19.65%	0.07

Table 129. E-6 Correlation Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Increase		Decrease			Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	40	7.17%	37	6.63%	0.27	60	8.56%	35	4.99%	0.00	43	7.48%	26	4.52%	0.00
Male	38	6.81%	37	6.63%	0.39	56	7.99%	33	4.71%	0.00	34	5.91%	26	4.52%	0.05
Female	21	3.76%	16	2.87%	0.09	26	3.71%	22	3.14%	0.16	25	4.35%	15	2.61%	0.01
Caucasian	24	4.30%	18	3.23%	0.07	38	5.42%	23	3.28%	0.00	35	6.09%	23	4.00%	0.01
African-American	17	3.05%	17	3.05%	0.44	21	3.00%	15	2.14%	0.05	16	2.78%	23	4.00%	0.03
Hispanic	17	3.05%	23	4.12%	0.06	26	3.71%	21	3.00%	0.11	28	4.87%	14	2.43%	0.00
Asian	24	4.30%	15	2.69%	0.01	19	2.71%	15	2.14%	0.12	11	1.91%	20	3.48%	0.00
Non-Caucasian	43	7.71%	25	4.48%	0.00	46	6.56%	30	4.28%	0.00	34	5.91%	23	4.00%	0.01
Low Tenure	25	4.48%	20	3.58%	0.11	44	6.28%	29	4.14%	0.00	34	5.91%	15	2.61%	0.00
High Tenure	40	7.17%	34	6.09%	0.13	41	5.85%	21	3.00%	0.00	38	6.61%	25	4.35%	0.00

Tables 130-132 present effect sizes of difficulty changes for the E-6 paygrade. For administrations with a 2-series gap, the vast majority of changes were negligible overall and for all groups. The proportion of items that had small or larger changes appeared to vary with group size. For example, the proportion of items with at least a small change (both easier and harder) was greater for smaller groups than for larger groups. As the length of time between administrations increased, the proportion of easier items decreased and the proportion of harder items increased. Though the vast majority of items exhibited negligible changes regardless of length of time between administrations, the proportion of items exhibiting small or larger differences generally increased as the length of time between administrations increased.

Table 130. E-6 Difficulty Change Effect Sizes - 2 Series Gap

	2 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	294	52.69%	268	25	1	0	264	47.31%	243	21	0	0
Male	298	53.41%	273	24	1	0	260	46.59%	238	22	0	0
Female	306	54.84%	245	60	1	0	252	45.16%	206	46	0	0
Caucasian	287	51.43%	257	29	1	0	271	48.57%	242	29	0	0
African-American	319	57.17%	262	56	0	1	239	42.83%	191	47	1	0
Hispanic	295	52.87%	259	35	1	0	260	46.59%	230	30	0	0
Asian	295	52.87%	218	74	2	1	263	47.13%	195	66	2	0
Non-Caucasian	323	57.89%	298	24	0	1	235	42.11%	212	23	0	0
Low Tenure	280	50.18%	257	22	1	0	278	49.82%	241	37	0	0
High Tenure	306	54.84%	267	38	1	0	252	45.16%	230	22	0	0

Table 131. E-6 Difficulty Change Effect Sizes - 3 Series Gap

	3 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	352	50.21%	312	38	2	0	349	49.79%	312	36	1	0
Male	353	50.36%	316	35	2	0	348	49.64%	308	39	1	0
Female	381	54.35%	293	83	5	0	320	45.65%	253	66	1	0
Caucasian	361	51.50%	304	55	2	0	340	48.50%	289	51	0	0
African-American	382	54.49%	292	86	3	1	316	45.08%	232	81	3	0
Hispanic	351	50.07%	291	58	2	0	350	49.93%	287	61	2	0
Asian	366	52.21%	271	92	2	1	331	47.22%	216	110	5	0
Non-Caucasian	362	51.64%	320	40	2	0	339	48.36%	294	44	1	0
Low Tenure	345	49.22%	292	51	2	0	356	50.78%	306	50	0	0
High Tenure	343	48.93%	302	39	2	0	358	51.07%	310	46	2	0

Table 132. E-6 Difficulty Change Effect Sizes - 4+ Series Gap

	4+ Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	265	46.09%	226	38	1	0	310	53.91%	268	37	3	2
Male	265	46.09%	225	39	1	0	310	53.91%	269	36	3	2
Female	300	52.17%	224	76	0	0	274	47.65%	208	60	4	2
Caucasian	271	47.13%	222	48	1	0	304	52.87%	253	46	3	2
African-American	290	50.43%	220	67	3	0	284	49.39%	218	59	5	2
Hispanic	256	44.52%	196	57	3	0	319	55.48%	251	62	4	2
Asian	258	44.87%	177	79	2	0	316	54.96%	210	98	6	2
Non-Caucasian	268	46.61%	225	42	1	0	307	53.39%	261	42	2	2
Low Tenure	264	45.91%	220	43	1	0	311	54.09%	261	46	2	2
High Tenure	272	47.30%	226	45	1	0	303	52.70%	250	47	5	1

Tables 133-135 present item-total correlation change effect sizes. Overall, the proportion of items that exhibited increases was higher after a 4-series gap than other lengths of time, but this was not consistent across demographic groups. The vast majority of items exhibited negligible changes, with a moderate proportion of small changes and few or no medium or larger changes. As group sizes decreased, changes appeared to be less stable, and the proportion of medium and large changes increased.

Table 133. E-6 Item-Total Correlation Change Effect Sizes - 2 Series Gap

	2 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	293	52.51%	275	18	0	0	265	47.49%	243	22	0	0
Male	300	53.76%	273	27	0	0	258	46.24%	234	24	0	0
Female	277	49.64%	179	92	6	0	279	50.00%	164	115	0	0
Caucasian	265	47.49%	208	57	0	0	293	52.51%	234	59	0	0
African-American	288	51.61%	178	105	5	0	266	47.67%	160	106	0	0
Hispanic	298	53.41%	204	93	1	0	256	45.88%	178	78	0	0
Asian	286	51.25%	135	119	32	0	258	46.24%	141	117	0	0
Non-Caucasian	305	54.66%	251	54	0	0	253	45.34%	208	45	0	0
Low Tenure	267	47.85%	205	62	0	0	291	52.15%	245	46	0	0
High Tenure	290	51.97%	232	58	0	0	265	47.49%	205	60	0	0

Table 134. E-6 Item-Total Correlation Change Effect Sizes - 3 Series Gap

	3 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	366	52.21%	327	39	0	0	335	47.79%	306	29	0	0
Male	383	54.64%	335	48	0	0	318	45.36%	285	33	0	0
Female	359	51.21%	213	135	10	1	337	48.07%	215	120	1	1
Caucasian	384	54.78%	297	86	1	0	317	45.22%	256	61	0	0
African-American	349	49.79%	214	129	6	0	343	48.93%	207	136	0	0
Hispanic	340	48.50%	231	107	2	0	359	51.21%	249	110	0	0
Asian	382	54.49%	208	141	33	0	287	40.94%	142	143	1	1
Non-Caucasian	359	51.21%	294	65	0	0	342	48.79%	282	60	0	0
Low Tenure	373	53.21%	270	102	1	0	328	46.79%	249	79	0	0
High Tenure	373	53.21%	291	80	2	0	327	46.65%	266	61	0	0

Table 135. E-6 Item-Total Correlation Change Effect Sizes - 4+ Series Gap

	4+ Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	308	53.57%	281	27	0	0	267	46.43%	241	26	0	0
Male	299	52.00%	267	32	0	0	276	48.00%	245	31	0	0
Female	313	54.43%	165	137	11	0	262	45.57%	157	99	3	3
Caucasian	286	49.74%	196	90	0	0	288	50.09%	221	67	0	0
African-American	298	51.83%	176	112	10	0	266	46.26%	148	118	0	0
Hispanic	289	50.26%	184	102	3	0	283	49.22%	179	104	0	0
Asian	301	52.35%	140	131	29	1	237	41.22%	120	113	2	2
Non-Caucasian	295	51.30%	229	66	0	0	280	48.70%	228	52	0	0
Low Tenure	280	48.70%	197	82	1	0	294	51.13%	224	70	0	0
High Tenure	309	53.74%	232	74	3	0	264	45.91%	197	67	0	0

7.1.1 Items Administered Prior to 2015

Table 136 shows the number of repeat items examined in Analysis 1 that were administered prior to 2015 for the E-6 paygrade.

Table 137 shows the numbers of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-6 paygrade. For items with a 2-series or 3-series gap between administrations, there was a non-significant difference overall. However, there was a significantly greater number of items that became harder among Low-Tenure candidates for the 2-series gap results and for Females and African-Americans for the 3-series gap results, as well as a significantly greater number of items that became easier among Asians for the 3-series gap results. For items with a 4-series gap between administrations, there was a significantly greater number of items that became harder than items that became easier overall and for Male, Caucasian, Asian, Non-Caucasian, and High-Tenure candidates. The change from non-significant differences after 2-series and 3-series gaps to a significant difference after a 4-series gap appears to be driven primarily by a stronger increase in the proportion of harder items as the gap between administrations increases.

Table 136. E-6 Repeat Items

Series	Administrations Between Series	# of Repeat Items
227-232	2	109
227-235	3	126
227-236	4	125
228-235	2	116
228-236	3	164
228-239	4	61
231-236	2	76
231-239	3	133
231-240	4	91
232-239	2	35
232-240	3	85
235-240	2	33
227-239	5	61
227-240	6	41
228-240	5	67
-	2 Combined	369
-	3 Combined	508
-	4+ Combined	446

Table 138 presents the numbers of items for which item-total correlation changed significantly from the initial within-scope administration to the next administration within the E-6 paygrade. For items with a 2-series or 3-series gap between administrations, there was a non-significant difference between the number of items with an increased item-total correlation and the number of items with a decreased correlation. This was also true for all demographic groups for the 2-series gap except Non-Caucasians and for all groups for the 3-series gap except Low-Tenure candidates. For items with a 4-series or larger gap between administrations, there was a significantly greater number of items with an increased item-total correlation than items with an decreased correlation overall and for all groups except Males, African-Americans, and Hispanics, all of whom exhibited non-significant differences, as well as for Asians, who exhibited a significantly greater number of items with a decreased item-total correlation.

Table 137. E-6 Difficulty Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Easier		Harder			Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	62	16.80%	57	15.45%	0.21	101	19.88%	106	20.87%	0.27	90	20.18%	107	23.99%	0.02
Male	53	14.36%	60	16.26%	0.13	93	18.31%	93	18.31%	0.47	79	17.71%	102	22.87%	0.00
Female	24	6.50%	22	5.96%	0.28	57	11.22%	40	7.87%	0.00	42	9.42%	37	8.30%	0.17
Caucasian	37	10.03%	44	11.92%	0.10	64	12.60%	61	12.01%	0.31	60	13.45%	72	16.14%	0.04
African-American	21	5.69%	25	6.78%	0.16	49	9.65%	39	7.68%	0.04	35	7.85%	28	6.28%	0.08
Hispanic	30	8.13%	27	7.32%	0.24	45	8.86%	49	9.65%	0.24	46	10.31%	46	10.31%	0.46
Asian	21	5.69%	19	5.15%	0.27	25	4.92%	39	7.68%	0.00	19	4.26%	30	6.73%	0.01
Non-Caucasian	50	13.55%	48	13.01%	0.34	72	14.17%	80	15.75%	0.14	62	13.90%	75	16.82%	0.03
Low Tenure	40	10.84%	56	15.18%	0.00	73	14.37%	71	13.98%	0.37	63	14.13%	70	15.70%	0.15
High Tenure	48	13.01%	45	12.20%	0.28	78	15.35%	88	17.32%	0.10	67	15.02%	93	20.85%	0.00

Table 138. E-6 Correlation Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Increase		Decrease			Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	20	5.42%	20	5.42%	0.44	35	6.89%	29	5.71%	0.11	30	6.73%	18	4.04%	0.00
Male	17	4.61%	17	4.61%	0.44	31	6.10%	26	5.12%	0.14	23	5.16%	20	4.48%	0.21
Female	11	2.98%	12	3.25%	0.31	17	3.35%	14	2.76%	0.17	16	3.59%	9	2.02%	0.01
Caucasian	12	3.25%	11	2.98%	0.31	24	4.72%	18	3.54%	0.06	28	6.28%	18	4.04%	0.01
African-American	10	2.71%	10	2.71%	0.42	14	2.76%	12	2.36%	0.23	13	2.91%	18	4.04%	0.07
Hispanic	9	2.44%	11	2.98%	0.19	13	2.56%	17	3.35%	0.11	17	3.81%	12	2.69%	0.06
Asian	12	3.25%	9	2.44%	0.12	10	1.97%	13	2.56%	0.13	5	1.12%	12	2.69%	0.00
Non-Caucasian	23	6.23%	12	3.25%	0.00	19	3.74%	22	4.33%	0.20	26	5.83%	17	3.81%	0.01
Low Tenure	15	4.07%	15	4.07%	0.43	29	5.71%	19	3.74%	0.01	24	5.38%	13	2.91%	0.00
High Tenure	20	5.42%	17	4.61%	0.19	20	3.94%	18	3.54%	0.27	28	6.28%	16	3.59%	0.00

Tables 139-141 present effect sizes of difficulty changes for the E-6 paygrade. As with the analysis using all items, the vast majority of items exhibited negligible changes overall and for all groups, and the proportion of items that had small or larger changes appeared to vary with group size. As the length of time between administrations increased, the proportion of easier items decreased, and the proportion of harder items increased.

