The Impact of Item Parameter Drift in Computer Adaptive Testing (CAT)

BY

NICOLE MAKAS RISK
B.A., California State University Fullerton, 2006
M.A., New York University, 2010

DISSERTATION

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Chicago, 2015

Chicago, Illinois

Defense Committee:

Everett V. Smith Jr., Chair and Advisor
Kimberly Lawless
Carol Myford
John Stahl, Pearson Vue
Yue Yin

This dissertation is dedicated to my husband—Jonathan Risk. You provided me with unwavering support when I needed it and also gave me space. Your strength and encouragement got me through more than you know. You are amazing.

To my mother—Diane Makas—for without her reassurance and belief in me I would not have pursued a doctoral degree.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to the chair of my dissertation committee—Dr. Everett V. Smith, Jr.—for giving his time and expertise to this research study. His guidance, understanding, and relaxed approach were essential for the successful completion of this study. I want to give a special thanks to my committee member—Dr. John Stahl—whose CAT expertise allowed me to complete a dissertation in this topic area. I also want to thank the other members of my dissertation committee—Dr. Carol Myford, Dr. Yue Yin, and Dr. Kimberly Lawless—for their professional guidance and support for this study.

I would also like to extend a big thank you to Dr. Kirk Becker, who graciously agreed to modify his CAT simulator program for my dissertation. He willingly worked on numerous modifications of the simulator to meet the needs of my research. Without his generous assistance many of the components of this study would not have been possible.

I want to thank my aunt and fellow PhD—Dr. Elaine Makas—for talking me down from numerous panic attacks and whose understanding of my journey made it possible for me to make it to the end. And my PhD-pursuing friends—Kristin Abner and Anna Colaner—who listened to me complain, understood my misery, and drank wine with me on numerous occasions. All of which allowed me to feel somewhat normal.

Finally, I would like to thank my husband again, who went above and beyond the emotional support of a partner by spending hours reading and editing countless drafts of this dissertation.

TABLE OF CONTENTS

LIST OF TABLES ………………………………………………………………………vii

LIST OF FIGURES……………………………………………………………………viii

LIST OF ABBREVIATIONS…………………………………………………………..ix

SUMMARY……………………………………………………………………………. x

CHAPTER PAGE

I. INTRODUCTION…………………………………………………………… 1

A. Background……………………………………………………………… 1

B. Purpose of the Study…………………………………………………….. 3

C. Research Questions……………………………………………………… 4

D. Approach of the Study…………………………………………………... 4

E. Significance of the Study….…………………………………………….. 5

II. REVIEW OF LITERATURE………………………………………………. 6

A. Computer Adaptive Testing (CAT)……………………………………… 6

1. What is CAT?.. 8

2. CAT methods and basic operations…………………………………... 8

a. Item selection……………………………………………….... 8

b. Test scoring…………………………………………………… 11

3. Advantages and disadvantages of CAT…………………………….. 14

a. Advantages…………………………………………………… 14

b. Disadvantages………………………………………………… 17

B. Item Parameter Drift (IPD)……………………………………………….. 19

1. What is IPD?.. 19

2. Reasons for drift………………………………………………….. 21

3. Impact of IPD……………………………………………………… 23

C. IPD and Fixed-Item Testing (FIT)…………………………………….. 24

1. Detecting IPD in FIT…………………………………………….. 24

2. IPD in FIT and test performance…………………………………… 30

D. IPD and CAT…………………………………………………………….. 41

1. Detecting IPD in CAT……………………………………………... 41

2. IPD in CAT and test performance…………………………………. 44

E. Summary of Literature and Proposed Study...……………………………… 48

III. METHODS…………………………………………………………………… 50

A. Research Questions……………………………………………………. 50

B. Overview……………………………………………………………….. 50

C. Independent Variables…………………………………………………. 51

1. Amount of drift……………………………………………………. 51

2. Magnitude of drift………………………………………………... 51

3. Size of the item pool……………………………………………… 52

D. Test Properties…………………………………………………………. 52

1. Item pool…………………………………………………………. 53

2. Examinee distribution…………………………………………… 53

E. CAT Simulation……………………………………………………….. 54

1. Ability estimation…………………………………………………….. 54
2. Item selection …………………………………………………….. 55
3. Constraints ……………………………………………………….. 55
4. Stopping rule………………………………………………………… 56
5. Direction of drift ………………………………………………… 56

F. Evaluation………………………………………………………………… 57
1. Precision ………………………………………………………….. 57
2. Classification accuracy…………………………………………… 57
3. Efficiency ……………………………………………………….. 58

G. Analysis Approach………………………………………………………… 59
1. Response model …………………………………………………… 59
2. Baseline Condition………………………………………………… 60
3. Research Questions………………………………………………… 60

IV. RESULTS…………………………………………………………………… 61

A. Precision………………………………………………………………….. 61
1. Correlation……………………………………………………….. 62
2. RMSE………………………………………………………………. 63
3. AAD……………………………………………………………….... 64
4. Bias………………………………………………………………… 65

B. Classification Accuracy…………………………………………………….. 67
1. Total misclassification…………………………………………….. 69
2. FP and FN…………………………………………………………….. 69

C. Test Efficiency……………………………………………………………… 69
1. Test length……………………………………………………… 71
2. Item exposure……………………………………………………… 71

D. Supplemental Analysis……………………………………………………… 72
1. Test efficiency……………………………………………………… 72
   a. 60% targeted probability…………………………………. 75
   b. Retargeted sample……………………………………………… 75
2. Precision and classification accuracy……………………………… 75
   a. Precision ………………………………………………….. 76
   b. Classification accuracy…………………………………… 79

V. DISCUSSION…………………………………………………………….... 81

A. Evaluation Criteria………………………………………………………... 81
1. Precision……………………………………………………………... 81
2. Classification accuracy……………………………………………… 82
3. Test efficiency………………………………………………………. 84

B. Supplemental Analysis……………………………………………………… 85

C. Research Questions…………………………………………………………. 87
1. Research Question 1………………………………………………… 87
2. Research Question 2………………………………………………… 87
3. Research Question 3………………………………………………… 88

D. Strengths and Limitations…………………………………………………… 88

1. Strengths……………………………………………………………….. 88
2. Limitations…………………………………………………………….. 89

E. Implications for Practitioners…………………………………………….. 91
F. Suggestions for Future Research……………………………………….. 92
G. Conclusion………………………………………………………………….. 93

REFERENCES…………………………………………………………….. 95

VITA……………………………………………………………………….. 106

LIST OF TABLES

TABLE PAGE

I. Reasons for Drift …..…………………………………………………….. 22

II. Test Properties and Item Characteristics ………………………………… 53

III. Evaluation Criteria ………………………………………….……… 59

IV. Measures of Precision………………………………. …………………… 62

V. Measures of Classification Accuracy…………………………………….. 68

VI. Measures of Test Efficiency……………………………………………… 70

VII. Test Efficiency Comparison for Baseline Conditions Across the Three Sets of Analyses……………………………………………………….. 73

VIII. Test Efficiency Comparison for the 500 Item Pool Across the Three Sets of Analyses ……………………………………………...………….. 74

IX. Precision and Classification Accuracy Comparison Between the Baseline Conditions for The Three Sets of Analyses………………………………. 76

X. Precision And Classification Accuracy Comparison for the 500 Item Pool Across the Three Sets of Analyses ……………………………………. 78

LIST OF FIGURES

FIGURE PAGE

I. RMSE Values for the Large Item Pool with 1000 Items…………………….. 63
II. RMSE Values for the Medium Item Pool with 500 Items…………………… 63
III. RMSE Values for the Small Item Pool with 300 Items.……………………... 64
IV. AAD Values for the Large Item Pool with 1000 Items……………………… 64
V. AAD Values for the Medium Item Pool with 500 Items…………………….. 65
VI. AAD Values for the Small Item Pool with 300 Items.………………………. 65
VII. Bias Values for the Large Item Pool with 1000 Items ………………………. 66
VIII. Bias Values for the Medium Item Pool with 500 Items……………………… 67
IX. Bias Values for the Small Item Pool with 300 Items.………………………... 67

LIST OF ABBREVIATIONS

1PL	One Parameter Logistic
2PL	Two Parameter Logistic
3PL	Three Parameter Logistic
AAD	Absolute Average Difference
CAT	Computer Adaptive Testing
FIT	Fixed-Item Testing
FN	False Negative
FP	False Positive
IPD	Item Parameter Drift
IRT	Item Response Theory
M	Mean
MH	Mantel-Haenszel
MS	Mean-Sigma
RMSE	Root Mean Square Error
SD	Standard Deviation
SE	Standard Error
SEM	Standard Error of Measurement

SUMMARY

In an attempt to better understand and further research on item parameter drift (IPD) in computer adaptive testing (CAT), I carried out a study looking at numerous aspects of IPD and its impact on measurement. The main goals of this research were to (1) gain an understanding of how IPD impacts measurement precision, classification accuracy, and test efficiency, and (2) to inform testing organizations about the potential threat of IPD to testing data and provide them with guidelines for handling IPD. I conducted a series of CAT simulations, varying the amount and magnitude of IPD, as well as the size of the item pool. To establish a baseline, I ran a series of simulations without drift and then compared the results of the altered IPD conditions to the non-altered baseline condition. I evaluated the effects of IPD on measurement precision, classification, and test efficiency, using a number of criteria. These included bias, root mean square error (RMSE), absolute average difference (AAD), total percentages of misclassifcation, the number of false positives and false negatives, the total test lengths, and item exposure rates.

The results revealed negligible differences when comparing the IPD conditions to the baseline condition for all measures of precision, classification accuracy, and test efficiency. The most relevant finding is that magnitude of drift has a larger impact on measurement precision than the number of items with drift. However, the findings overall suggest that under the studied conditions IPD present in a CAT item bank does not significantly impact the precision of scores or the classifications of examinees, and highlights the robustness of CAT to even large amounts of IPD. The results of this study have implications for future CAT research, and provide useful information as to how much and what magnitude of drift is most concerning when estimating ability. The findings can also help testing organizations that use CAT or plan to implement CAT make important decisions regarding the maintenance of their item banks.

I. INTRODUCTION

A. Background

Given advances in technology and the rise of the digital age, computerized assessment has assumed a more prominent role in testing. Assessment organizations and large-scale certification and licensure boards, such as the Educational Testing Service (ETS) and the National Council of State Boards of Nursing (NCSBN), are increasingly implementing computerized versions of their exams. Additionally, researchers conducting studies on the use of testing for making high-stakes decisions seek more accurate and reliable ways to assess examinee ability in order to maximize efficiency and optimize measurement. These changes in testing practice have led to the development of *computer adaptive testing* (CAT), a technique that tailors the administration of items to the ability of the examinee (Reckase, 2011).