Table 139. E-6 Difficulty Change Effect Sizes - 2 Series Gap

	2 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	183	49.59%	171	12	0	0	186	50.41%	174	12	0	0
Male	187	50.68%	174	13	0	0	182	49.32%	169	13	0	0
Female	198	53.66%	161	37	0	0	171	46.34%	140	31	0	0
Caucasian	179	48.51%	166	13	0	0	190	51.49%	170	20	0	0
African-American	201	54.47%	168	33	0	0	168	45.53%	132	36	0	0
Hispanic	184	49.86%	160	24	0	0	182	49.32%	163	19	0	0
Asian	184	49.86%	142	41	1	0	185	50.14%	136	48	1	0
Non-Caucasian	207	56.10%	193	14	0	0	162	43.90%	149	13	0	0
Low Tenure	169	45.80%	157	12	0	0	200	54.20%	175	25	0	0
High Tenure	196	53.12%	177	19	0	0	173	46.88%	159	14	0	0

Table 140. E-6 Difficulty Change Effect Sizes - 3 Series Gap

	3 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	238	46.85%	212	26	0	0	270	53.15%	244	25	1	0
Male	234	46.06%	211	23	0	0	274	53.94%	247	26	1	0
Female	265	52.17%	207	56	2	0	243	47.83%	191	51	1	0
Caucasian	250	49.21%	213	37	0	0	258	50.79%	219	39	0	0
African-American	271	53.35%	204	66	1	0	235	46.26%	169	64	2	0
Hispanic	241	47.44%	201	40	0	0	267	52.56%	220	45	2	0
Asian	250	49.21%	182	67	1	0	255	50.20%	169	82	4	0
Non-Caucasian	245	48.23%	216	29	0	0	263	51.77%	230	32	1	0
Low Tenure	245	48.23%	212	33	0	0	263	51.77%	226	37	0	0
High Tenure	227	44.69%	203	24	0	0	281	55.31%	244	35	2	0

Table 141. E-6 Difficulty Change Effect Sizes - 4+ Series Gap

	4+ Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	193	43.27%	168	24	1	0	253	56.73%	220	31	1	1
Male	193	43.27%	167	25	1	0	253	56.73%	220	31	1	1
Female	225	50.45%	174	51	0	0	220	49.33%	166	52	1	1
Caucasian	200	44.84%	166	33	1	0	246	55.16%	205	39	1	1
African-American	221	49.55%	170	50	1	0	224	50.22%	178	42	3	1
Hispanic	192	43.05%	151	40	1	0	254	56.95%	198	53	2	1
Asian	191	42.83%	133	57	1	0	254	56.95%	165	85	3	1
Non-Caucasian	198	44.39%	169	28	1	0	248	55.61%	212	34	1	1
Low Tenure	196	43.95%	164	31	1	0	250	56.05%	209	39	1	1
High Tenure	200	44.84%	170	29	1	0	246	55.16%	205	38	2	1

Tables 142-144 present item-total correlation change effect sizes. Overall, the vast majority of items showed negligible changes, with a moderate proportion of small changes and no medium or larger changes. As group sizes decreased, changes appeared to be less stable, and the proportion of medium and large changes increased.

Table 142. E-6 Item-Total Correlation Change Effect Sizes - 2 Series Gap

	2 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	197	53.39%	188	9	0	0	172	46.61%	155	17	0	0
Male	200	54.20%	186	14	0	0	169	45.80%	152	17	0	0
Female	180	48.78%	112	65	3	0	187	50.68%	106	81	0	0
Caucasian	174	47.15%	134	40	0	0	195	52.85%	148	47	0	0
African-American	185	50.14%	111	71	3	0	180	48.78%	100	80	0	0
Hispanic	206	55.83%	141	64	1	0	160	43.36%	109	51	0	0
Asian	191	51.76%	87	81	23	0	169	45.80%	89	80	0	0
Non-Caucasian	201	54.47%	164	37	0	0	168	45.53%	141	27	0	0
Low Tenure	180	48.78%	138	42	0	0	189	51.22%	158	31	0	0
High Tenure	188	50.95%	154	34	0	0	178	48.24%	133	45	0	0

Table 143. E-6 Item-Total Correlation Change Effect Sizes - 3 Series Gap

	3 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	258	50.79%	234	24	0	0	250	49.21%	225	25	0	0
Male	267	52.56%	236	31	0	0	241	47.44%	214	27	0	0
Female	260	51.18%	154	96	9	1	244	48.03%	149	93	1	1
Caucasian	272	53.54%	211	61	0	0	236	46.46%	186	50	0	0
African-American	249	49.02%	153	91	5	0	251	49.41%	146	105	0	0
Hispanic	238	46.85%	157	79	2	0	268	52.76%	182	86	0	0
Asian	272	53.54%	147	101	24	0	209	41.14%	98	109	1	1
Non-Caucasian	252	49.61%	211	41	0	0	256	50.39%	203	53	0	0
Low Tenure	271	53.35%	199	72	0	0	237	46.65%	176	61	0	0
High Tenure	265	52.17%	207	56	2	0	242	47.64%	191	51	0	0

Table 144. E-6 Item-Total Correlation Change Effect Sizes - 4+ Series Gap

	4+ Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	239	53.59%	218	21	0	0	207	46.41%	189	18	0	0
Male	231	51.79%	206	25	0	0	215	48.21%	193	22	0	0
Female	239	53.59%	121	109	9	0	207	46.41%	126	75	3	3
Caucasian	220	49.33%	148	72	0	0	225	50.45%	173	52	0	0
African-American	225	50.45%	132	86	7	0	212	47.53%	109	103	0	0
Hispanic	224	50.22%	148	73	3	0	219	49.10%	141	78	0	0
Asian	238	53.36%	107	107	24	0	182	40.81%	94	84	2	2
Non-Caucasian	225	50.45%	176	49	0	0	221	49.55%	182	39	0	0
Low Tenure	208	46.64%	147	60	1	0	237	53.14%	182	55	0	0
High Tenure	246	55.16%	189	54	3	0	198	44.39%	142	56	0	0

7.1.2 Items Not Administered Prior to 2015

Table 145 presents the number of repeat items examined in Analysis 1 that were not administered prior to 2015 for the E-6 paygrade.

Table 145. E-6 Repeat Items

Series	Administrations Between Series	# of Repeat Items
227-232	2	40
227-235	3	62
227-236	4	28
228-235	2	50
228-236	3	40
228-239	4	29
231-236	2	30
231-239	3	49
231-240	4	22
232-239	2	28
232-240	3	42
235-240	2	41
227-239	5	16
227-240	6	15
228-240	5	19
-	2 Combined	189
-	3 Combined	193
-	4+ Combined	129

Table 146 presents the number of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-6 paygrade. For items with a 2-series, 3-series, or 4-seies gap between administrations, there was a significantly greater number of items that became easier than items that became harder overall as well as for all demographic groups within the 2-series and 3-series results. This was also the case for all demographic groups in the 4-series results except Male, African-American, and Low-Tenure candidates, for each of which there was no significant difference.

Table 147 presents the numbers of items for which item-total correlation changed significantly from the initial within-scope administration to the next administration within the E-6 paygrade. For items with a 2-series gap between administrations, there was a non-significant difference between the number of items that exhibited increased item-total correlations than items that exhibited decreased item-total correlations overall. Female, Caucasian, Asian, Non-Caucasian, and Low-Tenure candidates had significantly more items with increased correlations than decreased correlations. For items with a 3-series gap between administrations, there was a significantly greater number of items that showed increased item-total correlations than decreased item-total correlations overall as well as for all demographic subgroups with the exception of Females. For items with a 4-series or more gap between administrations, there was a significantly greater number of items that showed increased item-total correlations than decreased item-total correlations overall as well as for Male, Hispanic, and Low-Tenure candidates.

Table 146. E-6 Difficulty Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Easier		Harder			Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	60	31.75%	30	15.87%	0.00	62	32.12%	30	15.54%	0.00	37	28.68%	29	22.48%	0.04
Male	51	26.98%	28	14.81%	0.00	61	31.61%	29	15.03%	0.00	36	27.91%	31	24.03%	0.13
Female	33	17.46%	14	7.41%	0.00	37	19.17%	16	8.29%	0.00	23	17.83%	13	10.08%	0.00
Caucasian	35	18.52%	22	11.64%	0.00	47	24.35%	16	8.29%	0.00	28	21.71%	21	16.28%	0.04
African-American	32	16.93%	19	10.05%	0.00	37	19.17%	14	7.25%	0.00	19	14.73%	14	10.85%	0.06
Hispanic	27	14.29%	14	7.41%	0.00	29	15.03%	20	10.36%	0.02	21	16.28%	15	11.63%	0.04
Asian	28	14.81%	9	4.76%	0.00	24	12.44%	9	4.66%	0.00	16	12.40%	10	7.75%	0.02
Non-Caucasian	54	28.57%	25	13.23%	0.00	46	23.83%	26	13.47%	0.00	29	22.48%	21	16.28%	0.03
Low Tenure	33	17.46%	24	12.70%	0.02	43	22.28%	25	12.95%	0.00	23	17.83%	22	17.05%	0.35
High Tenure	56	29.63%	22	11.64%	0.00	60	31.09%	22	11.40%	0.00	33	25.58%	20	15.50%	0.00

Table 147. E-6 Correlation Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Increase		Decrease			Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	20	10.58%	17	8.99%	0.18	25	12.95%	6	3.11%	0.00	13	10.08%	8	6.20%	0.03
Male	21	11.11%	20	10.58%	0.35	25	12.95%	7	3.63%	0.00	11	8.53%	6	4.65%	0.02
Female	10	5.29%	4	2.12%	0.00	9	4.66%	8	4.15%	0.28	9	6.98%	6	4.65%	0.08
Caucasian	12	6.35%	7	3.70%	0.02	14	7.25%	5	2.59%	0.00	7	5.43%	5	3.88%	0.13
African-American	7	3.70%	7	3.70%	0.40	7	3.63%	3	1.55%	0.01	3	2.33%	5	3.88%	0.08
Hispanic	8	4.23%	12	6.35%	0.06	13	6.74%	4	2.07%	0.00	11	8.53%	2	1.55%	0.00
Asian	12	6.35%	6	3.17%	0.01	9	4.66%	2	1.04%	0.00	6	4.65%	8	6.20%	0.15
Non-Caucasian	20	10.58%	13	6.88%	0.02	27	13.99%	8	4.15%	0.00	8	6.20%	6	4.65%	0.15
Low Tenure	10	5.29%	5	2.65%	0.01	15	7.77%	10	5.18%	0.04	10	7.75%	2	1.55%	0.00
High Tenure	20	10.58%	17	8.99%	0.18	21	10.88%	3	1.55%	0.00	10	7.75%	9	6.98%	0.29

Tables 148-150 present effect sizes of difficulty changes for the E-5 paygrade. As with the analysis utilizing all items, the vast majority of items exhibited negligible changes overall and for all groups, and the proportion of items that had small or larger changes appeared to vary with group size. The proportion of easier items was highest after a 3-series gap, but lowest after a 4-series gap.

Table 148. E-6 Difficulty Change Effect Sizes - 2 Series Gap

	2 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	111	58.73%	97	13	1	0	78	41.27%	69	9	0	0
Male	111	58.73%	99	11	1	0	78	41.27%	69	9	0	0
Female	108	57.14%	84	23	1	0	81	42.86%	66	15	0	0
Caucasian	108	57.14%	91	16	1	0	81	42.86%	72	9	0	0
African-American	118	62.43%	94	23	0	1	71	37.57%	59	11	1	0
Hispanic	111	58.73%	99	11	1	0	78	41.27%	67	11	0	0
Asian	111	58.73%	76	33	1	1	78	41.27%	59	18	1	0
Non-Caucasian	116	61.38%	105	10	0	1	73	38.62%	63	10	0	0
Low Tenure	111	58.73%	100	10	1	0	78	41.27%	66	12	0	0
High Tenure	110	58.20%	90	19	1	0	79	41.80%	71	8	0	0

Table 149. E-6 Difficulty Change Effect Sizes - 3 Series Gap

	3 Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	114	59.07%	100	12	2	0	79	40.93%	68	11	0	0
Male	119	61.66%	105	12	2	0	74	38.34%	61	13	0	0
Female	116	60.10%	86	27	3	0	77	39.90%	62	15	0	0
Caucasian	111	57.51%	91	18	2	0	82	42.49%	70	12	0	0
African-American	111	57.51%	88	20	2	1	81	41.97%	63	17	1	0
Hispanic	110	56.99%	90	18	2	0	83	43.01%	67	16	0	0
Asian	116	60.10%	89	25	1	1	76	39.38%	47	28	1	0
Non-Caucasian	117	60.62%	104	11	2	0	76	39.38%	64	12	0	0
Low Tenure	100	51.81%	80	18	2	0	93	48.19%	80	13	0	0
High Tenure	116	60.10%	99	15	2	0	77	39.90%	66	11	0	0

Table 150. E-6 Difficulty Change Effect Sizes - 4+ Series Gap

	4+ Series Gap											
	Easier						Harder					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	72	55.81%	58	14	0	0	57	44.19%	48	6	2	1
Male	72	55.81%	58	14	0	0	57	44.19%	49	5	2	1
Female	75	58.14%	50	25	0	0	54	41.86%	42	8	3	1
Caucasian	71	55.04%	56	15	0	0	58	44.96%	48	7	2	1
African-American	69	53.49%	50	17	2	0	60	46.51%	40	17	2	1
Hispanic	64	49.61%	45	17	2	0	65	50.39%	53	9	2	1
Asian	67	51.94%	44	22	1	0	62	48.06%	45	13	3	1
Non-Caucasian	70	54.26%	56	14	0	0	59	45.74%	49	8	1	1
Low Tenure	68	52.71%	56	12	0	0	61	47.29%	52	7	1	1
High Tenure	72	55.81%	56	16	0	0	57	44.19%	45	9	3	0

Tables 151-153 present item-total correlation change effect sizes. Overall, the vast majority of items exhibited negligible changes, with some small changes and no medium or larger changes. Generally, as group sizes decreased, changes appeared to be less stable, and the proportion of medium and large changes increased.