The popularity of CAT over *computer-based testing* (CBT), which is simply a computer version of a fixed-item paper and pencil test, is due to the numerous advantages that CAT provides. Compared to conventional testing, CATs can offer increased security, shorter tests, instantaneous score reporting, and more precise measurement (Weiss, 1982). CATs use a computer algorithm to administer items with difficulties matching the ability level of each examinee. This is accomplished through probabilistic or item response theory (IRT) models, which place the item difficulty and examinee abilities on the same scale and then relate the probability of getting a specific test item correct to an examinee's ability level on that scale (Lord, 1980). Thus, by administering items that yield a 50% probability of a correct response, CATs provide the maximum information about an examinee's ability level.

The principle of targeting items to a 50% probability is the key to achieving maximum efficiency or information (Wise, Roos, Plake, & Nebelsick-Gullett, 1994). Studies have shown that not only can a CAT result in smaller measurement errors and a more precise measure of ability, but it can also accomplish this using fewer items in comparison to conventional fixed-item testing (Drasgow, Luecht, & Bennett, 2006; Wainer, 1993; Weiss & Kingsbury, 1984). Minimizing the number of items is particularly advantageous for organizations that need to administer a large number of tests in a cost-effective manner. In a variable length CAT, test developers can specify a pre-determined level of precision so the exam ends only after achieving that level of precision. Examinees are only administered items that are of appropriate difficulty for each individual, which minimizes the standard error of measurement (SEM) and results in a more reliable test (Gershon, 2005).

In order for a CAT to fully realize these measurement advantages over conventional testing, a high-quality item pool is required. Thus, close monitoring of CAT item pools is essential to ensure that items are not flawed, obsolete, or overexposed, because the higher the quality of the item pool, the better the adaptive algorithm can perform (Flaugher, 2000). Maintaining the item pool is not only important to ensure that items are relevant and secure, but it is also necessary in order to evaluate any changes in the item parameters.

A change in item parameter values is known as *item parameter drift* (IPD) (Bock, Muraki, & Pfeiffenberger, 1988). IPD occurs when parameter estimates fluctuate over numerous administrations, despite treating common item estimates as fixed (Babcock & Albano, 2012). IPD can result from educational, technological, or cultural changes that could alter the functioning of an item (Wainer et al., 2010). Items exhibiting IPD can perform differently across groups of examinees who participate in different test administrations (Goldstein, 1983). Therefore, IPD has become a major concern in CAT research because when an item show's evidence of IPD, it may violate a fundamental IRT assumption: examinees of the same ability level have the same probability of answering an item correctly (Babcock & Albano, 2012). This is concerning because IPD has the potential to impact the measurement precision of examinee ability estimates. Additionally, since measurement precision is a pre-requisite for test score validity, IPD can have undue influence on the validity of test score interpretations. Wainer et al. (2010) claimed that investigating CAT item pools for IPD is a requirement to avert a threat to construct validity.

IPD research is limited to date, but a handful of researchers have investigated different aspects of IPD including detecting patterns of IPD (DeMars, 2004), the impact of dimensionality on IPD (Li, 2008), IPD in relation to scale drift (Babcock & Albano, 2012), and identification of factors that influence item drift (Bergstrom, Stahl, & Netzky, 2001). Bock et al. (1988) found that drift will affect item difficulty more than it will affect the slope of the item. Jones and Smith (2006) concluded that common factors affecting IPD include item exposure, too few or too many examinees, and the overall number of items in a bank. There is limited research investigating the impact of IPD on ability measures. A few researchers have demonstrated that the effect of IPD on theta estimation is minimal (Chan, Drasgow, & Sawin, 1999; Wells, Subkoviak, & Serlin, 2002). However, Jones and Smith (2006) looked at IPD impact on a certification exam's examinee pass-fail rate and concluded that IPD negatively influenced pass-fail consistency. Their results implied that given certain situations, there is potential for IPD to negatively impact an examinees ability estimate and passing status and therefore affect various aspects of test validity. Due to the limited and discrepant findings in these studies and the potential for negative impact on person ability measures, it warrants further research in this area. Additionally, Wainer et al. (2010) recommended taking steps to monitor and account for drift in all item pools because of the unknown but conceivable effect IPD might have on the validity of test score interpretations. As CATs become more frequently used, evaluating the effects of IPD on ability estimation and classification accuracy is crucial, given the high-stakes decisions that surround the exams.

B. Purpose of the Study

The purpose of this study is to examine various aspects of IPD and evaluate its effects on examinee ability estimation and pass-fail decisions in a CAT certification exam. I compared differing amounts and magnitudes of IPD and looked at the extent to which IPD jeopardized measurement precision, confidence in pass-fail decisions, and efficiency of the exam. Additionally, I examined the size of the item pool to assess how it impacted IPD and to ascertain how the size of an item pool might impact how much drift it could withstand. The main goals of this research are to gain an understanding of how

IPD impacts interpretations of classification accuracy and to inform testing organizations about the potential threat IPD could pose to the validity test users make from test scores.

C. Research Questions

The overarching research question for this study is “to what extent are examinee ability measures, classification accuracy, and exam efficiency for CAT impacted by IPD?” I addressed the following specific research questions:

1. What amount of drift can be present in the item bank before examinee ability estimates, pass-fail decisions, and the overall efficiency of the exam become compromised?
2. What magnitude of drift has the greatest impact on examinee ability estimates, pass-fail decisions, and overall exam efficiency?
3. Do the effects of IPD on examinee ability, pass-fail decisions, and overall exam efficiency vary by the size of the item pool?

D. Approach of the Study

To answer my research questions, I emulated an operational certification exam by simulating a CAT with the same content restrictions, item difficulty values, and ability distribution. I conducted a series of CAT simulations varying the amount and magnitude of IPD, as well as the size of the item pool. I then compared the results of the altered CATs to a non-altered CAT. The amount of drift varied by three levels: 100, 75, and 50 items with IPD in the bank. The magnitude of the drift varied by three levels: 0.5 logits, 0.75 logits, and 1.0 logits. The size of the item pool also varied by three levels (small: 300, medium: 500, and large: 1,000). Additionally, items drifted in both directions with an uneven split, (i.e., 75% of the items drifted easier, while 25% of the items drifted harder). The direction of drift remained constant across all the conditions. I used a crossed factorial design: 3(amount) × 3(magnitude) × 3(item pool), so that I could analyze both main and interaction effects.

I used several criterions to evaluate the effects of IPD on my variables of interest. To examine the precision and recovery of ability estimates, I used average bias, root mean square error (RMSE), and the absolute average difference (AAD). To investigate classification accuracy, I looked at the total number of misclassifications, the number of false positives and false negatives, and the number of misclassifications within a 95% confidence interval. Finally, to examine the efficiency of the exam, I compared the average numbers of test items administered and the overall item exposure rates.

E. Significance of the Study

Ensuring that test scores are fair and accurate measures of ability is essential for validity in all testing scenarios but is especially important in high-stakes testing and certification. The presence of IPD in an item calls into question the accuracy of ability estimates and the pass-fail decisions made from the scores. Therefore, investigating the impact of IPD on examinee ability estimation is important to safeguard the integrity of the exams and ensure that the inferences made from the scores are valid. Research on IPD in CAT is sparse, and research regarding the impact of IPD on ability estimation and pass-fail decisions is limited even further. Due to the extensive use of CAT and IPD's potential to impact accurate assessment of ability measures and validity of test score interpretation, it is imperative to investigate drift and understand how it affects ability estimation.

This study contributes to the literature by examining the effects of IPD on examinee ability estimates through a series of simulation studies set to model various amounts and magnitudes of drift. The results of this study have implications for future CAT research and provide useful information regarding how much and what magnitude of drift is most concerning when estimating ability. The findings can help testing organizations that use CAT or plan to implement CAT make important decisions regarding the maintenance of their item banks.

II. REVIEW OF LITERATURE

Few researchers have studied IPD and its effect on measurement precision, especially in CAT (McCoy, 2009). Additionally, research findings regarding the effects of IPD on the accuracy of measures of ability are mixed; thus, concern over IPD persists. The purpose of this study is to examine various aspects of IPD in a typical CAT licensure/certification exam, determine how it affects a CAT exam when IPD is not present in the item bank, and identify any substantial threats it might pose to measurement precision, accuracy of pass-fail decisions, and efficiency of the exam.

In this chapter, I will explore the experimental research on IPD and CAT. The review contains two main sections: (a) CAT and (b) IPD. I will start with some of the research that led to the development of CAT. I will give a brief description of the process and basic operations of CAT, while highlighting its advantages and disadvantages. Then I will move on to the topic of IPD. I will define IPD and identify factors that contribute to IPD. I will also review the literature on methods to detect IPD, as well as research into its influences on ability measurement and pass-fail decisions. Finally, I will summarize the research relating the two areas, focusing specifically on experimental studies that examine the role of IPD in CATs.

A. Computer Adaptive Testing (CAT)

The development of Rasch and item response theory (IRT) models, along with advances in modern technology, made computer adaptive testing possible (Gershon, 2005). CAT refers to a form of computer-based testing where each examinee takes a customized test. These customized tests offer numerous benefits to both examinees and testing companies and have led many large-scale testing programs to adopt computerized versions of their exams in recent years. Often-cited advantages of CAT include reduced testing cost, increased testing efficiency, improved testing security, and instant feedback to examinees (Wise & Plake, 1989).

The idea of adaptive testing has a long history with original roots in early mental and developmental testing (Wainer et al., 2010). The *Stanford Binet Intelligence Scales*™ is one of the earliest adaptive tests that Alfred Binet developed when he realized he could tailor individual tests by rank ordering the items in terms of difficulty (Linacre, 2000). Binet used an adaptive approach for the selection of items. He began with items that matched the examinee's age and then continued with harder questions if the examinee answered correctly or easier items if the examinee answered incorrectly. He ended the test when the examinee frequently failed items. Binet was able to estimate the examinee's ability based on this information. Other early tests that refined Binet's method include Lord's Flexilevel testing procedure and its variants, such as Henning's Step procedure and Lewis and Sheehan's Testlets (Linacre, 2000).

One of the first high-stakes, large-scale computer adaptive tests to be developed was the *Armed Services Vocational Aptitude Battery* (ASVAB). The CAT-ASVAB is a human proficiency exam that influences the qualification status of applicants for the U.S. Armed Forces and, as such, is one of the most researched tests in modern history (Segall & Moreno, 1999). In 1990, the Armed Forces introduced the CAT version of the exam in six locations. Today, only a handful of low-volume satellite locations continue to use the paper-and-pencil version. Another widely known large-scale CAT is the *Graduate Record Examination*® (GRE®), which began development in 1988. The first administration of the computerized version of the paper-and-pencil test was in 1992, and in 1993, the CAT became fully operational (Guernsey, 2000). The American Society of Clinical Pathologists (ASCP) exam was the first CAT certification examination to go live in 1993. Other high-stakes exams offered in a CAT format include the *Graduate Management Admission Test*® (GMAT®) and the *National Council Licensure Examination* (NCLEX®). The NCLEX shifted to a CAT format in 1994, and the GMAT moved exclusively to a computer adaptive administration in 1997 (Guernsey, 2000).

1. What is CAT?

In contrast to a traditional linear test that has a fixed number of test items, CAT is a special approach to the assessment of latent traits in which the test is specifically matched to the ability level of each examinee (Davey & Pitoniak, 2006). Although an examinee's true ability level for a given trait is unknown, CAT uses methodology that enables the estimation of the examinee's ability level by iteratively estimating the trait level during the entire testing process. The test is assembled as the examinee takes the test, and the computer selects test items based on the examinee's responses to items presented previously.

There are two main steps in the design and administration of an adaptive test: item selection and score estimation (Davey & Pitoniak, 2006). The first step involves choosing appropriate items given the examinee's performance level, and the second step involves using the examinee's response to each item to refine the score or ability estimate so that the next item selected targets the examinee's ability level even better. If the examinee answers the first item correctly, the next item presented is more difficult. If the item is answered incorrectly, then the next item presented is less difficult. The test continually updates the examinee's current ability estimate based on the responses to the items that he or she has already answered. This process continues until it reaches a test termination criterion, such as a pre-specified level of measurement precision, or a fixed number of administered items (Davey & Pitoniak, 2006). Thus, not all examinees will respond to the same number of items. This approach enhances measurement precision and reduces testing time.

2. CAT methods and basic operations

a. Item selection

The first step in creating and administering a CAT is item selection. The primary goal in selecting items for a CAT is to maximize the test information function (TIF)[1] and minimize measurement error in the examinee's score (Drasgow, Luecht, & Bennett, 2006). Selecting items involves three considerations: (a) optimizing test efficiency, (b) properly balancing the test, and (c) protecting items from overexposure. A CAT relies on the principles of Rasch or IRT models to determine the test items to administer. It selects the items according to an algorithm that attempts to maximize the efficiency of a test by providing the maximum amount of information. Items are administered as a function of examinee's knowledge level; and in order to maintain the appropriate level of item difficulty, each item is chosen to match the most current estimate of the examinee's ability. A CAT administers a more difficult set of items to an examinee who has shown a high probability of demonstrating mastery of the subject matter, and administers an easier set of items to an examinee who has exhibited a low probability of demonstrating mastery of the subject matter.

(1). <u>Selection of the first item</u>

The purpose of the test dictates how the test developer selects the very first item on the test. For norm-referenced tests, where the purpose of the test is to determine the ability level of an examinee in relation to other examinees' ability levels, the test developer typically selects the first item at the mean of the population being tested (e.g., the average ability level of entry-level medical assistants). Test developers can do this by specifying the mean ability level and selecting an item in the pool that reflects that point or by using Bayesian estimation methods (discussed later in the "Ability estimation" section). At this point, all the test developer knows is that the examinee's ability measure is likely to be either above or below the difficulty level of the first item. The test developer's goal is to estimate the true ability level of the examinee. However, the test developer knows relatively little about the examinee's ability after the first item. Since there is not enough information to compute a precise

[1] The TIF is the amount of information yielded by the test. Using item difficulty and examinee ability, the amount of information for a single item can be computed at any ability level. The TIF is simply the sum of all item information for the test and tells us how well each ability level is being estimated (Baker, 2001).

estimate of ability at this stage, sometimes the first few items are constrained to match the mean ability level of the population. In criterion-referenced, or mastery tests, the purpose of testing is to determine whether an examinee has acquired a certain amount of knowledge. In this type of testing, it is more efficient to select as the first item on the test an item with a difficulty level at or close to the pass point of the test. Regardless of whether the examinee answers the item correctly or incorrectly, continuing to select a few more items that have difficulty levels near the pass point of the test can optimize accuracy. Although the adaptive algorithm will estimate true ability by selecting items based on examinee ability, in this case, it is more efficient to constrain the items early on. The test developer is more concerned with whether the examinee's ability measure is above the pass point than whether the examinee's ability measure is far above (or far below) the pass point.

(2). Content balancing

Content balancing is another constraint that the test developer must consider in the item selection process (Gershon, 2005). Although the test might be measuring a unidimensional construct, it might also contain sub-content areas that are important, given the purpose of the test. For example, a test assessing math achievement might cover sub-content areas that include geometry, algebra, and statistics. Test developers must properly balance each examinee's test in terms of item content. Most adaptive tests attempt to maintain the same content distribution assembly process used in conventional fixed-item tests, with a test blueprint. Test specifications dictate the construction of conventional tests by outlining what content to include and the proportion of items that must address each sub-content area (Davey & Pitoniak, 2006). The test developer can specify the CAT algorithm to select items that not only maximize information about the examinee's ability level, but also conform to the desired content scheme. These specifications are considered constraints in CAT because they restrict the items that can comprise the test and alter the overall efficiency as compared to an unconstrained CAT.

(3). Item overexposure

Another important consideration in item selection is item exposure, which is how often examinees see an item . This constraint helps protect items from overexposure and ensures that less frequently used items are employed. Overused items become a threat to test security because items are administered to a large number of examinees, whereas underused items are a waste of resources because they are administered to too few examinees. However, test developers can address these concerns by incorporating a randomization factor into the CAT item selection procedure. Instead of selecting the item that best matches the examinee's ability level every time, the algorithm might choose from the best 5 or 10 items available in the pool and then randomly pick from those best 5 or 10 (Gershon, 1996). Another method of avoiding item overexposure is to select items from multiple sub-pools, where the first item is selected from one sub-pool, and the next item is selected from another sub-pool (Gershon, 2005). Again, this is another restriction on the overall efficiency compared to unconstrained CATs.

b. Test scoring

The second step in creating and administering a CAT is score estimation. During the CAT process, each time an examinee responds to an item, the test algorithm calculates an estimate of the examinee's ability. Naturally, the estimate is rough at the beginning, but it becomes refined as the test continues. The examinee's score is estimated using the response model that underlies the test. The mathematical foundation of a CAT is a Rasch or an IRT model. The most commonly used model is the Rasch dichotomous model; however, the two- and three-parameter IRT models and the rating scale and partial credit models are used in adaptive testing as well.

Rasch and IRT models determine ability measures using the mathematical function for the probability of observing a particular item response given the examinee's ability level and item parameters. These models put examinee ability and item difficulty on a single continuous scale, which suggests that there is a point on the scale where an examinee's ability level is equal to the difficulty of an item. The model then estimates an examinee's ability level based on this interaction between examinee ability and item difficulty. In order to maximize the information about an examinee's ability level from each item, the CAT will ideally administer items where the difference between difficulty and ability is close to zero. Administering items that fit these criteria also reduces the standard error of measurement (Gershon, 2005). In this regard, item-free ability estimation properties of the IRT model allow CATs to select items conditionally based on the examinee ability. Examinees do not need to be administered the same items to obtain their ability estimates. The item-free measurement allows score comparisons even though individual tests are composed of different items.

(1). Ability estimation

CATs use two common methods to estimate ability and calculate measurement errors: maximum likelihood estimation (MLE) and Bayes estimation. Both methods use a combination of the item parameters and the examinee's responses to items to determine the examinee's current ability estimate. MLE updates the ability estimate by taking into account the difficulty levels of the items that were previously administered and the examinee's response to the most recently administered item. The next item MLE chooses is one that will provide the maximum information—the item with a difficulty level that is closest to the examinee's current ability estimate. By contrast, Bayes estimation is based on a specified distribution of examinee ability estimates, which can be from a normal distribution or from a known typical population distribution for a particular exam (Gershon, 2005). This method selects items assuming that an examinee's true ability estimate is close to the mean or mode of the specified distribution. The examinee's response to the most recently administered item is then used to update the "prior" distribution before administering the next item and estimating ability.

(2). Stopping rule

The most crucial component of a CAT is the decision about when to stop the test. If the test is too short, the ability estimate may not be accurate; if the test is too long, then it could be a waste of resources (Linacre, 2000). When items are administered unnecessarily, they run the risk of becoming overexposed, and overexposed items need to be replaced more frequently. The stopping rules used to determine when to terminate a CAT result in two main types of CAT exams: fixed length and variable length. A fixed-length exam stops after a certain number of items are administered, which means that every examinee takes a test with the same number of items. Variable-length exams use different criteria and stop only after these criteria are satisfied. With variable-length exams, not every examinee is administered the same number of items. Variable-length exams terminate in one of two ways: (a) when a specified level of precision is reached, or (b) after a specified level of confidence in the pass-fail decision is reached (Bergstrom & Lunz, 1999). CATs sometimes use both fixed- and variable-length rules. For example, a CAT may administer a set minimum-number-of-items before the confidence-level or precision-level rules can terminate the exam, or the test may terminate after a set maximum number of items are administered, regardless of whether the confidence-level or precision-level rules are achieved.

Test developers need to take the purpose of the exam into account to determine the appropriate stopping rule. Test developers often use the pass-fail confidence-level rule for criterion-referenced testing (CRT) or credentialing exams, where the goal of the test is to determine whether the examinee has demonstrated a certain level of ability. A 95% confidence interval can at times be achieved rather quickly with the administration of a minimal number of items. It is easier for test developers to make confident pass-fail decisions for high- and low-ability examinees when their ability measures are clearly above or below the cut score. However, in high-stakes testing, it is often hard for a test developer to explain to an examinee that he or she passed or failed the exam when the examinee answered only 10 items. In these instances, the minimum number of items rule is used in conjunction with the confidence-level rule. On the other hand, it is almost impossible for a test developer to make a clear pass-fail decision when the examinee’s ability measure is near the cut score. The test could get lengthy because the examinee might bounce above and below the cut score as he or she answers additional items. In these instances, the test developer would use the maximum-number-of-items rule in conjunction with the confidence-level rule.

In achievement testing, when the purpose of the test is to provide an accurate estimation of ability, a precision-level stopping rule is typically used. These tests will continue until a specified standard error of measurement is reached to reflect the examinee’s most accurate measure of ability. The most accurate measure is obtained when the error of measurement is the smallest. Although these exams are also variable in length, they are more likely to administer a similar number of items to all examinees than pass-fail tests do, because a cut score is not used to compare against the examinee’s ability measure. Exams with a precision-level stopping rule also tend to be longer because obtaining more precise ability measures requires the administration of more items.

3. Advantages and disadvantages of CAT

a. Advantages

CATs have many cited advantages compared with traditional fixed-item testing, including increased efficiency, better measurement at the extremes of the examinee distribution, faster scoring, improved security, increased fairness, and cost benefits. Additionally, adaptive tests are often considered more reliable and are thought to enhance validity because they take advantage of technology and modern measurement theory (Gershon, 2005). As the practical constraints of CATs are decreasing for testing companies, their benefits are beginning to outweigh their disadvantages.