Table 151. E-6 Item-Total Correlation Change Effect Sizes - 2 Series Gap

	2 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	96	50.79%	87	9	0	0	93	49.21%	88	5	0	0
Male	100	52.91%	87	13	0	0	89	47.09%	82	7	0	0
Female	97	51.32%	67	27	3	0	92	48.68%	58	34	0	0
Caucasian	91	48.15%	74	17	0	0	98	51.85%	86	12	0	0
African-American	103	54.50%	67	34	2	0	86	45.50%	60	26	0	0
Hispanic	92	48.68%	63	29	0	0	96	50.79%	69	27	0	0
Asian	95	50.26%	48	38	9	0	89	47.09%	52	37	0	0
Non-Caucasian	104	55.03%	87	17	0	0	85	44.97%	67	18	0	0
Low Tenure	87	46.03%	67	20	0	0	102	53.97%	87	15	0	0
High Tenure	102	53.97%	78	24	0	0	87	46.03%	72	15	0	0

Table 152. E-6 Item-Total Correlation Change Effect Sizes - 3 Series Gap

	3 Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	108	55.96%	93	15	0	0	85	44.04%	81	4	0	0
Male	116	60.10%	99	17	0	0	77	39.90%	71	6	0	0
Female	99	51.30%	59	39	1	0	93	48.19%	66	27	0	0
Caucasian	112	58.03%	86	25	1	0	81	41.97%	70	11	0	0
African-American	100	51.81%	61	38	1	0	92	47.67%	61	31	0	0
Hispanic	102	52.85%	74	28	0	0	91	47.15%	67	24	0	0
Asian	110	56.99%	61	40	9	0	78	40.41%	44	34	0	0
Non-Caucasian	107	55.44%	83	24	0	0	86	44.56%	79	7	0	0
Low Tenure	102	52.85%	71	30	1	0	91	47.15%	73	18	0	0
High Tenure	108	55.96%	84	24	0	0	85	44.04%	75	10	0	0

Table 153. E-6 Item-Total Correlation Change Effect Sizes - 4+ Series Gap

	4+ Series Gap											
	Increase						Decrease					
	#	%	Neg.	Small	Med.	Large	#	%	Neg.	Small	Med.	Large
Overall	69	53.49%	63	6	0	0	60	46.51%	52	8	0	0
Male	68	52.71%	61	7	0	0	61	47.29%	52	9	0	0
Female	74	57.36%	44	28	2	0	55	42.64%	31	24	0	0
Caucasian	66	51.16%	48	18	0	0	63	48.84%	48	15	0	0
African-American	73	56.59%	44	26	3	0	54	41.86%	39	15	0	0
Hispanic	65	50.39%	36	29	0	0	64	49.61%	38	26	0	0
Asian	63	48.84%	33	24	5	1	55	42.64%	26	29	0	0
Non-Caucasian	70	54.26%	53	17	0	0	59	45.74%	46	13	0	0
Low Tenure	72	55.81%	50	22	0	0	57	44.19%	42	15	0	0
High Tenure	63	48.84%	43	20	0	0	66	51.16%	55	11	0	0

Summary

- Most items did not change in difficulty regardless of the length of time between administrations. The percentage of items that became harder increased as the length of time between administrations increased. After a 2-series or 3-series gap, the percentage of items that became easier was significantly higher than the percentage of items became harder, though the difference was non-significant after a 4 or more series gap. For items administered prior to 2015, only the 4 or more series gap exhibited a significant difference (a greater number of harder items), whereas for items not administered prior to 2015, there was a significantly greater number of easier items regardless of the length of time between administrations.
- Most items did not exhibit significant item-total correlation changes regardless of the length of time between administrations. For items administered prior to 2015, there was a significantly greater number of items exhibiting increased rather than decreased item-total correlations after a 4 or more gap, but there was a non-significant difference after a

2-series or 3-series gap. For items not administered prior to 2015, there was a significantly greater number of items exhibiting increased rather than decreased item-total correlations after a 3-series or 4 or more series gap.

7.2 Analysis 2: Item Parameter Changes for Repeat Test-Takers

Table 154 presents the number of repeat test-takers for each possible series pair for the E-6 paygrade. Note that within these tables, candidates can be counted in multiple series pairs if they took more than two exams within the rating in the scope of the study. However, only their initial and second viewing of an item were included in the analyses in this section.

Table 155 presents the numbers of items for which difficulty changed significantly from the initial within-scope administration to the next administration within the E-6 paygrade. For items with a 2-series, 3-series, or 4-series or more gap between administrations, there was a significantly greater number of items that became easier overall and for all demographic groups. The percentage of items that became harder was higher after a 2-series gap than after a 4-series gap overall and for all demographic groups, though both increases and decreases occurred after a 3-series gap. The percentage of items that became easier generally decreased as the length of time between administrations increased overall and for all demographic groups.

Table 156 presents the numbers of items for which item-total correlations changed significantly from the initial within-scope administration to the next administration within the E-6 paygrade. For items with a 2-series, 3-series, or 4-series or larger gap between administrations, there was a significantly greater number of items for which the item-total correlation increased overall and for all demographic groups with the exception of Female, Caucasian, African-American, Asian, and High-Tenure candidates within the 4-series gap results. The percentages of items for which item-total correlations increased were generally smaller as the length of time between administrations increased. The percentages of items for which correlations decreased for most groups from the 2-series gap results to the 3-series gap results, but increased for most groups from the 3-series gap to the 4-series or more gap results.

Table 154. E-6 Repeat Test-Taker Sample Sizes

Series	Admins. Between Series	Overall	Male	Female	Caucasian	African-American	Hispanic	Asian	Non-Caucasian	Low Tenure	High Tenure
227-232	2	2605	2123	482	951	493	616	285	1527	1494	1111
227-235	3	2350	1925	425	892	440	532	259	1349	1370	980
227-236	4	1693	1383	310	625	300	396	210	990	988	705
228-235	2	3643	2977	666	1389	667	813	384	2054	2514	1129
228-236	3	2631	2152	479	976	472	603	303	1519	1814	817
228-239	4	2225	1816	409	811	412	507	268	1309	1528	697
231-236	2	3243	2668	575	1209	543	766	356	1826	2464	779
231-239	3	2735	2260	475	993	482	647	312	1579	2071	664
231-240	4	2141	1780	361	782	366	515	242	1225	1620	521
232-239	2	3596	2969	627	1321	626	833	374	2013	2870	726
232-240	3	2810	2323	487	1020	488	664	297	1588	2245	565
235-240	2	3677	3042	635	1327	619	876	378	2067	3026	651
227-239	5	1438	1178	260	533	266	329	178	849	837	601
227-240	6	1149	949	200	414	211	276	153	690	678	471
228-240	5	1731	1423	308	624	320	411	211	1025	1188	543
-	2 Combined	16764	13779	2985	6197	2948	3904	1777	9487	12368	4396
-	3 Combined	10526	8660	1866	3881	1882	2446	1171	6035	7500	3026
-	4+ Combined	10377	8529	1848	3789	1875	2434	1262	6088	6839	3538

Table 155. E-6 Difficulty Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Easier		Harder			Easier		Harder			Easier		Harder		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	209	37.19%	36	6.41%	0.00	253	36.04%	56	7.98%	0.00	182	31.65%	33	5.74%	0.00
Male	197	35.05%	35	6.23%	0.00	243	34.62%	50	7.12%	0.00	172	29.91%	28	4.87%	0.00
Female	89	15.84%	28	4.98%	0.00	104	14.81%	25	3.56%	0.00	59	10.26%	14	2.43%	0.00
Caucasian	127	22.60%	29	5.16%	0.00	170	24.22%	33	4.70%	0.00	112	19.48%	22	3.83%	0.00
African-American	105	18.68%	18	3.20%	0.00	119	16.95%	25	3.56%	0.00	65	11.30%	17	2.96%	0.00
Hispanic	107	19.04%	23	4.09%	0.00	134	19.09%	38	5.41%	0.00	68	11.83%	14	2.43%	0.00
Asian	68	12.10%	19	3.38%	0.00	78	11.11%	14	1.99%	0.00	50	8.70%	13	2.26%	0.00
Non-Caucasian	174	30.96%	31	5.52%	0.00	194	27.64%	43	6.13%	0.00	133	23.13%	25	4.35%	0.00
Low Tenure	197	35.05%	25	4.45%	0.00	230	32.76%	38	5.41%	0.00	162	28.17%	25	4.35%	0.00
High Tenure	94	16.73%	27	4.80%	0.00	125	17.81%	14	1.99%	0.00	74	12.87%	22	3.83%	0.00

Table 156. E-6 Item-Total Correlation Changes

	2 Series Gap					3 Series Gap					4+ Series Gap				
	Increase		Decrease			Increase		Decrease			Increase		Decrease		
	#	%	#	%	*p*	#	%	#	%	*p*	#	%	#	%	*p*
Overall	68	12.10%	13	2.31%	0.00	77	10.97%	8	1.14%	0.00	52	9.04%	20	3.48%	0.00
Male	47	8.36%	8	1.42%	0.00	59	8.40%	7	1.00%	0.00	32	5.57%	19	3.30%	0.00
Female	30	5.34%	12	2.14%	0.00	14	1.99%	3	0.43%	0.00	10	1.74%	10	1.74%	0.42
Caucasian	29	5.16%	7	1.25%	0.00	30	4.27%	8	1.14%	0.00	18	3.13%	15	2.61%	0.18
African-American	23	4.09%	7	1.25%	0.00	26	3.70%	12	1.71%	0.00	14	2.43%	14	2.43%	0.43
Hispanic	23	4.09%	9	1.60%	0.00	27	3.85%	5	0.71%	0.00	24	4.17%	8	1.39%	0.00
Asian	23	4.09%	5	0.89%	0.00	18	2.56%	11	1.57%	0.02	18	3.13%	18	3.13%	0.44
Non-Caucasian	41	7.30%	14	2.49%	0.00	17	2.42%	8	1.14%	0.00	17	2.96%	9	1.57%	0.00
Low Tenure	35	6.23%	8	1.42%	0.00	44	6.27%	3	0.43%	0.00	26	4.52%	7	1.22%	0.00
High Tenure	28	4.98%	8	1.42%	0.00	21	2.99%	5	0.71%	0.00	13	2.26%	10	1.74%	0.13

Summary

- The percentage of items that became easier for repeat test-takers was significantly greater than the percentage of items that became harder. Overall, over a third of items became easier after a 2-series gap, but this decreased to just under a third of items after a 4-series gap.
- The percentage of items that exhibited an increased item-total correlation for repeat test-takers was significantly greater than the percentage of items that exhibited a decreased item-total correlation overall, but not for all demographic groups.

7.3 Analysis 3: Candidate Performance Differences for Initial vs. Repeat Exposures

Table 157 shows the number of candidates who performed better or worse on repeat items (i.e., items they had seen in prior administrations) vs. non-repeat items (i.e., items they had not seen in prior administrations) for candidates at the E-6 paygrade. A significantly greater number of candidates performed better on repeat items than on non-repeat items overall and in all demographic groups.

Table 157. E-6 Test-Taker Repeat Performance

	Repeat N	Sig. Better N	Sig. Better %	Sig. Worse N	Sig. Worse %	*p*
Overall	6,923	1,338	19.33%	99	1.43%	0.00
Male	5,685	1,132	19.91%	78	1.37%	0.00
Female	1,238	206	16.64%	21	1.70%	0.00
Caucasian	2,563	489	19.08%	37	1.44%	0.00
African-American	1,181	229	19.39%	13	1.10%	0.00
Hispanic	1,636	338	20.66%	27	1.65%	0.00
Asian	672	139	20.68%	4	0.60%	0.00
Non-Caucasian	3,860	773	20.03%	53	1.37%	0.00
Low Tenure	5,351	999	18.67%	72	1.35%	0.00
High Tenure	1,572	339	21.56%	27	1.72%	0.00

Summary

Most candidates at the E-6 paygrade do not perform significantly better or worse on repeat items compared to non-repeat items. Of those that do perform better or worse, a significantly greater number of candidates performed better on repeat items than on non-repeat items overall and in all demographic groups.

8.0 SUMMARY TABLES

Throughout the results presented up to this point, the results of note have come from examination of statistically significant differences in Analysis 1. This section summarizes these differences for each paygrade in three ways: overall and by administration date (i.e., pre/post-2015), by demographic group, and for the three ratings for which complete (i.e., E-4-E-7) data were available (i.e., HM, IT, and MA). For item difficulty results, "H"=Harder and "E"=Easier. For item-total correlation results, "I"=Increase and "D"=Decrease. For E-4, E-5, and E-6 results, "Shortest Gap" indicates a 2-series gap, "Medium Gap" indicates a 3-series gap, and "Longest Gap" indicates a 4 or more series gap. For E-7 results, "Shortest Gap" indicates a 1-series gap, while "Medium Gap" indicates a 2-series gap; data for longer gaps were not available for the E-7 results.

Table 158 summarizes statistically significant changes in item difficulty across all within-scope ratings for each paygrade. Results are presented for all items, as well as for only items administered for the first time prior to 2015 and for only items administered for the first time after 2015. For the E-4 and E-5 paygrades, there is a significantly higher number of items that became harder after medium or longer gaps, but for the E-6 and E-7 paygrades, there was a significantly higher number of items that became easier after a short or medium gap. The items administered prior to 2015 exhibited five instances of significantly greater numbers of harder items after a medium or longer gap across the E-4, E-5, and E-6 paygrades, whereas the items administered after 2015 showed five instances of significantly greater numbers of easier items across the E-5, E-6, and E-7 paygrades.

Table 158. Item Difficulty Changes

	Paygrade	Shortest Gap	Medium Gap	Longest Gap
All Items	E-4		H	H
	E-5			H
	E-6	E	E	
	E-7	E	E	--
Items administered prior to 2015	E-4		H	H
	E-5		H	H
	E-6			H
	E-7		E	--
Items administered after 2015	E-4		H	
	E-5	E		
	E-6	E	E	E
	E-7	E		--

Table 159 summarizes statistically significant changes in item-total correlation across all within-scope ratings for each paygrade. There was no clear pattern for item-total correlation changes. Instances of significantly greater numbers of items with increased correlations were far more common than instances of significantly greater numbers of items with decreased correlations for all times and for items divided by administration date, but the pattern across paygrades or gap lengths was not apparent.

Table 159. Item-Total Correlation Changes

	Paygrade	Shortest Gap	Medium Gap	Longest Gap
All Items	E-4			D
	E-5	I	I	
	E-6		I	I
	E-7			--
Items administered prior to 2015	E-4			D
	E-5	I	I	D
	E-6			I
	E-7		I	--
Items administered after 2015	E-4	I		
	E-5	I	I	I
	E-6		I	I
	E-7	I	D	--

Table 160 summarizes statistically significant changes in item difficulty across all within-scope ratings for each paygrade and demographic group. The pattern observed in the overall results applied generally well across demographic groups, with instances of greater numbers of harder items more likely as gap length increased for lower paygrades, and instances of greater numbers of easier items more likely as gap length decreased for higher paygrades. However, there were some exceptions to this, such as Female and African-American candidates in the E-6 paygrade exhibiting a significantly greater number of easier items after 4 or more series gaps, and low-tenure candidates exhibiting a significantly greater number of harder items across all gap lengths.

Table 161 summarizes statistically significant changes in item-total correlation across all within-scope ratings for each paygrade and demographic group. All instances of significantly greater numbers of item-total correlation decreases occurred after the longest gaps between administrations. Instances of significantly greater numbers of item-total correlation increases were common regardless of gap length, but the patterns across demographic groups were unclear and inconsistent.

Table 160. Item Difficulty Changes by Demographic Group

	Paygrade	Shortest Gap	Medium Gap	Longest Gap
Male	E-4		H	H
	E-5			H
	E-6	E	E	H
	E-7	E	E	--
Female	E-4		H	H
	E-5			H
	E-6	E	E	E
	E-7	E	E	--
Caucasian	E-4		H	H
	E-5			
	E-6		E	
	E-7	E	E	--
African-American	E-4		H	H
	E-5			H
	E-6		E	E
	E-7	E	E	--
Hispanic	E-4		H	H
	E-5	E	E	H
	E-6	E		
	E-7	E	E	--
Asian	E-4			H
	E-5	E	E	
	E-6	E		
	E-7	E		--
Non-Caucasian	E-4		H	H
	E-5	E		H
	E-6	E		
	E-7	E	E	--
Low Tenure	E-4	H	H	H
	E-5	H	H	H
	E-6		E	
	E-7	E	E	--
High Tenure	E-4		H	H
	E-5	E	E	H
	E-6	E	E	
	E-7	E	E	--

Table 161. Item-Total Correlation Changes by Demographic Group

	Paygrade	Shortest Gap	Medium Gap	Longest Gap
Male	E-4	I	I	
	E-5	I	I	
	E-6		I	I
	E-7			--
Female	E-4	I		D
	E-5	I	I	D
	E-6			I
	E-7	I		--
Caucasian	E-4			D
	E-5	I	I	
	E-6		I	I
	E-7		D	--
African-American	E-4		I	
	E-5			
	E-6			D
	E-7		I	--
Hispanic	E-4			D
	E-5	I	I	
	E-6			I
	E-7	I		--
Asian	E-4			
	E-5			D
	E-6	I		D
	E-7	I	I	--
Non-Caucasian	E-4			D
	E-5	I	I	
	E-6	I	I	I
	E-7	I	I	--
Low Tenure	E-4			D
	E-5		I	D
	E-6		I	I
	E-7			--
High Tenure	E-4	I		D
	E-5	I	I	
	E-6		I	I
	E-7		I	--

Table 162 summarizes statistically significant changes in item difficulty within the three ratings for which complete E-4 to E-7 paygrade data were available. Across all items, the HM rating exhibited significantly greater numbers of easier items for the E-5, E-6, and E-7 paygrades regardless of gap length; results were largely similar for items administered prior to 2015 and items administered after 2015. The IT and MA ratings exhibited predominately instances of harder items for the E-4 and E-5 paygrades, a pattern that appears more consistent among items administered prior to 2015 than items administered after 2015. Within the E-6 and E-7 paygrades, the IT and MA ratings exhibited non-significant differences or significantly greater

numbers of items that became easier across all items, though the patterns were not clear for these results or when dividing items by administration date.

Table 162. Item Difficulty Changes

	Rating	Paygrade	Shortest Gap	Medium Gap	Longest Gap
All items	HM	E-4	E		H
		E-5	E	E	E
		E-6	E	E	E
		E-7	E	E	--
	IT	E-4		H	H
		E-5	H		H
		E-6		E	
		E-7	E		--
	MA	E-4		H	H
		E-5	E	H	H
		E-6			E
		E-7			--
Items administered prior to 2015	HM	E-4	E	H	H
		E-5		E	
		E-6	E	E	E
		E-7	E	E	--
	IT	E-4		H	H
		E-5	H		H
		E-6			H
		E-7	E	E	--
	MA	E-4	H	H	H
		E-5		H	H
		E-6	H	H	
		E-7			--
Items administered after 2015	HM	E-4		E	
		E-5	E	E	E
		E-6	E	E	E
		E-7	E	H	--
	IT	E-4		H	
		E-5			H
		E-6		E	
		E-7	E		--
	MA	E-4		H	
		E-5	E	H	
		E-6	E		E
		E-7			--

Table 163 summarizes statistically significant changes in item-total correlation within the three ratings for which complete E-4 to E-7 paygrade data were available. As with the results for all ratings, the pattern within these ratings across paygrades or gap lengths was not apparent.

Table 163. Item-Total Correlation Changes

	Rating	Paygrade	Shortest Gap	Medium Gap	Longest Gap
All items	HM	E-4	I		D
		E-5	I	I	
		E-6		I	I
		E-7	I	I	--
	IT	E-4	D		
		E-5		D	D
		E-6		I	
		E-7			--
	MA	E-4			
		E-5	I	I	I
		E-6		I	I
		E-7	D	D	--
Items administered prior to 2015	HM	E-4		D	D
		E-5	I	I	
		E-6			I
		E-7	I	I	--
	IT	E-4	D		
		E-5	I	D	D
		E-6		I	D
		E-7			--
	MA	E-4	I		
		E-5	I	I	
		E-6		I	I
		E-7	D	D	--
Items administered after 2015	HM	E-4	I	I	
		E-5	I	I	
		E-6		I	
		E-7	I		--
	IT	E-4	D		
		E-5		I	D
		E-6		I	
		E-7			--
	MA	E-4			
		E-5		I	I
		E-6	I	I	
		E-7		D	--

9.0 BEST PRACTICES AND RECOMMENDATIONS

In this section, we discuss current best practices as reviewed in previous sections and how they can be applied to managing a testing environment, including detection and control of item exposure, creation and implementation of parallel forms, applications for alternative testing methods, as well as designing and using situational judgment testing as a supplement to existing test conditions. We also discuss how the results of the analyses conducted for the current project and described above inform these recommendations in the testing context of the NEAS process.

9.1 Detecting Item Exposure

Item exposure should be understood and addressed when developing a testing program that relies on repeated uses of test content. The first step in addressing this is to accurately detect item exposure. Unfortunately, the research on this topic is still fairly novel and under development (Belov, 2017). As a result, common methods currently in use may mischaracterize items as compromised that are not and mischaracterize items as not compromised that are (Type I and Type II errors; Wainer, 2014). Given that this is still an emerging field of study, there is no single preferred way to fix these issues and correctly detect item exposure (Belov, 2017). However, there are several general recommendations for overcoming subsets of them. Because each test for detecting item exposure has its own limitations, they are susceptible to categorizing exposure incorrectly so a key recommendation is to use multiple sources of evidence to draw conclusions regarding whether an item has been compromised (Boughton et al., 2017; Kantrowitz & Gutierrez, 2013). Limitations that any one test may have can be offset by using multiple tests and therefore multiple sources of information.

Overall, there are two primary methods of detecting issues with item exposure: careful monitoring of item exposure rates and the use of statistical techniques that look for aberrations in responding (Zara, 2006). These two methods should be used in combination in order to detect whether items may be compromised. Kantrowitz and Gutierrez (2013) recommend methods that account for repeated responses patterns, changes in pass rates, and response latency changes as being the best for identifying suspicious data. Of the statistical techniques that can be used for detection, the l_z statistic has been shown to have the highest degree of stability compared to other methods, but also has been shown to have smallest degree of sensitivity to uncertainty (Belov, 2016). The scale-purified DGM has been shown to have better performance on average than the majority of other statistics, but does not have the same degree of stability found in statistics such as l_z (Belov, 2016; Eckerly 2016b). Again, it is wise to use multiple tools in a coordinated effort to address as many factors as possible.

It is also important to understand the source of the problem driving effects of item exposure. Identifying the problem and its source enables examiners to address the issue. If the problem is in a certain subpopulation of those being tested, then examiners must detect and identify the key characteristics of that subpopulation (Belov, 2014). For example, if individuals know the correct responses due to some theft or unauthorized access to the item pool, examiners can take steps to bolster security of the item pool. This can include protecting both physical and virtual areas in which the item pool is stored as well as encrypting and protecting the security of transmission of item pools (Way, 1998). Further, if examiners can identify which subset of items being known or otherwise compromised, then they can identify which items fall into this subset, where they were stored, and which individuals had access to this subset of information (Belov, 2014; Way, 1998).

9.2 Controlling for Item Exposure

There are three primary options for controlling item exposure:

- Optimizing tests and test banks,
- Developing parallel forms,and
- Switching to computerized multi-stage testing environments.

9.2.1 Optimizing Tests and Test Banks

On a basic level, the key to optimizing test banks is to increase the number of items. Longer tests have been shown to be more resistant to test compromise (Guo et al., 2009). Having more items within a test decreases the impact of individuals having preknowledge of items because each exposed item contributes to a smaller degree of the examinee's score. Similarly, the likelihood of item exposure will go down when there are more items in a test bank because examinees would likely have preknowledge of a smaller portion of the test pool. That is, the more items that exist in the overall pool, the more resistant the test becomes to item exposure.

Another method to optimize test banks is to use multiple or rotating item pools (Lim, 2011; Stoking & Swanson, 1993; Veldkamp & van der Linden, 2004; Way, 1998; Zhang & Chang, 2005). This is done by splitting the overall test bank into several smaller item pools, and then creating a schedule rotation by which these item pools can be used to develop tests. Consistently rotating the pool of items from which the test can be developed makes it less likely that individuals will be exposed to any specific item or subset of items, thus increasing overall test security.

9.2.2 Creating Parallel Forms

Another common way to help control for exposure of items in conventional paper and pencil testing settings is through the development of parallel forms (Hetter & Sympson, 1997), whereby equivalency can be calculated as described below. Having different examinees see alternative forms of the test reduces the amount of exposure for the item subset on any given test while still allowing examiners to test the same constructs across individuals and over multiple administrations. Ideally, parallel forms should have a larger range of item difficulties than traditional test forms, functionally independent items, unidimensionality within items, and parallel forms should have items that have gone through thorough subject matter expert item reviews (Segall, Moreno, & Hetter, 1997). Because parallel forms test for the same competencies with different items, it is important to have a gradient of item difficulties so that no subset of items on any particular examination are harder or easier than its alternative forms. Likewise, having functionally independent, unidimensional items ensures that examiners are testing for all of the competencies they are trying to assess across test forms.

When developing parallel forms, the best method for creating equivalent exams is to use item-by-item parallelism to ensure that if there is any multidimensionality in the original items, it will be adequately replicated across all forms (Clause et al., 1998). This type of item cloning procedure is the most likely to result in the same features replicated across test forms and has the best outcomes for creating exams that are equivalent across most metrics. However, this technique takes the most time and effort, and is the most difficult to implement and is therefore not recommended when there are limited resources available for parallel forms development. In

these cases of limited resources, it is better to use item-set parallelism because these tests are easier to develop and are still able to meet statistical targets (Leucht, 2003; Samejima, 1977).

When using statistical techniques to create parallel forms, both CTT and IRT are possible options. However, in the majority of cases, IRT will be the preferred method (Zickar & Broadfoot, 2009). It is generally recommended to use IRT because it allows the test maker to focus on the particular range of each construct, conduct goodness-of-fit studies, and utilize other statistical tools to supplement development and analyses (DeMars, 2018; Lord, 1983; Zickar & Broadfoot, 2009). CTT, on the other hand, has several limitations that can cause problems both for interpreting item statistics and person statistics. Fundamentally, CTT is a test-level theory as opposed to IRT– which can give item-level information (DeMars, 2018; Lord, 1983). In CTT, the indices are group dependent, aggregating across items and individuals. This leads to there being issues regarding generalizability across different groups (Lord, 1977).

Item Response Theory can estimate statistics for individual items and individual people, without aggregating across them. When using IRT, the key is to focus on TIF and TRF (Armstrong et al., 1992) to ensure that the same information about the examinee is being gained regardless of test form. Because both of these statistics are calculated by combining across individual item-level statistics, it is important to consider the item information functions and item response functions across the different forms in addition to the test-level statistics of the forms. In item information functions, the properties of items are compared to the variance of the item to see how much information a given item is contributing. These individual item information functions can then be summed together to create the TIF and provide a general overview of the exam. With item response functions (or item characteristic curves), items that are more difficult indicate higher levels of ability in candidates who answer correctly. These item response functions can then be summed or averaged together to create the TRF. Both TIF and TRF can be used to compare different parallel forms to determine how similar items behave across the forms. Because these functions provide information about how items are functioning and how participants are responding, ensuring that these functions are similar across forms helps verify that the parallel forms are effectively equivalent. However, several strong assumptions need to be satisfied before IRT can be used. These include:

- Unidimensional traits for individual items,
- Local independence of items, and
- The ability to mathematically model individual's IRFs.