(1). Increased efficiency

CAT offers improved testing efficiency. That is, test developers can obtain more precise examinee ability estimates using fewer items than are required when using nonadaptive tests (Drasgow et al., 2006). The item selection algorithm is the mechanism that makes this possible: Items that are too easy or too hard for an examinee are not administered. With an adequate item pool, a CAT can be only half as long as a parallel nonadaptive test and obtain the same measurement precision (Drasgow et al., 2006). Although CATs rarely achieve optimal efficiency due to practical constraints (e.g., content balancing), they are still significantly shorter than fixed-item tests. Consequently, the administrative efficiency and measurement precision that CATs offer has led many measurement professionals to consider them superior to conventional tests (Zwick, 2006).

(2). Better measurement

Conventional tests are designed to do a good job of measuring the abilities of examinees that are in the midrange of performance. That is, they are good at measuring ability at the center of a distribution but are often poor at measuring ability at the extremes of the distribution (Davey & Pitoniak, 2006). By contrast, CATs individually adapt the test's difficulty level to the examinee's ability and administer only items of appropriate difficulty (i.e., items that are closely matched to the examinee's ability level). When items are targeted to the examinee's ability level, the standard error of measurement is reduced. The minimal standard error maximizes the precision of ability measures for every examinee (Gershon 1996; 2005).

(3). Immediate score reporting

CATs are able to provide examinees with their test results immediately after they finish the test. Having immediate access to the score report offers examinees such conveniences as allowing them to meet tight application deadlines or to register for retesting if their performance is substandard (Davey & Pitoniak, 2006). Immediate scoring also saves test organizations the hassle and cost of mailing examinee score reports.

(4). Test security

Test security is one of CAT's often-cited benefits for a few reasons (Davey & Pitoniak, 2006; Gershon, 1996). First, it is more feasible to send secure electronic files than hard copies of the exam to testing centers. Additionally, it is very difficult for examinees taking CATs to share or copy answers. However, most importantly, CATs greatly diminish item exposure. Because CATs are assembled individually for each examinee, multiple test forms are created for various ability levels. Not only do examinees rarely see the same sets of items, but they also are exposed to only a fraction of the total available items in the bank (Wainer et al., 2010). This also decreases the impact of examinees memorizing and sharing items.

(5). Increased fairness

CATs also have characteristics that enhance fairness for examinees, because humans are taken out of the selection and construction of test forms (Gershon, 2005). When the item bank is large and well-constructed, the exams can be individually tailored to meet the ability level of each examinee. Thus, all examinees have the same opportunity to demonstrate their abilities (Gershon, 2005). In addition, since most CAT programs have more than one item bank, they can be easily changed out to remove compromised items quickly. Monitoring compromised items and maintaining and refreshing item banks further enhances the fairness of the exam for all examinees.

(6). Cost reduction

A number of cost-reduction benefits can result when moving from a paper-and-pencil test to a CAT. First, the costs and time associated with creating multiple test forms are often greatly reduced. This results from piloting a large number of items at one time when there is a large testing population. Because more items can be piloted and seeded into CAT tests (as an alternative to the inclusion of a smaller number of pretest items on a single paper-and-pencil test form), item statistics can be obtained, and item banks can be built at a faster rate. Additionally, testing organizations can administer their tests more frequently and in multiple locations, because they can easily generate multiple test forms. Second, the individualized tests that CATs produce have shorter testing times, which decrease expenses and inconvenience. Finally, test scoring costs are virtually nonexistent because CATs are scored during the administration process, and the final score report can be generated at the time of testing.

b. Disadvantages

Despite the numerous advantages that CATs can offer, they do have some limitations and practical constraints, including several technical and procedural issues.

(1). Cost and feasibility

Although testing organizations can benefit financially in the long run by moving to CATs, the upfront costs can be prohibitive. The biggest cost considerations are the large calibrated item pool needed to administer a CAT and the large examinee sample sizes required to carry out IRT item calibrations. A large item pool is necessary to fulfill the content and item exposure constraints in order to ensure that there is an adequate number of items with appropriate difficulties for the entire range of ability levels. Testing organizations often have difficulty trying to balance the content representation with appropriate item exposure rates, and reduced efficiency of the exam becomes a problem. Additionally, not closely monitoring the test's item exposure rate can compromise test security.

Other limitations are the cost of computer hardware and access to a computer network that has the system requirements necessary to administer CATs (Luecht & Sireci, 2012). CATs use a large amount of data and require a high degree of computation to estimate ability during test administration. The computer used to administer the test must have high speed and a large storage capacity in order to implement the item selection and scoring algorithm. This is particularly problematic for Internet-based testing networks that rely on central computer servers.

(2). Inability to review

Generally, CATs do not allow examinees to skip items or to review their answers to previous items. Examinees frequently complain about the inability to review items; however, some researchers argue that allowing item review can lead to upward biased scores for some examinees (Wainer, 1993). A test-wise examinee could detect incorrect answers or trick the adaptive test into building an easy exam if the examinee were to deliberately answer items incorrectly. Reviewing and correcting their prior responses could lead to an erroneously high score. Despite this claim, some research shows that the level of precision in ability estimation attained is not impacted by item review (Olea, Revuelta, Ximenez, & Abad, 2000), and there are testing organizations that permit item review, like the American Society for Clinical Pathology (ASCP), which has allowed item review since it first started administering CATs (Lunz, Stahl, & Bergstrom, 1993).

(3). Psychometric issues

CATs provide less control over the tests that are administered to examinees because neither subject matter experts nor psychometricians are able to review a test "form" before it is administered. Luecht and Sireci (2012) have suggested that this could result in examinees receiving suboptimal tests with respect to measurement precision. Additionally, Embretson and Reise (2000) have raised concern about the adaptive procedures that CATs use. According to Rasch and IRT assumptions, a test should produce equivalent ability measures for an examinee regardless of the order of the items on the test. The invariance assumption assumes that item difficulty values and item discrimination statistics remain consistent for examinees with different levels of ability and when items appear in a different order on the test. However, several researchers have found that item position can impact ability measures, which violates the assumption of item parameter invariance (Doris & Sarason, 1955; Gershon, 1992; Munz & Smouse, 1968; Whitely & Dawis, 1976; Yen, 1980). Their findings also suggest that parameter estimates can change due to item order. Since IPD is defined as a change in parameter estimate, this finding implies that IPD can impact test performance as a result of item order. Other researchers have failed to find a relationship between ability measure and item order when they investigated the effect of changing test item positions on examinee performance (Klein, 1981; Neely, Springston, & McCann, 1994; Plake, 1980). If, in fact, item parameters are sensitive to item order and item order affects test performance, estimated ability measures may be biased. The examinee's responses to items that appear at the beginning of an exam could affect that examinee's score. This suggests that ability measures estimated from CATs could be more susceptible than fixed-item tests (FIT) to the effects of IPD, since CATs are assembled individually, and items often appear in different positions on each test. This is especially true if items with drift appear at the beginning of the exam because the CAT might estimate an examinee's ability incorrectly and administer inappropriate subsequent items. Thus, the examinee may receive a lower or higher score than he/she deserves.

The next section reviews the literature on IPD and its impact on test performance.

B. Item Parameter Drift (IPD)

One of the many benefits of IRT is the invariant measurement scale that it establishes. Invariance is the concept that parameter values are identical in separate examinee populations or across separate measurement conditions, which is necessary to infer generalizability (Rupp & Zumbo, 2006). Parameter invariance also assumes that examinees of a given ability have the same probability of answering an item correctly (Babcock & Albano, 2012). Thus, the invariance property plays an important role in testing because it offers the ability to build measurement scales that can be expected to maintain their measurement characteristics, even when test forms are modified or adaptive tests are implemented (Kingsbury & Wise, 2011). However, it is not always possible to satisfy this property in practice, because items tend to "drift" over time due to a variety of factors.

1. What is drift?

In testing, *drift* simply implies that there was a difference or shift in the test item's score scale (i.e., difficulty) or the construct being measured when the item was administered on two testing occasions or to two groups (i.e., the item difficulty values drifted). One can define IPD as a change in one or more of an item's parameters over time (Goldstein, 1983). *Parameter drift* can occur when an item's parameters vary systematically across time (Hatfield & Nhouyvanisvong, 2005) or vary over subsequent testing occasions (Bock et al., 1988). Bergstrom et al. (2001) defined IPD as a standardized logit difference greater than or equal to 2.00 logits between pretest and active items.

Another kind of drift is *scale drift*. Scale drift is often referred to as the accumulation of random equating error over multiple test administrations (Livingston, 2004). According to Babcock and Albano (2012), *scale drift* is a change in the measurement scale resulting from a shift in the construct or content across cohorts or across time (e.g. one 5th grade class does better than another 5th grade class on the geometry portion of an exam due to that teacher's emphasis on geometry material). *Construct shift* refers to an actual change in the construct measured across different exam forms (Martineau, 2004, 2006). This might occur if testing organizations conduct a new job analysis and revise the performance criteria on an exam to reflect changes to the job role or tasks that incumbents perform. Some researchers have referred to IPD as a type of *differential item functioning* (DIF) because items perform differently across groups who participate in different test administrations (Babcock & Albano, 2012). However, DIF is characterized as a change in parameter values for different subgroups. It is a measure of how differently an item operates for various subgroups of a population, where the probability of a correct response on that item is different for examinees of equal ability who are from different groups (e.g., racial/ethnic groups). By contrast, IPD is characterized as a differential shift in parameter estimates over time and a measure of how differently an item operates for a population relative to that time (McCoy, 2009).

In IRT, *item parameters* include difficulty, discrimination, and guessing parameters. A change in item difficulty or item discrimination values can result in IPD. There are three types of drift: a-, b-, and ab-drift. b-drift refers to an item's increase or decrease in difficulty over time from its initial calibration, a-drift refers to a change in the item's discrimination parameter, and ab-drift arises when both the discrimination and difficulty parameters change. Depending upon the IRT model that is used one may obtain measures of different types of drift . Within a Rasch model framework, only measures of b-drift are obtained, (i.e., items can become more or less difficult relative to other items over the period of time in which the items are used). When using 2PL or 3PL models one obtains measures of all three types of drift if changes in item discrimination and/or item difficulty occur.

2. Reasons for IPD

Item parameter invariance is susceptible to a number of threats. Several researchers have investigated and documented factors that influence IPD and its effects (Kingston & Dorans, 1984; Leary & Dorans, 1985; Whitely & Dawis, 1976; Yen, 1980). The most common type of IPD is b-drift, or changes in item difficulty (Bock et al., 1988). As Table 1 indicates, this can happen for a number of reasons, such as changes over time in examinee's knowledge, skills, and abilities within a certain discipline or vocation; changes in attitudes within a population; changes in curriculum; or item disclosure/overexposure. When an item becomes easier to answer over time, it may indicate overexposure, a security breach, or test-wise training. When an item becomes more difficult to answer over time, it can be due to educational, technological, or cultural changes (e.g., changes in instructional practice, changes in curriculum changes in the definition of the). Poor initial calibration of items can result in both easier and harder shifts. Additionally, changing the location of an item in a test can contribute to an item becoming easier or harder, when the location of the item on a pre-test is different from its location on the operational form. An item's discrimination parameter would be expected to shift when there is a change in the reliability of scoring a constructed-response item (Donoghue & Isham, 1998), thus producing a-drift in the item. In any of these cases of IPD, the interpretation of test scores and program judgments based on those scores are jeopardized. In practice, testing experts typically recommend removing or re-estimating the drift items (Kolen & Brennan, 1995).