Additionally, IRT requires large samples in order to estimate properly. As such, IRT has some limitations that could lead to certain scenarios where CTT would be a better statistical approach to parallel form development. Zickar and Broadfoot (2009) recommend that CTT be used when samples are small and the overarching data is multidimensional. When using CTT, the key is to optimize overall test reliability to ensure that test results are consistent across all forms (Armstrong et al., 1992).

9.2.2.1 Improving Parallel Forms to Reduce Item Exposure

Although there are several methods for developing parallel forms, random assignment within each construct being measured is typically the preferred option (Lievens & Sackett, 2007). This method requires that several large item pools are built for each domain being assessed. Then, using these item pools, a predetermined number of items from within each area are randomly

assigned to different test forms. This predetermined number would not have to be the same for each domain, as long as the items provide enough information to determine the applicants' knowledge. Further, the item pool would not necessarily have to be the same size across domains, as long as there were enough unique items available to populate all parallel forms in use. Because this method of parallel form development allows for the simple creation of tests without the need for pretesting, it is one of the simplest ways to create parallel forms. Additionally, because items are randomly assigned from each subject area, multidimensional constructs can still be equally assessed. Finally, this method can account for retest effects because individuals retaking the examination would be assessed on the same overarching topics without seeing any of the same items (Lievens & Sackett, 2007).

Using randomly generated tests allows test developers to control for item exposure while maintaining test integrity and information gathered from testing. Previous work done by PDRI has used this technique to develop parallel forms with high reliability and validity (Bruskiewicz & Lammlein, 1999; Lammlein, Stellmack, Bruskiewicz, & Duehr, 2010). In this work, tests were developed to have the same number of items within each content area being assessed, with a large range of more and less difficult items, with items on each form being representative of the overall item pools.

However, this technique involves creating equivalency at a testlet, or item-set, level. If the goal is the highest level of equivalency between different forms (item-by-item equivalency) other techniques are available. For item-by-item equivalency, parallel forms would be developed utilizing a separate item pool for each item. Due to the time and expense involved in developing parallel forms at that level, item-by-item equivalency becomes progressively more challenging to develop the longer the test, and can be prohibitively expensive to implement on an ongoing basis.

9.2.3 Utilizing CAT

Parallel forms are easier to develop, taking less time, money, and resources than computer based approaches. However, CAT is a significantly better method for limiting item exposure. Although other techniques can be used to prevent large-scale cheating as well as other outcome related issues, CAT is far better at resisting small-scale cheating (Guo et al., 2009) and is much more secure (Barrada et al., 2009). Additionally, CAT accounts and adjusts for item preknowledge because of its adaptive format (Barrada et al., 2009). That is, if examinees start doing well due to preknowledge of some items, they will be confronted with different and more difficult items later in the test that are less likely to have been compromised. In the military context, CAT has been successfully applied to other examinations and has shown several advantages over conventional testing formats (McBride, 1997). Specifically, the CAT-ASVAB had much more precise measurement of aptitudes (Divgi & Mayberry, 1991), with fewer items and shorter test lengths (McBride, 1997; Moreno & Segall, 1997; Moreno et al., 1984; Segall et al., 1997).

There are two stages involved in creating a CAT from a traditional test. The first is score equating development (SED; Segall, 1997). In this step, individuals will take both the paper-and-pencil version and the CAT version of the test in non-operationally motivated conditions. Data gathered from these examinees can then provide an interim method of equating scores between the traditional and CAT versions of the test. The second step is score equating verification (SEV; Segall, 1997). Here, individuals take only one mode of the examination in operationally motivated conditions. Then, score equating is updated based on the verification. This

equipercentile procedure was used successfully for the transformation of the ASVAB from paper-and-pencil testing to CAT (Segall, 1997), and allows for more precise measurement with less item exposure and little difference in overall examinee performance (Moreno et al., 1984; Segall et al., 1997; Wolfe, Moreno, & Segall, 1997).

9.2.3.1 Improving CAT to Reduce Item Exposure

Several techniques exist to detect for item exposure in CAT formats. Although some of these techniques rely solely on observing response times, other methods incorporate statistical elements such as IRT to understand if and when individuals have preknowledge of test items. Overall, when all data are available for examinees, mixture models that incorporate multiple aspects across these techniques are the best at detecting when individuals have item preknowledge (Lee, 2018). Specifically, the MRM-RT, which looks to classify examinees' response behaviors into solution vs. rapid guessing behaviors, has been found to be the most successful in correctly identifying individuals with preknowledge (Lee, 2018). However, when only limited data are available to examiners, it is not appropriate to use mixture models and in these cases, examiners must rely on response time models or item models.

The simplest and most helpful method for test designers to limit item exposure when using CAT is to increase the randomness at the beginning of the test (Barrada et al., 2009; Davis & Dodd, 2003). Traditionally in CAT, all individuals will see the same items at first before the test starts to differentiate and give examinees different items based on individual ability. However, this common practice leads all examinees to be exposed to the same beginning items and increases the overall likelihood of these items being compromised. However, if items of similar difficulty and assessing the same content are randomly selected to appear before different individuals, then the amount to which these beginning items are exposed and at risk for compromise becomes less.

Several techniques exist to control for item exposure within CAT, spanning across randomization approaches (Davis, 2002, 2004; Davis & Dodd, 2005; Revuelta & Ponsoda, 1998; Stocking, 1993), conditional selection approaches (Chang, et al., 2000; Chen, 2010; Chen et al., 1999; Chen & Lei, 2005; Davey & Parshall, 1995; Pastor et al., 2002; Stocking, 1993; Stocking & Lewis, 1995a, 1995b, 1998), stratified approaches (Barrada et al., 2009; Chang & Ying, 1999; Parshall, Harmes, & Kromrey, 2000; Yi & Chang, 2003), and several combinations thereof (Barrada et al., 2009; Chang & Ying, 1999; Eggen, 2001; Georgiadou et al., 2007; Yi, 2002). While each of these approaches offer methods of controlling for item exposure in their own way, none of the direct comparisons between them appear to promote any one approach over the others (Davis & Dodd, 2005). Rather, the newest techniques that look to improve upon the weaknesses in older methodologies often outperform their predecessors. For example, while the SH method was one of the first attempts to control for item exposure issues in CAT, all of the other methods of conditional selection outperform SH because they were designed to improve upon the issues existent within SH to make a better technique for controlling for item exposure (Parshall et al., 1998). As such, the best practice for controlling item exposure within CAT would be to use the newest technique that can be easily implemented within the given context.

9.3 Utilizing and Improving Situational Judgment Tests

Previous work done by PDRI and others has shown that SJTs can help improve testing in several ways. Because they can simulate contextualized, job-related scenarios, SJTs can be used to measure several constructs across practice (Bruskiewicz, et al., 1997; Hanson & Borman, 1989, 1990; Lievens, Peeters, & Schollaert, 2008). These constructs can include leadership skills,

interpersonal skills, managerial skills, teamwork skills, emotional intelligence, judgment, personality, and integrity (Lievens & De Soete, 2015). As such, SJTs have been used in a variety of applications across the Army, Navy, and Air Force (Bruskiewicz, Logan, Hedge, & Hanson, 1997; Legree & Psotka, 2006; Lievens et al., 2008; Ployhart & Weekley, 2006; Schmitt & Chan, 2006). Due to their widespread application, there has been a great deal of research on the various strengths and weaknesses associated with implementing SJTs in the military context.

Some strengths include that SJTs are highly predictive of overall job performance, demonstrate incremental validity over cognitive ability and personality tests, exhibit fewer subgroup differences than other measures, give the appearance of being effective to examinees, flexible to administer, highly stable and reliable, and enable large-scale applicant testing at once (Lievens et al., 2008; Ployhart & Weekley, 2006; Schmitt & Chan, 2006). Some weaknesses of the methodology are that SJTs can be susceptible to faking (examinees giving socially desirable answers rather than honest answers), are vulnerable to practice and coaching effects and lack cross-cultural generalizability. Additionally, due to their multi-faceted nature, there is a question regarding what constructs they are measuring, indicating issues with construct validity (Chan & Schmitt, 2005; Lievens et al., 2008; McDaniel & Nguyen, 2001; McDaniel, Whetzel, Hartman, Nguyen, & Grubb, 2006; Oostrom et al., 2015; Ployhart & Weekley, 2006; Schmitt & Chan, 2006; Whetzel & McDaniel, 2009).

However, in developing and implementing SJTs, choices must be made dependent on the context in which they will be used (Guenole et al., 2017; Krumm, Lievens, Hüffmeier, Lipnevich, Bendels, & Hertel, 2015). For example, factors about the situation and the individual doing the choosing affect implementation. In the military context, a choice that may be acceptable when in active combat may not be acceptable during peacetime. Likewise, a choice that might be considered the best response to a situation for a Captain may not be for a Seaman. However, there are many factors that contribute to the development and implementation of SJTs that are not context-dependent. Regardless of context, SJTs must be developed using a structured process that draws on SME knowledge (Pollard & Cooper-Thomas, 2015). Items should measure job-relevant knowledge and skills, be based in a job analysis and/or another method of capturing requirements and competencies of the job, and utilize SMEs in some capacity to help inform developers regarding the effectiveness of the SJTs (Krumm et al., 2015; Lievens et al., 2008; Pollard & Cooper-Thomas, 2015; Weekley et al., 2006). Finally, after an SJT is implemented, it must be continually monitored over time to track for any changes in scores or validity of the assessment over time (Bruskiewicz et al., 1997). Thus, the same administrative requirements of oversight and review that are required to address item exposure apply here, as well.

9.4 Recommendations Based on Analyses

The results of the study using archival NWAE data described above provide some insight into considerations to keep in mind in planning the way forward.

In analysis 1, a relatively consistent finding across the E-4, E-5, and E-6 results was that the proportion of items that became easier was relatively stable across different lengths of time between administrations, whereas the proportion of items that became harder increased. Given that a common primary concern is that item exposure will result in decreased item difficulty (i.e., items becoming easier), these results do not provide support for modifying the length of time between administrations. While it is not possible to explain the increase in the proportion of items that became harder, potential contributing factors could be the qualitative (e.g., item

content decreasing in relevance over time) or quantitative (e.g., item selection procedures to create test forms). Additionally, there was a higher percentage of items with decreased item-total correlations as the length of the gap between administrations increased for the E-4 and E-5 paygrades, which could be a result of similar factors. We recommend ensuring qualitative and quantitative review and selection factors be used to maintain item quality in form creation.

In analysis 2, the percentage of items that became easier was consistently, significantly higher than the percentage of items that became harder. For the E-4, E-5, and E-6 paygrades, the percentage of items that became easier decreased as the length of time between administrations increased. This may suggest that increasing the length of time between administrations will reduce the likelihood that repeat test-takers recall items they have seen before. However, there are alternative explanations for these results. First, the number of candidates who saw repeat items decreased as the length between gaps increased, which lowers power to detect differences in item parameter changes. Additionally, there may be proficiency differences between candidates who see items after shorter and greater lengths of time (e.g., less proficient candidates may be more likely to be continuing to take the exam after 4 or more administrations since originally seeing an item, and these candidates may be less likely to recall items they have seen before). Thus, the results are not clear indicators that increasing the length of time between administrations would be beneficial.

Analysis 3 indicates that of candidates who exhibit performance changes between repeat and non-repeat items, a significantly greater number perform better on repeat items. This is consistent with expectations and does not provide evidence for altering the testing approach.

9.5 Current Air Force Research

Beyond the present work conducted for the Navy to evaluate the NWAE, research has been done for the Air Force, in conjunction with HumRRO, analyzing the Specialty Knowledge Test (SKT) and the Promotion Fitness Examination (PFE) as part of its Weighted Airman Promotion System (WAPS); (Waugh, Walion, Burgoyne, & McCloy, 2019). That research similarly focused on test security, item exposure, test compromise, and test forensics.

In that effort, the Air Force sought to understand the importance of test security in terms of safeguarding validity, ensuring test fairness, and limiting item compromise (Waugh et al., 2019). Specifically, Waugh and colleagues (2019) investigated the effects lack of test security can have on overall cost, time, and effort, as well as providing recommendations for mitigating negative impacts. Overall, the Air Force research provided three key recommendations for protecting test security:

- Explain security risks and procedures to those involved in test development
- Limit the sharing of information between examinees
- Regularly conduct security audits of the testing materials and practices

In constructing and administering exams, organizations must be careful about how and when items are used and the impact on test security. Limiting opportunities for item compromise to occur is critical. This can be done by designing exams to have fewer security risks by using short testing windows, electronic exams, multiple test forms, and randomization of item content. Further, limiting the ability of examinees to share information or otherwise gain preknowledge of items can be achieved by banning cell phones and other technology from the testing area, formally training test proctors on behaviors to watch for, securing test materials, and prohibiting

the sharing of item content. Lastly, it is important to use methods of tracking exam security over its use by utilizing forensic statistics and analyzing changes in item difficulty over time.