Table 1

Reasons for Drift

Easier Shifts	**Harder Shifts**
Overexposure of items	Changes in curriculum, practice, or policy
Disclosure of item content due to cheating or a security breach	Changes in instruction, knowledge, or skills
Test-wise training	Historical Events
Changes in curriculum, practice, or policy	Cultural changes
Changes in instruction, knowledge, or skills	Poor initial calibration
Historical Events	Item location
Cultural changes	
Poor initial calibration	
Item location	

When trying to diagnose why items appear easier than their bank value indicates, practitioners and/or researchers should consider questions such as the following: Is the item content more stressed in practice now compared to when the item bank value was determined? Does the item cover relatively new content? Was the opportunity to learn the new content lacking? If the answer to any of these questions is yes, then item difficulty should be re-estimated, and the new estimate should replace the existing item's value in the item bank. In the event that the security of the item has been compromised or the item has been overexposed through past examinations, the items may be candidates for deletion from the bank (Kolen & Brennan, 1995). In situations where items become more difficult than their item bank values, practitioners and/or researchers should consider the following questions to determine why these items are becoming harder: Is the item content less stressed in practice now compared to when the item bank value was determined? Is the item content dated? Are any of the distracters no longer plausible? If one or more substantive reasons for IPD can be determined, then item difficulty should be re-estimated, and the new estimate should replace the existing value in the item bank. If the content is dated, the item should be removed from the bank (Kolen & Brennan, 1995).

Other reasons identified for drift in items include: a flawed original item calibration (Jones & Smith, 2006); the position of the item in a common item-equating design; changes in answer sheet design;

changes in item location on the exam; changes in font or pagination used on the exam; administration of the exam under nonstandard conditions (i.e., accommodations such as extra time, Braille, large font); or changes in the content domain that render the original correct response less correct or one of the distracters more correct (Kolen & Brennan, 1995). Kolen and Brennan also suggested considering the following questions when investigating the potential reason for item drift: Were there any text changes to the item or rearrangement of the options in the item from its original bank format? Was there any change in how long examinees had to take the test? Is it possible that motivation conditions changed for the examinees (e.g. the item content makes examinees uncomfortable)? If any of these conditions occur, the item should be considered a new item, and the item difficulty should be re-estimated and replaced for use with future administrations (Kolen & Brennan, 1995). Unfortunately, IPD is likely to occur even when item pools are maintained with quality items and protected by good security procedures. Thus, instituting measures to detect item drift is of critical importance.

3. Impact of IPD

The presence of IPD may violate the fundamental IRT assumption of invariance and therefore poses a threat to measurement. IPD may confound or exaggerate measurement errors and impact the underlying construct or the content validity of the affected items. Prior parameter estimates may no longer accurately model these items; and in the presence of drift, theta estimates may no longer be considered to be measurements of the original construct. If an item's parameter estimate changes, inferences using estimates of examinee ability based on the initial parameter estimates become less valid (McCoy, 2009). Thus, changes in parameter estimates threaten the validity of score-based decisions because they introduce trait-irrelevant differences over time (Donoghue & Isham, 1998).

Items exhibiting IPD can also impact the validity of many IRT procedures, including equating and adaptive testing. IPD can increase equating error by either incorrectly including or excluding an IPD item from among the common items. Typically, an analyst would want to remove the item from the set of common equating items if IPD was due to construct-irrelevant factors but keep the item among the set of common equating items if IPD was related to the construct being measured (e.g. the item difficulty changed as a result of a change in the job task that the question addresses). In adaptive testing, pretest items are calibrated by fixing the operational item parameters at their original values. Thus, the accuracy of pretest item calibrations is jeopardized if there are a substantial number of operational items with IPD (Meng, Steinkamp, & Matthews-Lopez, 2010). Additionally, IPD poses a threat to measurement applications that require a scale to be stable over time (Wells et al., 2002). Scale stability is important to ensure stable score reporting and allow comparability of scores from different test administrations (Guo & Wang, 2003). In paper-and-pencil tests (PPT), unstable scales result from equating a new test form to one or more existing forms. In CAT, scale drift occurs from errors in item calibrations and parameter scaling of new items over time. Thus, IPD not only impacts the stability and accuracy of the scale, but it also compromises the validity of cut scores, which can lead to inaccurate and invalid inferences, regarding examinee performance.

C. IPD and Fixed-Item Testing (FIT)

1. Detecting IPD in FIT

Most researchers studying IPD have investigated drift in fixed-length or FIT. They have employed a variety of methods to detect drift, such as time-dependent models, analysis of covariance (ANCOVA) models, and even DIF models. They used various exams, conditions, and different IRT models. In the next section I discuss the methods employed in a number of studies that have focused on detecting drift.

Bock et al. (1988) investigated a method for maintaining and updating an IRT scale while accounting for IPD on the College Board Physics Achievement Test and the English Achievement Test. The authors evaluated the stability of item parameter estimates using a 3PL time-dependent model, estimating item parameters and parameter trends concurrently. They used analysis of variance (ANOVA) to examine two-way interactions between items and occasions over a 10-year period. In general, the results indicated that drift is relatively systematic over time in large populations. The results also revealed that IPD had a greater effect on the item locations (difficulties) them on the item slopes (discrimination) for both content areas. Lastly, they reported statistically significant drift in item difficulty over time for some of the physics items, but not for the English items. The authors attributed the drift in the physics items to changes in instruction over the time period. Because the study focused on the development of a statistical model for detecting IPD, there was no discussion of how the drift impacted examinee scores.

Cook, Eignor, and Taft (1988) used the 3PL model to investigate the invariance requirement of item parameters and study the effect of instruction on item parameter estimates. They used two forms of a biology exam administered over three time points. The older form was administered in the fall to one group of examinees. The newer form was administered in the fall of the next year and then again to a different group of examinees in the spring. The authors employed three equating methods: equipercentile equating with a common item set, linear equating, and item response theory equating. The results from both the classical test theory (CTT) and IRT analyses indicated that item difficulty estimates were not stable across the fall and spring administrations. However, the item estimates were stable between the two fall administrations, despite one exam being an older form. The authors concluded that recency of instruction influenced item performance because item difficulty estimates differed between the fall and spring administrations only.

Stone and Lane (1991) also examined the impact of IPD by implementing a model-testing approach using the 2PL model. They investigated the stability of item parameter estimates over time with 19 items from a math achievement test across two test administrations. The authors compared two types of models: (a) a completely unrestricted model and (b) a completely restricted model, where item discriminations and difficulties were constrained to be equal across groups. The unrestricted model provided a better model fit. Although some items had unstable item difficulties and discriminations, the parameters of the majority of items remained stable over time. The authors acknowledged that they only investigated the stability of item parameters and recommended in the future researchers should look at the impact of IPD on the validity of inferences from test scores.

Sykes and Fitzpatrick (1992) investigated the stability of item parameter estimates calculated using the Rasch model. They used empirical data from a 285-item professional licensure exam administered over a 5-year period. The authors employed ANCOVA methods to analyse the data and they considered possible explanations for changes in b-values or item difficulty estimates, which included item position, item content, item type, and elapsed time between administrations. Item difficulty estimates showed directional drift with some items becoming more difficult over time. They conclude that the drift was not associated with changes in item position or item type. One of the four content areas showed greater change in b-values for items than the other content areas. The authors theorized that the changes in the difficulty estimates were attributable to changes in curricular emphasis (i.e., the content area showing the greatest change in item difficulty had experienced pronounced changes in curriculum). Unfortunately, since the study emphasis was on investigating covariates of drift, there was no discussion of the magnitude of item difficulty changes on examinee scores.

Sykes and Ito (1993) used the Rasch model to explore the impact of IPD on the equating process. They analyzed data from administrations of two licensure exams over an 8-year period. They used ANCOVA to look for differences in item difficulties and then to determine whether any differences they found were related to elapsed time or changes in item position. The authors examined whether systematic, non-zero differences existed between pairs of item difficulties in the two item banks. . Some pairs of item difficulties were significantly different, and those differences appeared to be related to elapsed time but not to changes in item position. Additionally, both exams mean item bank difficulty (e.g. the mean difficulty for all items in the bank) had a noticeable change or drift in value. Thus, the authors concluded that the elapsed time between exams has a greater influence on the equating process than changes in item position.

Donoghue and Isham (1998) investigated IPD using the 3PL model. Using 12 DIF methods they looked at the extent to which three common measures (IRT-based, MH-based, and chi-square based) could detect drift in the items across two occasions. They simulated data to exhibit both positive and negative drift in the difficulty and/or discrimination parameters. The 12 DIF methods detected IPD in 75% of the b-drifting items and 44% of the a-drifting items. Overall, the method using the Lord's chi-squared measure was the most effective at identifying items with drift. However, it was only accurate when the c-parameter was constrained to be equal, or only when the parameters associated with a 2PL model were estimated. The other most effective methods included Raju's exact unsigned interval, the NAEP BILOG/PARSCALE chi-square by subgroup method, and the method using Kim and Cohen's closed-interval signed-area measure. Although these methods appeared to work, they require empirical estimates of critical values for the test statistics to function properly, which ultimately diminishes their usefulness. The authors also noted that the ability to detect IPD increased as test length increased, but did not increase with the number of drifting items.

Chan, Drasgow, and Sawin (1999) investigated the effect of time on the psychometric properties of items from the ASVAB (a cognitive ability test battery) across 5 time points within a 16-year period. The analysis included 200 items from eight subtests. The authors plotted and studied item characteristic curves (ICCs) and test characteristic curves (TCCs) to determine whether the items and tests changed significantly over time. Of the 200 items studied, only 25 (12.5%) showed significant changes in difficulty estimates. Tests of general skills and principles had fewer items that exhibited DIF over the years, and tests with more semantic knowledge content had higher rates of items with significant DIF over time. Some tests showed differential test functioning (DTF), but the effect sizes were relatively small and resulted in only a handful of items that needed to be removed to eliminate the DTF. The findings revealed that time does have an effect on the psychometric effectiveness of psychological items and tests. The results also suggest that semantically laden cognitive ability measures are more susceptible to the effects of time compared to other types of cognitive ability tests.

Glas (2000) used simulated data to compare two methods for evaluating parameter drift: the CUSUM method (cumulative sum) and the Lagrange multiplier statistic (LM). The results indicated that both methods were effective in detecting drift, but had different advantages and disadvantages. The LM method supports detection of specific model violations and has the advantage of known asymptotic distributions for the statistics from which it is based. The CUSUM method does not have known distributions for these statistics, but an appropriate critical value can be found via simulations. This actually provides an advantage to researchers, because they can modify the procedure to fit the specific needs of the situation. For example, researchers can choose an effect size to reflect the magnitude of parameter drift that they judge to be relevant in a particular situation. Glas concluded that both approaches are practical tools to monitor parameter drift.