Similarly, research was done for the Air Force, in conjunction with HumRRO, analyzing the application of SJTs to the overall WAPS process as a means of augmenting and supporting their current system (Sullivan, Whetzel, & McCloy, 2019). They discussed general guidelines and reviewed best practice recommendations for doing so. As in the current report, they reviewed guidelines for developing scenarios and response options, selecting appropriate response instructions, understanding different response formats, and utilizing various scoring approaches. Sullivan and colleagues (2019) emphasized the importance of using SMEs throughout each portion of the SJT development process, and specifically point to the necessity of obtaining diverse perspectives from individuals with operational experience with the content being assessed. Additionally, they noted that it is difficult to assess the reliability of SJTs due to their multidimensional nature and that SJTs have high concurrent and predictive validity for both selection and promotion within the military. They also discussed methods to improve the use of SJTs, including reducing adverse impact, faking, and the effects of coaching.

10.0 SUMMARY

10.1 Literature Review

When using exams and test items over a period of time, it is critical to maintain awareness of how frequently items are being exposed to test takers and to understand what effects this can have on the accuracy, validity, and fairness of the testing program. Several issues can arise, such as when individual items become compromised in that they are over-utilized and become known to test takers. Another is identifying individuals who have preknowledge of items on an exam. Further, the combination of these two issues is critical; identifying individuals who have preknowledge and of which items. Further, examiners must make efforts to detect groups of individuals working together to gain preknowledge of exam items.

Several forensic tools can be used to detect item exposure as defined in these categories. Strategies for detecting item compromise include response pattern modeling, response-time modeling, speed/ability distributions, item compromise probabilities, and utilizing IRT. All of these strategies fall into one of two broad types of methods for detecting these issues. The first involves careful monitoring of item exposure rates over time and then using that knowledge in conjunction with test taker performance over time. However, these patterning and modeling approaches are often deficient in some way and fail to capture all of the various idiosyncrasies involved with item compromise and preknowledge. There can be a variety of reasons why performance, speed, or any number of other factors used in these methods are changing over time, and these reasons may not necessarily have anything to do with item exposure rates. The other type of method for detecting item exposure involves the use of statistical techniques that look for aberrations in responding. However, these methods are mostly, if not always, performed ad hoc and cannot consistently account for the cause of the aberrations. Rather, these methods only tell the examiners that there is an issue rather than of what is causing it. As a result, both types of methods for detecting issues with item exposure are deficient when used alone. However, when used in combination, they can often overcome their individual deficiencies. Consequently, examiners should employ multiple forensic tools to maximize their effectiveness in detecting item exposure issues when present.

In addition to detecting item exposure after it occurs, there are techniques to preclude the resulting negative effects altogether. The most common is parallel forms, as having different versions of the same exam content limits how often items are used and therefore exposed. There are several means by which parallel forms can be created. While strictly parallel tests with item-level equivalence are ideal, they are simply too difficult and resource-intensive to create. In the majority of cases, weakly parallel tests created at the test level using IRT prevail as being the most efficient means of creating parallel forms. The key to creating these forms is to have a gradient of item difficulties that appropriately discriminate between individuals of differing abilities so that no subset of items on any particular examination are harder or easier than its alternative forms, ideally within a range between .25 or .75 for level of difficulty. Likewise, having functionally independent, unidimensional items ensures that examiners are testing for all of the competencies they are hoping to assess across test forms. Although there are several methods of developing parallel forms, random assignment within each construct being measured is considered best for limiting item exposure.

Parallel forms is the more commonly used method, but computer adaptive approaches are significantly more effective in limiting item exposure. Because CAT takes into account

individual ability level and adjusts the test accordingly for each individual such that all test takers are presented with functionally different items, exposure is significantly minimized. However, CAT is not in itself a complete solution for remedying item exposure. In the most basic versions of CAT, all examinees are presented with the same first item(s). This issue can be remedied simply by increasing randomness at the beginning of the exam. Other issues can be fixed through a variety of item selection and randomization techniques that are constantly evolving and changing to adapt and overcome the issues of their predecessors. As such, the newest of these techniques are generally the most recommended to alleviate limitations that arise in CAT.

In addition to the item exposure research summarized above, we also provided a review of how SJTs can be used to assess candidate's abilities. In SJTs, a number of choices need to be made in their development and implementation. Regardless of context, SJTs should be developed using a structured process that draws on SME knowledge. Items should measure job-relevant knowledge and skills, be based in a job analysis and/or another method of capturing requirements and competencies of the job, and utilize SMEs in some capacity to help inform developers regarding the effectiveness of the SJTs.

Following our review of the literature, we analyzed data from eight E-4/5/6 administrations (March 2015 – September 2018) and four E-7 administrations (January 2015 – January 2018) for the NWAEs to evaluate the effects of item exposure on item effectiveness and Sailor performance. Three primary analyses were conducted, as summarized below. Although the report groups analyses by grade and then type of analysis within grade in the report above, here, for clarity, analyses are grouped by type of analysis and then by grade within analysis.

10.2 Analyses of NWAE Data

To identify possible effects of item exposure under the current testing system, we conducted three sets of analyses.

10.2.1 Analysis 1

For the first analysis, we evaluated item parameter changes over time, looking specifically at whether items changed in difficulty over time and whether item-total correlations changed over time. Analyses were conducted across different time gaps (2 series, 3 series, and 4+ series) and were split by whether the items had or had not been administered prior to 2015.

Overall, the majority of items, across all paygrades, did not exhibit significant changes in difficulty regardless of the length of time between administrations. There were, however, some modest differences within each grade, which are highlighted below.

- **E-4:** As the length of time between administrations increased, the percentage of items that became significantly harder increased, with significantly greater numbers of harder items after 3-series and 4 or more series gaps. This suggests that test takers within this grade had more difficulty with items presented on the exam the longer the period of time had passed.
- **E-5:** After a 2-series gap, the percentage of items that became easier was higher than the percentage of items that became harder. The percentage of items that became harder increased as the length of time between administrations increased, and the percentage of items that became harder was significantly greater after a 4-series or larger gap. Thus,

while a short duration between exams led to test takers within this grade having an easier time with items, longer durations led to test takers struggling more with items.

- **E-6:** The percentage of items that became harder increased as the length of time between administrations increased. After a 2-series or 3-series gap, the percentage of items that became easier was significantly higher than the percentage of items that became harder, though the difference was non-significant after a 4 or greater series gap. This suggests that test takers within this grade had more ease with items presented on the exam after time had passed, though this ease fades the longer the duration. For items administered prior to 2015, only the 4 or more series gap exhibited a significant difference (a greater number of harder items), whereas for items not administered prior to 2015, there was significantly more easier items regardless of the length of time between administrations. This indicates that there is a difference between how test takers performed on these different types of items, with the newer items giving greater ease to test takers regardless of length of time.
- **E-4/5/6 (combining across E-4, E-5 and E-6):** The percentage of items that become easier was relatively stable as the time between administrations increased. After a 2-series gap, the percentage of items that became easier was higher than the percentage of items that became harder. The percentage of items that became harder increased as the length of time between administrations increased, and the percentage of items that become harder was significantly greater after a 3-series or 4-series or larger gap. Thus, although a short duration between exams led to test takers within these combined grades having an easier time with items, longer durations led to test takers struggling more with items.
- **E-7:** A higher percentage of items became easier than became harder regardless of the length of time between administrations. However, this was non-significant.

Similar to difficulty changes, most items across all paygrades did not exhibit significant item-total correlation changes over administrations regardless of grade. Once again, there were some modest differences within each grade.

- **E-4:** For items administered prior to 2015, the proportion of items with decreased item-total correlations increased as the length of time between administrations increased, but the pattern was less clear for more recent items not administered prior to 2015. This indicates that there was a stronger difference between how test takers within this grade performed on the older items the longer the duration for item use was.
- **E-5:** For items administered prior to 2015, there was a significantly greater number of items exhibiting increased than decreased item-total correlations after a 2-series or 3-series gap. However, there was a non-significant difference after a 4 or more series gap. This means that there was a weaker difference between how test takers within this grade performed on the older items for short durations between item use, but not for longer durations. For items not administered prior to 2015, there was a significantly greater number of items exhibiting increased than decreased item-total correlations regardless of the length of time between administrations. Thus, there was a weak difference between how test takers within this grade performed on the newer items regardless of time.
- **E-6:** For items administered prior to 2015, there was a significantly greater number of items exhibiting increased than decreased item-total correlations after a 4- series or larger

gap, but there was a non-significant difference after a 2-series or 3-series gap. This indicates that there was a weaker difference between how test takers within this grade performed on the older items only when the duration between use if there was 4 or greater series. For items not administered prior to 2015, there was a significantly greater number of items exhibiting increased than decreased item-total correlations after a 3-series or 4 or more series gap. This indicates that there was a weaker difference between how test takers within this grade performed on the older items the longer the duration for item use was.

- **E-4/5/6:** A greater percentage of items not administered prior to 2015 exhibited correlation increases than items administered prior to 2015. Across these grades, older items had a stronger difference between how test takers performed on them than newer items.
- **E-7:** There were no strong differences from the overall pattern. Item-total correlations were fairly stable across administrations.

Across all paygrades, the majority of items did not exhibit significant changes in difficulty nor in item-total correlations regardless of the length of time between administrations. Although there were some differences across grades, these were mostly non-significant.

10.2.2 Analysis 2

For the second analysis, we evaluated item parameter changes for repeat test-takers to see if those who took more than two exams within the rating in the scope of the study had changes in performance that could be accounted for by item exposure. Analyses were once again conducted across different time gaps (2 series, 3 series, and 4+ series).

Overall, the percentage of items that became easier for repeat test-takers was significantly greater than the percentage of items that became harder for all paygrades.

- **E-4:** The percentage of items that became harder was relatively consistent across the different intervals between administrations, but the percentage of items that became easier decreased as the time interval between administrations increased.
- **E-5:** The percentage of items that became harder and the percentage of items that became easier decreased as the time interval between administrations increased.
- **E-6:** Overall, over a third of items became easier after a 2-series gap, but this decreased to just under a third of items after a 4-series gap.
- **E-4/5/6:** Similar to the pattern for E-6, over a third of items became easier after a 2-series gap, but this decreased to approximately one quarter of items after a 4-series gap.
- **E-7:** The percentage of items that became easier increased as the time interval between administrations increased, while the percentage of items that became harder decreased as the interval between administrations increased.

Just as with difficulty, the percentage of items that showed an increased item-total correlation for repeat test-takers was significantly greater than the percentage of items that showed a decreased item-total correlation for all pay grades with some minor exceptions.

- **E-4:** There were no strong differences from the overall pattern.

- **E-5:** The percentages of items for which correlations increased declined as the length of time between administrations increased.
- **E-6:** The percentage of items that exhibited an increased item-total correlation for repeat test-takers was significantly greater than the percentage of items that exhibited a decreased item-total correlation overall, but not for all demographic groups.
- **E-4/5/6**: There were no strong differences from the overall pattern.
- **E-7:** There were no strong differences from the overall pattern.

Across all paygrades, the percentage of items that became easier and the item-total correlation increases for repeat test-takers were significantly greater than the percentage of items that became harder and the item-total correlation decreases. While there were some differences across grades, the majority were non-significant with there being almost no differences across grades for item-total correlations.

10.2.3 Analysis 3

Analysis 3 examined whether there were differences across candidate performance for initial item exposure compared to repeat exposures. Overall, most candidates across paygrades did not perform significantly better or worse on repeat items compared to non-repeat items. Where there were differences, a significantly greater number of candidates performed better on repeat items than on non-repeat items overall across all demographic groups.

10.3 Implications of Changes to Item Exposure Rules

There are a variety of findings in these results, and a number of interpretations. Unfortunately, the lack of consistency creates some limitations to their applicability for driving changes to current practices. Absent more specific information regarding current practices and the availability of resources to make changes, as well as the possibilities that could be realistically implemented, it is difficult to make specific recommendations. We explore this further below.

For the first analysis, item parameter changes over time were examined and across different time gaps (2 series gap, 3 series gap, and 4+ series gap) and paygrades, the majority of items did not exhibit significant changes in difficulty or item-total correlation regardless of the time between administrations. However, the proportion of items that became easier was relatively stable compared to the proportion that became harder, which increased. As such, further research is needed to better understand the reason for the difficulty changes across time gaps.

For the second analysis, item parameter changes were examined across repeat test-takers to determine whether individuals who took more than two exams within a given rating had changes in performance that could be accounted for by item exposure. Across different time gaps (2 series gap, 3 series gap, and 4+ series gap), the percentage of items that became easier and had increased item-total correlations was significantly greater than the percentage of items that became harder and had decreased item-total correlations for repeat test-takers. Although these findings indicate a possible relationship between difficulty and repeat test-taking as well as between item-total correlation changes and repeat test-taking, more experimentation would be required to understand exactly what is causing these relationships and how to best build this into future test design.

The third analysis showed that those candidates who exhibited performance changes between repeat and non-repeat items performed significantly better on repeat items. Just as with the first

and second analyses, further research is needed to understand the specific reasons behind this finding and how to best use this information in the NWAE testing program. While the results of the analyses provide novel information regarding overall NWAE testing and possible effects of item exposure, they do not provide clear, readily actionable support for making changes to the testing program. Additional research is needed to determine the specifics of why the repeated items lead to improved performance, beyond intuiting that simply having seen the items before yields improvement. This would include more focused individual level analyses of item performance and experimental designs of evaluations of items and test forms going forward. Such analyses would yield stronger conclusions and possibly more specific guidance regarding best practices going forward.

10.4 Conclusion

Several tools are available to detect possible negative effects of item exposure. The key is to use multiple techniques over time to ensure consistent and reliable information that accurately identifies, and can be used to address, those effects. Further, there are many methods of controlling for item exposure, such as modifications to tests and test banks, the addition of alternate forms, and changes to the testing environment. Each of these control methods have different limitations and benefits, so the selection of which to use must be carefully considered in the broader context of the overall testing process. Additionally, alternative methods of testing, such as SJTs, can be used to enhance and add incremental information to the testing process due to their ability to capture realistic examples of potential future performance in a reliable manner.

In the archival NWAE data analyzed in the present study, it was initially anticipated that items would become easier over time. However, data analyses indicated a much more complicated picture.

In the first analysis, a consistent finding was that the proportion of items that became easier was relatively stable across different lengths of time between administrations, whereas the proportion of items that became harder increased. In the second analysis, the percentage of items that became easier was consistently, significantly higher than the percentage of items that became harder. Further, the percentage of items that became easier decreased as the length of time between administrations increased. The third analysis was consistent with expectations and showed that of candidates who exhibited performance changes between repeat and non-repeat items, a significantly greater number of them performed better on repeat items. There are several interpretations of these results, as discussed above, and while they do provide some unique insight, they do not provide clear, readily actionable support for making simple changes to the testing program such as increasing the length of time between administrations. Further research may yield greater insights into the specifics of what drives some of the findings in the archival data.

Overall, the work compiled in this report underscores the need to be particularly careful about consistently analyzing item performance and ensuring that trends in the data do not change in unexplainable ways over time. If shifts are identified, there are several tools available to investigate what may be causing those changes, and solutions are available. However, the application of the appropriate tool is not always straightforward, and may have unanticipated secondary effects. Therefore, any actions taken must be considered in the context of the overall testing program to ensure fair, valid, and accurate testing.

11.0 REFERENCES

Ackerman, T. A. (1989). An alternative methodology for creating parallel test forms using the IRT information function. *Paper presented at the 1989 annual meeting of the National Council on Measurement in Education, San Francisco, CA.* Retrieved from ERIC database. (ED306279)

Adema, J. J. (1992). Methods and models for the construction of weakly parallel tests. *Applied Psychological Measurement*, *16*(1), 53-63.

Alf, E. F., & Stapleton, J. W. (1981). *Postenlistment Mental Qualification Verification: Calendar Year 1979* (No. NPRDC-SR-82-9). Navy Personnel Research and Development Center San Diego CA.

Algina, J., & Penfield, R. D. (2009). Classical test theory. *The Sage handbook of quantitative methods in psychology*, 93-122.

Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical Care*, I7-I16.

Angoff, W. H. (1957). The "equating" of non-parallel tests. *The Journal of Experimental Education*, *25*(3), 241-247.

Ariel, A., Veldkamp, B. P., & Breithaupt, K. (2006). Optimal testlet pool assembly for multistage testing designs. *Applied Psychological Measurement*, *30*(3), 204-215.

Armstrong, R. D., Jones, D. H., & Wang, Z. (1994). Automated parallel test construction using classical test theory. *Journal of Educational Statistics*, *19*(1), 73-90.

Armstrong, R. D., Jones, D. H., & Wu, L. (1992). An automated test development of parallel tests from a seed test. *Psychometrika*, *57*(2), 271-288.

Ayhan, A. S. (2015). Comparability of scores from cat and paper and pencil implementations of student selection examination to higher education (Doctoral dissertation, Bilkent University).

Barrada, J., Olea, J., Ponsada, V., Abad, F., Ponsoda, V., & Abad, F. J. (2009). Test overlap rate and item exposure rate as indicators of test security in CATs. In *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved from www.psych.umn.edu/psylabs/CATCentral*.

Belov, D. I. (2012). Detection of Large-Scale Item Preknowledge in Computerized Adaptive Testing via Kullback–Leibler Divergence. (Report No. 12-01). Newtown, PA: Law School Admission Council, Inc.

Belov, D. I. (2014). Detecting item preknowledge in computerized adaptive testing using information theory and combinatorial optimization. *Journal of Computerized Adaptive Testing*, *2*(3), 37-58.

Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, *40*(2), 83-97.

Belov, D. I. (2017). Identification of item preknowledge by the methods of information theory and combinatorial optimization. *Handbook of quantitative methods for detecting cheating on tests*, 164-176.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, *14*(3), 223-235.

Bobko, P., Roth, P. L., & Buster, M. A. (2007). The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis. *Organizational Research Methods*, *10*(4), 689-709.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational Statistics*, *15*(2), 129-145.

Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, *53*(3), 225-250.

Boughton, K. A., Smith, J., & Ren, H. (2017). Using response time data to detect compromised items and/or people. *Handbook of quantitative methods for detecting cheating on tests*, 177-192.

Brown, J. D. (2013). Classical theory reliability. *The companion to language assessment*, *3*, 1165-1181.

Bruskiewicz, K. T., & Lammlein, S. E. (1999). *OCC job knowledge test development and validation report* (Institute Report #333). Minneapolis, MN: Personnel Decisions Research Institutes, Inc.

Bruskiewicz, K. T., Logan, K. K., Hedge, J. W., & Hanson, M. A. (1997). Annotated Bibliography of Research Relevant to the Development and Validation of the Situational Test of Aircrew Response Styles Inventory (No. AL/HR-TR-1997-0015). Tampa, FL: Personnel Decisions Research Institutes, Inc.

Campion, M. C., Ployhart, R. E., & MacKenzie Jr, W. I. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance*, *27*(4), 283-310.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment* (Vol. 17). Sage publications.

Cascio, W. F. & Aguinis, H. (2011). *Applied psychology in human resource management,* 7th ed. Upper Saddle River, NJ: Prentice Hall.

Chan, D. and Schmitt, N. (2002) Situational judgment and job performance. *Human Performance,* 15, 233–254.

Chan, D. and Schmitt, N. (2005) Situational judgment tests. In A. Evers, N. Anderson and O. Voskuijil (Eds), Handbook of personnel selection (pp. 219–246). Oxford: Blackwell.

Chang, S. W., Ansley, T. N., & Lin, S. H. (2000). Performance of Item Exposure Control Methods in Computerized Adaptive Testing: Further Explorations. *Paper presented at the 2000 annual meeting of the American Educational Research Association, New Orleans, LA.* Retrieved from ERIC database. (ED442837)

Chang, S. W., & Twu, B. Y. (1998). A Comparative Study of Item Exposure Control Methods in Computerized Adaptive Testing. *Paper presented at the 1998 annual meeting of the*

American Educational Research Association, San Diego, CA. Retrieved from ERIC database. (ED420722)

Chang, H. H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 211-222.

Chen, S. Y. (2010). A procedure for controlling general test overlap in computerized adaptive testing. *Applied Psychological Measurement, 34*(6), 393-409.

Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (1999). Exploring the Relationship between Item Exposure Rate and Test Overlap Rate in Computerized Adaptive Testing. (Report No. ACT-RR-99-5). *ACT Report Series, Iowa City, IA.* Retrieved from ERIC database. (ED435643)

Chen, S. Y., & Lei, P. W. (2005). Controlling item exposure and test overlap in computerized adaptive testing. *Applied Psychological Measurement, 29*(3), 204-217.

Choe, E. M., Zhang, J., & Chang, H. H. (2017). Sequential Detection of Compromised Items Using Response Times in Computerized Adaptive Testing. *Psychometrika*, 1-24.

Clause, C. S., Mullins, M. E., Nee, M. T., Pulakos, E., & Schmitt, N. (1998). Parallel test form development: A procedure for alternate predictors and an example. *Personnel Psychology*, *51*(1), 193-208.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*(3), 410.

Corstjens, J., Lievens, F., & Krumm, S. (2017). Situational judgement tests for selection. *The Wiley Blackwell handbook of the psychology of recruitment, selection and retention*, 226-246.

Davey, T., & Parshall, C. G. (1995). New algorithms for item selection and exposure control with computerized adaptive testing. *Paper presented at the 1995 annual meeting of the American Educational Research Association, San Francisco, CA.* Retrieved from ERIC database. (ED421525)

Davis, L. L. (2002). Strategies for controlling item exposure in computerized adaptive testing with polytomously scored items (Doctoral dissertation). Austin, TX: University of Texas at Austin.

Davis, L. L. (2004). Strategies for controlling item exposure in computerized adaptive testing with the generalized partial credit model. *Applied Psychological Measurement, 28*(3), 165-185.

Davis, L. L., & Dodd, B. G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement*, *27*(5), 335-356.

Davis, L. L., & Dodd, B. G. (2005). Strategies for controlling item exposure in computerized adaptive testing with the partial credit model. *Journal of Applied Measurement, 9*(1), 1.

DeMars, C. E. (2018). Classical Test Theory and Item Response Theory. The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development, 49-73.

Divgi, D. R., & Mayberry, P. W. (1991). *An Analysis of CAT-ASVAB Scores in the Marine Corps JPM Data* (No. CRM-91-161). Center for Naval Analyses. Alexandria, VA: Operations and Support DIV.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67-86.

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). Development of the tailored adaptive personality assessment system (TAPAS) to support army personnel selection and classification decisions. Drasgow Consulting Group. Urbana, IL.

Eckerly, C. A. (2016a). Detecting Preknowledge and Item Compromise. *Handbook of Quantitative Methods for Detecting Cheating on Tests*, 101.

Eckerly, C. (2016b). Increasing the Robustness of the Deterministic Gated IRT Model: Using a Scale-Purified Approach to Identify Compromised Items and Examinees with Item Preknowledge. The University of Wisconsin-Madison.

Eggen, T. J. H. M. (2001). Overexposure and underexposure of items in computerized adaptive testing. *Measurement and Research Department Reports, 1*.

Eignor, D. R. (1993). Deriving comparable scores for computer adaptive and P&P tests: An example using the SAT. Princeton, NJ: Educational Testing Service.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. Educational and Psychological Measurement,58(3).357-382

Feinberg, R. A., Raymond, M. R., & Haist, S. A. (2015). Repeat testing effects on credentialing exams: are repeaters misinformed or uninformed?. *Educational Measurement: Issues and Practice*, *34*(1), 34-39.

Fisher, J. M. (2018). An Analysis of Paper-Based Assessment vs. Computer-Based Assessment: A Quantitative Inferential Research Study (Doctoral dissertation, Northcentral University).

Flanagan, J. C. (1954). The critical incident technique. *Psychological bulletin, 51*(4), 327.

Folk, V. G., & Smith, R. L. (2002). Models for delivery of CBTs. *Computer-based testing: Building the foundation for future assessments*, 41-66.

Georgiadou, E. G., Triantafillou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning and Assessment, 5*(8).

Geving, A. M., Webb, S., & Davis, B. (2005). Opportunities for repeat testing: Practice doesn't always make perfect. *Applied HRM Research, 10*(2), 47-56.

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.

Giordano, C., Subhiyah, R., & Hess, B. (2005). An Analysis of Item Exposure and Item Parameter Drift on a Take-Home Recertification Exam. *Paper presented at the 2005 annual meeting of the American Educational Research Association, Montreal, Canada.*

Guenole, N., Chernyshenko, O. S., & Weekly, J. (2017). On Designing Construct Driven Situational Judgment Tests: Some Preliminary Recommendations. *International Journal of Testing, 17*(3), 234-252.

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.

Haley, S. M., Coster, W. J., Andres, P. L., Kosinski, M., & Ni, P. (2004). Score comparability of short forms and computerized adaptive testing: simulation study with the activity measure for post-acute care 1. *Archives of physical medicine and rehabilitation, 85*(4), 661-666.

Han, N. (2003). Using moving averages to assess test and item security in computer based testing. *Center for Educational Assessment Research Report*, (468).

Han, N., & Hambleton, R. (2007). Using moving averages to detect exposed test items in computer-based testing. *Real Data Analysis*, 277.

Hanson, M. A., & Borman, W. C. (1990). A situational judgment test of supervisory knowledge in the US Army. In *32nd Annual Conference of the Military Testing Association, Orange Beach, AL*.

Hanson, M. A., & Borman, W. C. (1989). Development of a situational judgment test to be used as a job performance measure for first line supervisors in the US Army. In *4th annual conference of the Society for Industrial and Organizational Psychology, Boston, MA*.

Hertz, N. R., & Chinn, R. N. (2003). Effects of Item Exposure for Conventional Examinations in a Continuous Testing Environment. *Paper presented at the 2003 annual meeting of the National Council on Measurement in Education, Chicago, IL*. Retrieved from ERIC database. (ED476722)

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), Computerized Adaptive Testing: From Inquiry to Operation (pp. 141–144). Washington, DC: American Psychological Association.

Hwang, D. Y. (2002). Classical Test Theory and Item Response Theory: Analytical and Empirical Comparisons. *Paper presented at the 2002 annual meeting of the Southwest Educational Research Association, Austin, TX*. Retrieved from ERIC database. (ED466799)

Kantrowitz, T. M., Gutierrez, S. (2013). The Security of Employment Testing: Practices That Keep Pace with Evolving Organizational Demands and Technology Innovations. *The Industrial-Organizational Psychologist, 50(4),* 33–42.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277-298.

Kolen, M. J., & Tong, Y. (2005). Classical Test Score Equating. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (Vol. 1, pp. 282-287). West Sussex, UK: John Wiley & Sons, Ltd.

Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in situational judgment tests?. *Journal of Applied Psychology, 100*(2), 399.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79-86.

Lammlein, S. E., Stellmack, A. L., Bruskiewicz, K. T., & Duehr, E. E. (2010). *Job analysis and validation study for OCC entry-level selection measures* (Institute Report #674). Minneapolis, MN: Personnel Decisions Research Institutes, Inc.

Lee, S. Y. (2018). *A Mixture Model Approach to Detect Examinees with Item Preknowledge* (Doctoral dissertation, The University of Wisconsin-Madison).

Legree, P. J., & Psotka, J. (2006). *Refining situational judgment test methods*. Army Research Institute for the Behavioral and Social Sciences Arlington VA.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of educational statistics*, *4*(4), 269-290.

Lievens, F., & De Soete, B. (2015). Situational judgment tests. In *International encyclopedia of the social & behavioral sciences* (Vol. 22, pp. 13-19). Elsevier.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, *37*(4), 426-441.