Stahl and Muckle (2007) investigated displacement in Winsteps® (Linacre, 2013) as a means of detecting item drift when using the Rasch model. They simulated test data using a normal distribution for three different sample sizes of test-takers (200, 500, and 1,000) and three different test lengths (30, 100, and 200 items), looking at both the percentage of items in the bank with drift (10%, 20%, 50%) and the direction of drift (easier or harder). Additionally, they looked at the distribution of drift type where there was systematic drift (e.g. item drift in all one direction), an even number of items that become easier or harder over time as well as an asymmetric distribution, where 70% of the items drifted easier and 30% drifted harder over time. Stahl and Muckle then examined the displacement statistic to determine whether the item reflected actual drift or contained an element of statistical artifact (e.g. an anomaly in the difficulty estimate that was the result of the way the item statistic was calculated). Items that have a statistical artifact appear as if they have drift even though their item characteristics have remained stable. The researchers found artificial positive displacement in stable items when systematic drift occurred in one direction, and this was more pronounced in conditions with longer tests and more drifting items. However, the artificial positive drift was not affected by examinee sample size. The asymmetric conditions showed a pattern of positive displacement similar to the conditions with systematic drift, but they were not as pronounced. The artifact of positive displacement was detected more frequently in data sets with large numbers of drifting items. In data sets with balanced drift conditions, where the number of easier and harder drifting items was equal, the artifact of positive displacement was completely ameliorated.

DeMars (2004) also used a time-dependent IRT model to detect IPD over multiple test administrations. He conducted simulations using the 3PL model with 100 items where 10 exhibited drift. He then compared three methods (KPC, CUSUM, and a linear procedure in BILOG-MG) to detect IPD trends across multiple time intervals. There were six drift conditions across five time points that included linear, uneven, and sudden shifts in drift. He found that the linear drift procedure in BILOG-MG and the modified KPC method were very effective in detecting both discrimination and difficulty drift for the magnitudes used in the study (.25, .5, and 1). All three methods had false alarm rates, but all the methods had acceptable error rates that fell within the nominal alpha or criterion level of .01. The BILOG-MG method for detecting difficulty drift was fairly accurate but somewhat overestimated item difficulties in the sudden shift conditions of .5 and 1 logit. The opposite was true for the CUSUM procedure, where difficulty drift detection rates were higher when drift occurred as a sudden shift rather than as a gradual shift for both the linear and uneven conditions. The CUSUM procedure only detected discrimination drift when the difficulty drift was small and the discrimination drift was large and negative. For all three methods, there was a relationship between the detection rate and the amount of drift. Detection of both difficulty and discrimination drift was higher when there were larger amounts and higher magnitudes of drift. Overall, when detecting drift, the BILOG and modified KPC procedures were almost always more powerful than the CUSUM procedure.

Li (2008) investigated the effect of dimensionality on IPD using the 3PL model to analyze the Examination of the Certificate of Proficiency in English (ECPE). The responses for over 70,000 examinees to 30 linking items measuring grammar and vocabulary were evaluated for drift in difficulty and discrimination parameters across three test administrations. The author examined arbitrary combinations of items to evaluate the effects of models that used different dimensionality structures on IPD. The effects of multidimensionality on IPD were explored using models with four structures (one-, two-, and three-dimensional—one with three individual unidimensional structures and one with an underlying three-dimensional structure). Overall, the results indicated that the item difficulty estimates showed a high degree of invariance for all the items; however, multidimensionality did not result in violation of the invariance property. Models that had structures with fewer dimensions showed less invariance in the estimation of item difficulty parameters than models that had structures with more dimensions. The results suggest that the estimation of item difficulty parameters are robust and remain stable in models that have both unidimensional and multidimensional structures. However, the item discrimination parameter was found to be less invariant than the item difficulty parameter. The invariance of item discrimination increased as the dimensions in the models increased. This suggests that some item discrimination estimates are more stable when there are multiple dimensions. . Thus, there is evidence that multidimensionality affects item discrimination parameter invariance (i.e., models with multidimensional structures exhibit less variation in the discrimination parameter than models with unidimensional structures. This finding suggests that the choice of model and dimensional structure for calibration and linking has an effect on the IPD detection. Li concluded that high amounts or magnitudes of IPD detected in tests with a unidimensional structure might be indicative of an inadequate representation of the dimensionality of the test as opposed to IPD.

2. IPD in FIT and test performance

Questions concerning the impact of IPD on accurate measurement led to studies investigating both the nature and effects of IPD. These studies considered how IPD impacts equating methods, scale score drift, and accuracy in ability measurement and pass-fail decisions, with varying conditions and across different IRT models. Although this line of research has the potential for a large concentration of studies , there are only a handful of studies that have this focus.

Stahl, Bergstrom, and Shneyderman (2002) examined the impact of item drift on examinee measurement in a simulation study using FIT or CBT. The authors used the Rasch model to simulate test data for 200 items and 200 examinees, both normally distributed. Various magnitudes of drift (.1, .2, .3, and .5 logits), percentages of items with drift (5%, 10%, 15%, and 25%), and directions of drift (easier, harder, and both) were examined. The recovery of ability estimates ranged from .98 to .99, and only 3 of the 324 total misclassifications (i.e., examinees misclassified as passing or failing) across all the conditions fell outside the 95% confidence interval (CI) band of the cut score. The authors concluded that examinee measures estimated with the Rasch model were robust, even in the presence of extensive item drift, and that undetected item drift has a minimal impact on pass-fail decisions.

Wells et al. (2002) also examined the effect of IPD on examinee ability estimates for two FIT lengths: 40 items and 80 items. Using the 2PL model, the authors simulated test data for two sample sizes: 300 and 1,000. For both test lengths and sample sizes, three types of drift (a-drift, b-drift, and ab-drift) as well as four percentages of item drift (5%, 10%, 15%, and 20%) were evaluated. They used RMSE to determine the recovery of parameter estimates. RMSEs were found to be similar across the conditions for both percentage and type of drift; however, RMSEs were smaller for the larger sample size. The effect of test length was similar for the recovery of the a-parameter estimates, but for b-parameter recovery RMSEs were higher for the 80-item test. Additionally, RMSEs were higher for larger amounts of drift, and the authors reported a larger effect on theta estimates in conditions with 1,000 examinees. Although the authors observed that IPD had minimal impact on ability estimates, these results do indicate that sample size and percentage of IPD impact theta estimates and the measurement of drift. All types of drift have a greater impact on theta estimates when there is a larger percentage of drifting items, and both a- and ab-drift have a greater impact on theta estimates with larger sample sizes.

Witt, Stahl, Bergstrom, and Muckle (2003) looked at the impact of item drift using non-normal distributions. They used the Rasch model to examine the impact of IPD on estimations of examinee ability and pass-fail status using empirical item parameters from a credentialing exam and simulated test responses. Eighteen drift conditions were simulated for two test scenarios: (a) a 100-item test for 187 examinees and (b) a 200-item test for 260 examinees. The authors evaluate the effect of IPD on ability estimates using correlations between true and estimated abilities, and they evaluate the effect on pass-fail status by noting the number of misclassifications—false positives (FP) and false negatives (FN)—in comparison to the baseline data (i.e., examinee test response data without the presence of IPD). The results showed that under the baseline condition, correlations between estimated and true abilities ranged from .81 to .92 for the 100-item test and from .95 to .97 for the 200-item test. Correlations for the drift conditions were very similar to the baseline, ranging from .85 to .94 for the 100-item test and from .96 to .97 for the 200-item test. Across all 18 conditions, the total number of examinee misclassification as passing or failing was 187 for the 100-item test and 260 for the 200-item test. Of these misclassifications only 4 occurred outside the normal error rate (i.e., 95% CI band) for the 100-item test, and only 7 for the 200-item test. In all cases, the number of misclassifications was within the range of what would be expected as a result of measurement error alone. In licensure and certification testing, FP classifications (i.e., those classified as passing who should have failed) are considered worse than FN classifications (i.e., those who fail but should have passed). For this study, FNs outnumbered FPs under all drift conditions for both tests. Additionally, the authors concluded that a fairly large number of item difficulties must be altered (25%) before even a hint of possible distortion in ability estimates appears. These results provide further evidence of the robustness of the Rasch model to estimate ability in the face of undetected drift, even when items and examinees are not normally distributed.

In a theoretical study, Rupp and Zumbo (2003) evaluated the robustness properties of the 1PL, 2PL, and 3PL models under IPD. The authors looked at the drift across all items and examined the overall effect on examinee ability parameter estimation by calculating the cumulative effect of the per-item differences in probability. They found examinee ability estimates were minimally changed except in cases where IPD was large. This held true for all three models. In another study Rupp and Zumbo (2006) looked at the effects of parameter invariance on examinee ability estimation in unidimensional IRT models. The authors used simulated data and characterized IPD as a lack of invariance (LOI), where *LOI* implies parameter values that are not identical in separate examinee populations or across separate measurement conditions (Rupp & Zumbo, 2006). To investigate different types and magnitudes of effects introduced by LOI they used the mathematical formulization of parameter invariance (i.e., the presence of parameter invariance in the mathematical equation of the measurement model) to examine three linear transformations of LOI: algebraic, numeric and visual. When looking at pairs of items different forms and magnitudes of LOI effects are produced under different transformations. The researchers used these three linear relationships (algebraic, numerical, and visual) between pairs of item parameters to examine the magnitude and effect of IPD on examinee response probabilities. The researchers found that LOI and item parameters have a complex relationship due to the numerous differences in item difficulty and discrimination found in practice. Therefore, theoretical representations of LOI cannot yield a general answer as to the effect of LOI on the estimation of examinee ability, because the type and magnitude of the effect are dependent on the actual examinee and item characteristics. However, the results did suggest that inferences about examinee ability based on IRT measurement models are robust for low-to-moderate amounts of LOI across a wide range of theoretical conditions.

Jones and Smith (2006) examined the impact of IPD on pass-fail decision making in a certification exam. Using the Rasch model, they observed the proportion of items that had drifted as a function of average item exposure (i.e., measures in drift over time), the distribution of drift magnitudes, the direction of drift (i.e., easier versus harder), and the consistency of observed drift (i.e., once drift is observed how consistent is that drift condition over time). They reported that in most cases the impact of IPD on examinee scores was minimal—just a fraction of the SEM (0.03−0.2) was attributing to the presence of IPD in items. For two conditions, the average impact was 0.5 of the SEM, and for only one condition was the impact large (1.43 of the SEM), reflecting a significantly overexposed exam. Although the numbers of items that drifted easier and harder were symmetrical (14 to 13), the impact on pass-fail decisions was in one direction only (i.e., all misclassified examinees went from pass to fail). Item drift values were more extreme when items drifted harder compared to when items drifted easier (i.e., the change in item difficulty was larger with harder drifting items). The biggest impact on the pass-fail decisions was at the cut score of +1 logits, where there was only a 92% pass-fail agreement between the original ability estimate and the recalibrated estimate. IPD impacted the pass-fail classification for 8% of the examinees, whose ability estimates were all outside the 95% confidence interval. When testing 500 examinees, this would translate to a misclassification of 40 people.