Lievens, F., & Sackett, P. R. (2007). Situational judgment tests in high-stakes settings: Issues and strategies with generating alternate forms. *Journal of Applied Psychology*, *92*(4), 1043.

Lim, E. Y. (2011). *The effectiveness of using multiple item pools to increase test security in computerized adaptive testing*. (Doctoral dissertation). Urbana-Champaign, IL: University of Illinois at Urbana-Champaign.

Lin, C. J. (2008). Comparisons between classical test theory and item response theory in automated assembly of parallel test forms. *The Journal of Technology, Learning and Assessment*, *6*(8).

Livingston, S. A. (1972). Criterion-Referenced Applications of Classical Test Theory 1, 2. *Journal of Educational Measurement*, *9*(1), 13-26.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, *14*(2), 117-138.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, *48*(2), 233-245.

Luecht, R. M. (2003). Exposure Control Using Adaptive Multi-Stage Item Bundles. *Paper presented at the 2003 annual meeting of the National Council on Measurement in Education, Chicago, IL*. Retrieved from ERIC database. (ED475831)

McBride, J. R. (1997). Research antecedents of applied adaptive testing. In W.A. Sands, B.K. Waters, and J.R. McBride. (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 47-57). American Psychological Association.

McDaniel, M. A., Hartman, N. S., & Grubb III, W. L. (2003). Situational judgment tests, knowledge, behavioral tendency, and validity: A meta-analysis. *Paper presented at 18th Annual Conference of the Society for Industrial and Organizational Psychology. Orlando, FL*.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, *86*(4), 730.

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment*, *9*(1-2), 103-113.

McDaniel, M. A., Whetzel, D. L., Hartman, N. S., Nguyen, N. T., & Grubb, W. L. (2006). Situational judgment tests: Validity and an integrative model. In J.A. Weeklet & R.E. Ployhart (Eds.) *Situational judgment tests: Theory, measurement, and application*, (pp. 183-204).

McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, *27*(2), 121-137.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and pencil cognitive ability tests: a meta-analysis. *Psychological Bulletin*, 114(3), 449-458.

Mead, A. D., & Meade, A. W. (2010). Test construction using CTT and IRT with unrepresentative samples. *Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology in Atlanta, GA* (Vol. 56).

Moreno, K. E. (1997). CAT-ASVAB operational test and evaluation. In W.A. Sands, B.K. Waters, and J.R. McBride. (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 199-205). American Psychological Association.

Moreno, K. E., & Segall, D. O. (1997). Reliability and construct validity of CAT-ASVAB. In W.A. Sands, B.K. Waters, and J.R. McBride. (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 169-174). American Psychological Association.

Moreno, K. E., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests. *Applied Psychological Measurement*, 8, 155-163.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, *75*(6), 640.

Muñiz, J. (2005). Classical Test Models. Encyclopedia of Statistics in Behavioral Science.

Murphy, K. R. (2002). Can conflicting perspectives on the role of g in personnel selection be resolved?. *Human Performance*, *15*(1-2), 173-186.

Nye, C. D., Drasgow, F., Chernyshenko, O. S., Stark, S., Kubisiak, U. C., White, L. A., & Jose, I. (2012). *Assessing the tailored adaptive personality assessment system (TAPAS) as an MOS qualification instrument* (No. ARI-TR-1312). Personnel Decisions Research Institutes (PDRI) Inc. Tampa, FL.

O'Leary, L. S., & Smith, R. W. (2017). Detecting candidate preknowledge and compromised content using differential person and item functioning. *Handbook of quantitative methods for detecting cheating on tests*, 151-163.

O'Neill, T. R., Sun, L., Peabody, M. R., & Royal, K. D. (2015). The impact of repeated exposure to items. *Teaching and learning in medicine*, *27*(4), 404-409.

Oostrom, J. K., De Soete, B., & Lievens, F. (2015). Situational judgment testing: A review and some new developments. In I. Nikolaou, & J. K. Oostrom (Eds.) *Employee recruitment, selection, and assessment: Contemporary issues for theory and practice* (pp. 172-189). Sussex, UK: Psychology Press.Oswald, F. L., Shaw, A., & Farmer, W. L. (2015). Comparing simple scoring with IRT scoring of personality measures: The Navy Computer Adaptive Personality Scales. *Applied Psychological Measurement*, *39*(2), 144-154.

Paek, P. L. (2005). Recent trends in comparability studies using testing and assessment to promote learning. *Pearson Educational Measurement.* Retrieved from http://www.pearsonedmeasurement.com/research/research.htm

Parshall, C. G., Davey, T., & Nering, M. L. (1998). Test development exposure control for adaptive testing. *Paper presented at the 2002 annual meeting of the National Council on Measurement in Education, San Diego, CA.* Retrieved from ERIC database. (ED421526)

Parshall, C., Harmes, J. C., & Kromrey, J. D. (2000). Item exposure control in computer-adaptive testing: The use of freezing to augment stratification. *Florida Journal of Educational Research*, *40*(1), 28-52.

Pastor, D. A., Dodd, B. G., & Chang, H. H. (2002). A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model. *Applied Psychological Measurement*, *26*(2), 147-163.

Ployhart, R. E., & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment*, *11*(1), 1-16.

Ployhart, R. E., & Weekley, J. A. (2006). Situational judgment: Some suggestions for future science and practice. In J.A. Weeklet & R.E. Ployhart (Eds.) *Situational judgment tests: Theory, measurement, and application*, (pp. 345-350).Pollard, S., & Cooper-Thomas, H. D. (2015). Best Practice Recommendations for Situational Judgment Tests. *The Australasian Journal of Organisational Psychology*, *8(E7), 1-10*.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. Journal of Technology, Learning, and Assessment, 2(6). Available from http://www.jtla.org.

Pulakos, E. D. (2005). Selection assessment methods. United stated of America: Society for Human Resource Management (SHRM) Foundation.

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using Response Time to Detect Item Preknowledge in Computer-Based Licensure Examinations. *Educational Measurement: Issues and Practice*, *35*(1), 38-47.

Rasch, G. (1960). Studies in mathematical psychology: I. *Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.

Raymond, M. R., Kahraman, N., Swygert, K. A., & Balog, K. P. Scores Gains on Performance Tests for Repeat Examinees: An Evaluation of Construct and Criterion-Related Evidence. *Paper Presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.*

Raymond, M. R., Neustel, S., & Anderson, D. (2007). Retest effects on identical and parallel forms in certification and licensure testing. *Personnel Psychology, 60*(2), 367-396.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(4), 311-327.

Robin, F. (2005). Item Exposure. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (Vol. 2, pp. 974-978). West Sussex, UK: John Wiley & Sons, Ltd.

Sanders, P. F., & Verschoor, A. J. (1998). Parallel test construction using classical item parameters. *Applied Psychological Measurement, 22*(3), 212-223.

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika*, *42*(2), 193-198.

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct. In J.A. Weeklet & R.E. Ployhart (Eds.) *Situational judgment tests: Theory, measurement, and application*, (pp. 135-155).Segall, D. 0.(1997). Equating the CAT-ASVAB. In W.A. Sands, B.K. Waters, and J.R. McBride. (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 181-198). American Psychological Association.

Segall, D. O. (2002). An item response model for characterizing test compromise. *Journal of Educational and Behavioral Statistics*, *27*(2), 163-179.

Segall, D. O., Moreno, K. E., & Hetter, R. D. (1997). Item pool development and evaluation. In W.A. Sands, B.K. Waters, and J.R. McBride. (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 117-130). American Psychological Association.

Segall, D. O., Moreno, K. E., Kieckhaefer, W. F., Vicino, F. L., & McBride, J. R. (1997). Equating the CAT-ASVAB. In W.A. Sands, B.K. Waters, and J.R. McBride. (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 181-198). American Psychological Association.

Shu, Z., Henson, R., & Luecht, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika*, *78*(3), 481-497.

Sinharay, S. (2017). Which Statistic Should Be Used to Detect Item Preknowledge When the Set of Compromised Items Is Known?. *Applied Psychological Measurement*, *41*(6), 403-421.

Society for Industrial and Organizational Psychology (SIOP). (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Smith, R. M. (1990). Theory and practice of fit. *Rasch Measurement Transactions*, *3*(4), 78.

Stage, C. (2003). Classical test theory or item response theory: The Swedish experience (Vol. 42). *Umeii, Sweden: Umeii Universitet Department of Educational Measurement.*

Stevens, M. J., & Campion, M. A. (1999). Staffing work teams: Development and validation of a selection test for teamwork settings. *Journal of Management, 25*(2), 207-228.

Stocking, M. L. (1993). Controlling item exposure rates in a realistic adaptive testing paradigm. *ETS Research Report Series, 1993*(1), i-31.

Stocking, M. L., & Lewis, C. (1995a). A new method of controlling item exposure in computerized adaptive testing. *ETS Research Report Series, 1995*(2), i-29.

Stocking, M. L., & Lewis, C. (1995b). Controlling item exposure conditional on ability in computerized adaptive testing. *ETS Research Report Series, 1995*(2), i-31.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. Journal of Educational and Behavioral Statistics, 23, 57-75.

Stocking, M., Smith, R., & Swanson, L. (2000). An investigation of approaches to computerizing GRE subject tests. New Jersey: Educational Testing Service.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*(3), 277-292.

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., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*(3), 365-384.

Veldkamp, B. P., & van der Linden, W. J. (1999). Designing Item Pools for Computerized Adaptive Testing. (Research No. 99-03). *AE Enchede, The Netherlands.* Retrieved from ERIC database. (ED434150)

Wagner-Menghin, M., Preusche, I., & Schmidts, M. (2013). The effects of reusing written test items: A study using the Rasch model. *ISRN Education, 2013*.

Wainer, H. (2014). Cheating: Some ways to detect it badly. In *Test Fraud* (pp. 24-36). Routledge.

Wang, X., Liu, Y., & Hambleton, R. K. (2017). Detecting item preknowledge using a predictive checking method. *Applied psychological measurement, 41*(4), 243-263.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology, 68*(3), 456-477.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*(4), 17-27.

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52*(3), 679-700.

Weekley, J. A., & Ployhart, R. E. (2013). Situational judgment tests: Theory, measurement, and application. Psychology Press.

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. *Situational Judgment Tests: Theory, Measurement, and Application*, *26*, 157-182.

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review*, *19*(3), 188-202.

Wolfe, J. H., Moreno, K. E., & Segall, D. O. (1997). Evaluating the Predictive Validity of CAT-ASVAB. In W.A. Sands, B.K. Waters, and J.R. McBride. (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 175-179). American Psychological Association.

Wood, T. J. (2009). The effect of reused questions on repeat examinees. *Advances in Health Sciences Education*, *14*(4), 465-473.

Wyse, A. E. (2011). The potential impact of not being able to create parallel tests on expected classification accuracy. *Applied Psychological Measurement*, *35*(2), 110-126.

Yi, Q. (2002). Incorporating the Sympson-Hetter exposure control method into the a-stratified method with content blocking. *Paper presented at Annual Meeting of the American Educational Research Association (AERA), New Orleans, LA.*

Yi, Q., & Chang, H. H. (2003). a-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology*, *56*(2), 359-378.

Zara, A. (2006). Defining item compromise. In Unpublished paper presented at the 2006 annual meeting of the National Council on Measurement in Education, San Francisco, CA. Retrieved from https://www.ncsbn.org/2006.04_Zara_-_AERA_-_Defining_Item_Compromise.pdf.

Zheng, Y., & Chang, H. H. (2014). Multistage testing, on-the-fly multistage testing, and beyond. In Y. Cheng & H. Chang (Eds.) *Advancing methodologies to support both summative and formative assessments*, (pp. 21-39).

Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical test theory and item response theory. In C.E. Lance & R.J. Vandenberg (Eds.) *Statistical and methodological myths and urban legends*, (pp.37-61).

12.0 SYMBOLS, ABBREVIATIONS, AND ACRONYMS

ASVAB	Armed Services Vocational Aptitude Battery
CAST	Computerized Adaptive Sequential Testing
CAT	Computerized Adaptive Testing
CBT	Computer-Based Testing
CSH	Sympson & Hetter Conditional Procedure
CTT	Classical Test Theory
DGM	Deterministic Gated Item Response Theory Model
DIF	Differential Item Functioning
DIFMIN	Minimization of Differences
DP	Davey & Parshall
DPF	Differential Person Functioning
DTIC	Defense Technical Information System
ERT	Effective Response Time Model
FLOR	Final Log Odds Ratios
H-IRTRT	Joint Model within a hierarchical framework
HumRRO	Human Resources Research Organization
1PL	One-Parameter Logistic
IRF	Item Response Function
IRT	Item Response Theory
KLD	Kullback-Leibler Divergence
Ln-RT	Lognormal Response Time Model
MA	Maximin
MADI	Maximin with Minimization of Differences
MI	Minimax

MIDI	Minimax with Minimization of Differences
MRM	Mixture Rasch Model
MRM-RT	Mixture Lognormal Model of Response Times
MST	Multi-Stage Testing
NEAS	Navy Enlisted Advancement System
NWAE	Navy-Wide Advancement Examinations
PDRI	Personnel Decisions Research Institutes
PFE	Promotion Fitness Examination
SED	Score Equating Development
SEV	Score Equating Verification
SH	Sympson & Hetter
SJT	Situational Judgment Test
SKT	Specialty Knowledge Test
SL	Stocking & Lewis Multinomial
SME	Subject Matter Expert
STR-C	a-str Design with Content Blocking
STR-SH	Communication of the a-str with the Sympson & Hetter Strategy
TAPAS	Tailored Adaptive Personality Assessment System
TCC	Test Characteristic Curve
TEC	Targeted Exposure Control
TIF	Test Information Function
TRF	Test Response Function
USN	United States Navy
WAPS	Weighted Airman Promotion System
WX	Wang & Xu