In their research on IPD and equating, Sykes and Ito (1993) also examined the effect of differences in item difficulties on past exam cut scores and pass rates. They found that differences between cut scores from the actual forms and cut scores adjusted for item drift ranged from one to five raw score points. They also noted that the pass rates were unstable over the eight years of test administration. These results highlight the negative impact IPD can have on examinee test scores and suggest that IPD can threaten the validity of decisions based on these scores.

Using simulated data with the 3PL model, Huang and Shyu (2003) also looked at the impact of IPD on equating and ability estimates. They examined both a-parameter (discrimination) and b-parameter (difficulty) changes in mean scaled scores and passing rates if the item drift were ignored. Their results indicated that the presence of both a- and b-parameter drift in test data had statistically significant effects on scale scores and pass rates. Additionally, they found that sample sizes of examinees and percentages of items with drift had significant impacts on scaled scores and passing rates. Although the effects of a-parameter drift on scale scores and passing rates were statistically significant, they had no practical significance. The scaled scores and passing rate changes were non-existent or non-meaningful changes. However, the effects of b-parameter drift, sample sizes of examinees, and percentage of common items with drift on changes in scaled scores and passing rates did have practical significance.

Skorupski (2006) also investigated the effect of IPD on equating test scores with the 3PL model. The author looked at the impact of drift while trying to recover differences in group characteristics across administrations for 5,000 examinees and 50 items. Skorupski used simulated responses for tests containing equating items exhibiting various levels of drift. The study design included four drift conditions: direction (positive and negative) and two different magnitudes (.5 and 1.0 logits). It also incorporated three item difficulty levels (easy, medium, and hard) and three changes in the mean ability level to represent ability growth across years (0.0, 0.25, and 0.5). The outcome of interest was the impact of drift on the recovery of differences in the ability distributions, which Skorpski evaluated using the mean-sigma (MS) equating method. According to the findings, IPD can create large errors when recovering mean differences between groups across two administrations. The MS equating method recovered true growth ability reliably in the no-drift condition, but growth ability estimates were substantially over- or underestimated in the IPD conditions. Across all conditions, negative b-parameter drift caused the MS method to overestimate changes in the ability distribution, and positive b-parameter shift caused group differences to be overestimated. In the 0.0 and 0.25 logit ability growth levels, easy items showed the most bias, medium difficulty items showed less bias, and the hard linking items showed the most bias. Conversely, in the 0.5 growth level, the hardest linking items showed the most bias, the medium difficulty items showed less bias, and the easier items showed the lease bias. These results imply that drift in harder items has a greater impact on higher ability levels, and drift in easier items has a greater impact on lower ability levels.

Song and Arce-Ferrer (2009) compared three methods for detecting IPD in a common-item, non-equivalent group equating design. They simulated test response data using the 3PL model for 1,000 examinees and 50 items. Three factors of drift were examined: two percentages of items with drift (10% and 25%), three types of drift (a-drift, b-drift, and ab-drift), and three logit-shift decreases for each type of drift (.8, .4, or .2 for b-drift; .5, .3, or .15 for a-drift; and .5 and .8, .3 and .4, or .15 and .2 for ab-drift). Using the MH, Raju's signed area (SA), and unsigned area (UA) methods, the authors evaluated the effects of IPD using three criteria: bias, RMSE, and classification of examinees into below, proficient, and advanced categories. Overall, differences between the performance of the three methods were negligible when there was only a small amount of drift (i.e., none of the methods were particularly good or accurate at identifying drift items and recovering linking coefficients). The three methods did not perform equally effectively in the conditions with the largest amount of drift (i.e., the .8 logit shift b-drift condition, the .5 logit shift a-drift condition, and the .5/.8 logit-shift ab-drift condition). For b-drift, use of the MH method resulted in the fewest FP and FN classifications and the smallest bias and RMSE values. The MH method also performed best in terms of recovery of the item intercepts, but it was the most ineffective at identifying a-drift. The SA method was the least effective in detecting drift in any condition. The UA method performed the best in the a-drift conditions and in the recovery of the item slopes. These results indicate that the methods for IPD detection differed in their effectiveness. When it came to classifying examinees, all three methods underestimated the percentage of examinees classified into the below category and overestimated the percentages of examinees in the proficient and advanced categories. There were more changes in classification for examinees in the below and proficient categories (i.e., a range of 1.96‒5.4%) than for examinees in the advanced category (i.e., 0% to 1.33%). The performance of the three methods in classifying examinees was similar to their effectiveness in detecting target drift items. The MH method was most effective in detecting b-drift, and the UA method was most effective in detecting a-drift. For the ab-drift conditions, the MH method was better at the recovery of item intercepts, and the UA method was better at the recovery of item slopes. The SA method was again the least effective in classifying examinees for all drift conditions. Overall, the differences in passing rates were small—less than one half of a percentage point. When drift in item difficulty was small, passing rates were minimally impacted, but bigger differences in passing rates occurred when a-drift and ab-drift were large.

Wollack, Sung, and Kang (2005) investigate the longitudinal effects of IPD on scale scores using empirical data from a German placement test that had 55 items. The test was administered over a 7-year period to 750‒1,500 examinees. The authors also evaluated the impact of IPD on examinee ability under ten different IRT linking designs. Their results indicated that the choice of linking or IPD model could have a large effect on ability estimates and on passing rates. When using the TCC method with indirectly linked test forms, the results between true and estimated thetas were consistently different in the common item equating model. The authors also detected these differences when they used the common item TCC linking model to determine whether the current and anchor form showed evidence of IPD. From a theoretical perspective, the differences in performance of the ten models were large, but they were not large from a practical point of view. That is, the total number of drifting items and the magnitude of drift were small. Furthermore, differences between the expected true scores and the ability estimates that the models produced were negligible. This result might suggest that IRT is sufficiently robust to IPD and can still estimate examinee ability reliably when IPD is present. However, the authors note that they could not determine from their results whether one of the ten models used in the study was robust to IPD, or how the models would perform under different types, amounts, or magnitudes of IPD.

Using the Rasch model, Meyers, Miller, and Way (2009) evaluated item position and changes in item difficulty in an IRT-based common item equating design. They used empirical data from a 27-item math test and a 48-item reading test administered in grades 3‒8. They employed regression analysis to examine changes in Rasch item difficulty (RID) estimates for both math and reading as a function of item position change, grade level, objective, and time between field and live testing. The authors reported that item position change accounted for 56% of the variance in math RID changes and 73% of the variance in reading RID changes. This implies that placing items near the end of the test has a greater effect on their item difficulty estimation than placing them at the beginning. The authors also found that position change from field to live testing impacted RID. However, the observed effects were mitigated because of the way the testing programs ordered the items: easier items at the beginning and end of the test, and difficult items in the middle of the test. The authors then followed up with simulated data to illustrate further the effects of these changes on difficulty estimates. They used 281 items and various examinee sample sizes (500; 1,000; 2,000; 2,500; 5,000; 10,000; 20,000; and 100,000). The results from the simulation study revealed measurable effects for test equating when items were ordered from easiest to hardest, where the effects modeled the relationship between field and live testing. Difficult items became more difficult when they were moved toward the end of the test, and easier items became easier as they were moved toward the beginning of the test. Thus, when items are ordered from easy to hard, the test appeared more difficult than it truly was for students of higher ability and easier than it truly was for students of lower ability. In such cases, higher ability students would benefit from inflated ability measures, and lower ability students would be disadvantaged because of the underestimated ability measures. However, when items were ordered in the simulation in the same way as in the operational test (easier at the beginning and end of the test), the results mimicked those from the regression analysis. The item position difficulty effects that were due to position change between field and live testing canceled each other out.

Kingsbury and Wise (2011) conducted a long-term study of the stability of item parameter estimates for the purpose of creating a K‒12 adaptive test. They used empirical data from the Measures of Academic Progress (MAP) with 3,091 math items and 1,728 reading items administered to 100,000 examinees from grades 2 through 10 in ten school districts across seven different states. They conducted two main analyses using a Rasch model framework: a study of scale drift and a study of impact. They evaluated scale drift by correlating new and original difficulty estimates, bias estimates, and mean absolute differences. The impact analysis examined the extent to which the scale changes affected examinee test scores (i.e., differences between examinee's original and new scores). They evaluated the effects on examinee test scores using bias estimates, mean differences, and maximum changes between the original and new scores. The results showed high correlations between the original and new item difficulty estimates: .967 for math and .976 for reading. The average change in difficulty estimates was -0.11 for math and -0.17 for reading. The authors' analysis of the impact of length of time indicated that no substantial drift occurred in the scale values, and the new difficulty estimates did not vary systematically as a function of time. The largest change between the original and new examinee test scores was only 1.1 points for both math and reading and 99% of the expected changes were only less than 1 point. Therefore, the drift that did occur in the items had an almost non-existent impact on examinee scores.

In an attempt to define when to reset an exam's score scale, Babcock and Albano (2012) evaluated scale drift over time with the Rasch model. Using simulated data with 500 examinees, 220 items, and two test forms, they examined how drift affects item parameter recovery, scoring, and classification of examinees at both the item and trait level. The study was a fully crossed design and included five levels of item drift proportions (.00, .05, .10, .50, .20), three levels of direction of drift (increase, decrease, and both), and four levels of changes in the latent trait (i.e., observed increase or decrease in the latent trait ability level over time) (0%, 1%, 5%, 10%). They investigated pass rates, misclassifications, and person fit statistics to determine the effect of scale drift on examinee classification. In addition, they assessed theta value recovery with RMSE and bias statistics. They used a factorial ANOVA to analyze their simulated data and calculated effect sizes. The results indicated that when the latent trait changed by 10% there was a substantial impact on the recovery of the theta estimates. As expected, the RMSE increased as the proportion of drift increased. Also, similar to findings in other studies, bias cancelled out when the drift occurred in both directions. When item drift increased or items became harder, more of a biasing effect was seen than when items drifted easier. Because the item pool was already fairly easy, the item drift in the easier direction did not have as significant an impact on examinee scores. However, both easy and hard drift had a substantial impact on the theta estimation. The results for the effects on classification accuracy were similar. Compared to the baseline condition, there was no difference in classification accuracy when drift occurred in both directions. However, more failing classifications occurred in the harder drift conditions, and fewer failing classifications occurred in the easier drift conditions. The true pass rate for the exam was high because the cut point was low relative to the distribution of examinee ability. Therefore, there was a higher risk of classifying examinees as failing as opposed to classifying them as passing. Thus, the easier drift condition had fewer "risky" failing classifications. The opposite would have been true had the pass rate been low and the cut point been high. Items drifting easier would have inflated true ability scores, causing a higher risk of classifying examinees as passing.

Additionally, as the number of years between testing occasions and proportion of drift increased, there was only a slight decrease in examinee fit statistics. Thus, it was determined that the examinee fit statistic is not very sensitive to overall shifts in the measurement scale. The authors concluded that under a small amount of item drift and small to moderate changes in the latent trait, a Rasch scale may remain stable for 15 years (+/−3). However, they cautioned that substantial item drift or large changes in the latent trait could drastically reduce the longevity of the scale. These findings provide evidence that IPD, when unaccounted for, has the potential to seriously impact the effectiveness of the exam scale and thus threaten the validity of examinee ability measures and accuracy in pass-fail decisions.

In a study investigating the impact of compromised anchor items on IRT equating, Jurich, DeMars, and Goodman (2012) found that compromised items can have a substantial impact on estimating examinee ability. Their study used simulated data with 100 items and 3,000 examinees under a non-equivalent anchor test design. The study included conditions with four proportions of cheating examinees (5, 10, 25, and 50), two proportions of compromised items (25 and 100), four ability distributions (M: 0, SD: 1; M: −.5, SD: 1; M: 0, SD: 1.25; and M: −.5, SD: 1.25), and two anchor item methods (external versus internal scoring to the test). Their results showed that an increase in the proportion of compromised anchor items or cheaters resulted in positively biased equated scores. As one would expect, if the item was part of the internal anchor item set (i.e., the item was included in calculating the examinee's score), cheaters received inflated number correct scores and thus inflated ability estimates. However, inflated scores still occurred when compromised anchor items were part of the external anchor item set (i.e., the item was not included in calculating the examinee's score), and honest test takers benefited as well. The authors concluded that this was because the B scaling constant was overestimated when cheating occurred on items used to scale the test form. Therefore, the difficulties of the anchor items were underestimated for the group taking the new form (NF). Overestimating the B scaling constant caused the difficulties of the unique items to increase artificially. The inflated NF's b-parameters then caused the unique items to appear more difficult, which increased the ability estimate when examinees responded correctly to the items regardless of whether the examinee was cheating. The extent of the bias at even moderate levels of cheating was somewhat large, which suggests that equated scores obtained from even slightly compromised test forms will overestimate examinees' true abilities. These results are another example of how IPD has the potential to negatively influence ability estimates and pass-fail decisions.

D. IPD and CAT

1. Detecting IPD in CAT

Unfortunately, the procedures used to detect IPD in FIT previously described are not appropriate for CAT. In most cases, the methods require that items be recalibrated, but CAT data is not ideal for continuous recalibration because many items are only taken by examinees who have similar ability levels. The following research studies have extended IPD research by detecting or examining IPD in CAT environments.

By looking at items through their life cycle (pretest, active, retired), Bergstrom et al. (2001) identified factors that influence IPD in CAT with the Rasch model. They used empirical data from an adaptive licensure exam with 1,000 examinees and four operational item banks. Each test contained approximately 70–140 items generated from one of the four item banks. Some items were present in more than one item bank. The authors looked at six changes in the state of item use (pretest-pretest, pretest-active, active-active, active-pretest, pretest-retired, active-retired). They detected drift by looking at changes in item difficulty values using several statistics: the mean-centered difference, the standardized difference, and the cumulative sum of the standardized difference. The mean difference indicates the magnitude of drift, the standardized difference indicates how important the shift in difficulty is given the standard error of the item calibration, and the cumulative sum of the standardized difference allows for observation of trends in IPD. The results revealed that mean item exposure (i.e., the average rate examinees are exposed to items) varied for items in the pretest banks and less drift was found in situations where fewer items were pretested and the time period was only one year after initial calibration. The standardized difference of item difficulty values increased from pretest to active. This was due to exposure to a more appropriate sample and a better definition of item difficulty at the ends of the item distributions (i.e., difficult items got more difficult). Exposure also played a role in IPD. Items with drift that were exposed to too few examinees were not identified, and items exposed to too many examinees were classified as drifting, even if the logit shift was relatively small. Additionally, higher item counts within a bank and higher volumes of examinees impacted the ability to monitor and account for drift. Thus, the size of the item bank, the number of examinees, and the exposure rates of items all influenced IPD detection.

Lu and Hambleton (2003) proposed an item fit analysis method to detect item drift in CAT. They used simulated data to model drift in some items and then examined statistics based on item residuals and likelihoods to identify differences in item parameters. This approach does not identify drift in items across time, but rather points out aberrant response patterns that likely result from IPD. The results showed that the item fit analysis detected over 70% of the drifting items with a 5% type I error rate.

Han (2003) also investigated IPD in CAT by using a technique that looks at moving averages of item difficulty. Plots of item p-values are generated within successive time intervals and used to evaluate drift. When the p-values for examinees in later test administrations differ from the p-values for examinees in earlier administrations, it indicates that item drift has occurred. Although this approach was successful in identifying drifting items, it is not always plausible in practice, because it assumes that comparable populations of examinees from one test administered to another.

Hatfield and Nhouyvanisvong (2005) examined parameter drift in a high-stakes CAT licensure exam with anchor items. They used 440 registered nurse (RN) items and 447 practical nurse (PN) items from the NCLEX to examine the degree to which IPD was evident in the anchor items. Using a hierarchical linear model (HLM) approach, the authors tried to determine whether item parameters tended to increase or decrease across time and whether those changes were related to specific factors. Their results showed no evidence that b-values, point-biserial correlations, or item response times systematically increased or decreased over time. Thus, the authors concluded that systematic drift was not compromising the validity of the NCLEX exam.

Masters, Muckle, and Bontempo (2009) compared methods to recalibrate drifting items in CAT, using empirical data with 450 examinees and 152 operational items. They examined whether applying the displacement statistic to drifted items could account for the drift. The authors assessed whether calculating a new difficulty value (i.e., adding the displacement to the original calibration) or recalibrating the item in another pretest better accounted for drift. They then compared the adjusted calibrations to the new calibrations. Their results showed a high correlation between the adjusted and new calibrations for drifted items. The difficulty measures of 40 of the 152 items were statistically significantly different. However, the actual magnitudes of the differences were small. Overall, the findings provided mixed support for use of the displacement value to adjust the calibrations of drifted items.

Studies detecting drift in CAT typically involve only two time points; however, Deng and Melican (2009) used a 3PL CAT program to evaluate IPD at multiple time points. Using empirical data, they examined operational items in a placement exam over a 4-year testing period. Each exam administration had over 200 examinee responses. IPD was found in a very small number of items, even over the four-year period. The authors claimed that this result was not surprising, given the nature of the exam: a low-stakes student placement test.

Meng et al. (2010) investigated IPD in a fixed-length adaptive test under the 3PL model. They used three years of real test data with 252 items and sample sizes of 10,706, 11,693, and 11,895. Drift was evaluated using a non-compensatory differential item functioning (NCDIF) index for two calibration methods: fixed-item and fixed-person. In total, 44% of the items showed severe drift and needed recalibration. The fixed-person method identified more items than the fixed-item method and the ICC's from the fixed-person method were more analogous to the observed item success rates (i.e., the rate examinees correctly answered the item). In addition, the fixed-person calibration method outperformed the fixed-item method for analyzing the simulated data. That is, the fixed-person method had more power, and the index identified more items with IPD. However, the fixed-item method was better at identifying items that drifted harder and became more discriminating. Based on these results, the authors recommended that both fixed-item and fixed-person NCDIF methods should be used to identify and evaluate drift.

Although slightly different from IPD, a study by Guo and Wang (2003) found minimal impact on score stability and scale drift in their evaluation of online calibration and scale stability from a large-scale operational CAT program. The study objectives included developing an online data collection method to study scale stability using both real and simulated data. The authors used real CAT data to obtain the ability distribution. They generated the item parameters for the simulations from the 31 items that were linearly administered at two time points 20 months apart. They evaluated the stability of the scale using the TCCs and the ICCs by comparing the online calibrations for the two time points. The findings indicated good scale stability for this particular CAT program, and the authors claimed that it is acceptable to apply the design and methods used in this study to monitor scale stability over time to other CAT programs.

2. IPD in CAT and test performance

Despite the important implications of IPD, few researchers have investigated how IPD impacts the measurement of ability and the classification of examinees. Since not all examinees are exposed to the same items during an adaptive exam (McCoy, 2009), the presence of IPD is even more important to monitor because it increases the potential for bias. This section details the few studies that have explored the impact of IPD on examinee ability estimates and classification accuracy in CAT.

The study by Guo and Wang (2003) previously described also explored the potential impact of *scale drift* on test scores. Scale drift refers to the instability of a test score across two or more test forms (Gou & Wnag, 2003). Using the observed and simulated scores, the authors evaluated the bias in ability measures and changes in test scores at two time points. The results showed that the tests were slightly easier at the second time point, which led to lower ability estimates at the second time point. The changes in ability estimates indicate that bias was present, but the overall effect was small and trivial in a practical sense. They also observed changes in test scores, indicating that the scale drift did impact test scores. However, these changes were also small (i.e, just over have a point lower between the two time points). Despite the minimal impact to the test scores the authors point out that by using the observed changes in TCCs, it is not difficult to infer how the score impact would change if the scale drift had been larger. If an exam's scale difference between two test forms was larger than the scale difference observed in this study, the measurement precision for the exam would be noticeably different at two time points. Therefore, in order to compare scores from the two testing time points it would be necessary to adjust the scale. This finding suggests that even ability measures obtained in CAT are susceptible to error due to severe drift.

McCoy (2009) determined whether IPD was present in a CAT item bank and assessed its impact on examinee ability measures using the Rasch model. He analyzed empirical data from a high-stakes licensure exam with 2,555 examinees and 1,270 items across eight content areas. McCoy applied a Rasch Longitudinal Model (RLM) framework, a Rasch modified HLM approach that extended the Rasch model to control for IPD. The RLM model showed a better model fit than the Rasch model to this date. He reported a minimal-to-moderate presence of IPD in each content area, with a range of 2–8 items demonstrating IPD in each subscale. The hematology subscale had the most items showing IPD, but there was no evidence that one subscale was significantly more prone to drift than the others. The RLM model also provided more accurate examinee ability estimates than the Rasch model when IPD was present. His research results also demonstrated that changes in pass-fail decisions could result when examinee ability estimates are computed accounting for IPD, even after adjusting for measurement error. Although he detected changes in pass-fail decisions, they were minimal across all scales with the largest change being only 0.6%. Thus, the presence of IPD would have resulted in pass-fail decision changes for only 9-24 of the total 2,555 examinees.

In their analysis of IPD in eCAT (a computerized adaptive test to assess the written English level of Spanish speakers) Abad, Oleo, Aguado, Ponsoda, and Barrada (2010) found IPD to have a negative impact on ability estimates. They used the 3PL model to analyze their data. The sample included 7,254 examinees and 3,224 items. They evaluated drift by comparing test administrations at multiple time points and conducted a DIF study using the original and new item calibrations. The authors found significant item drift in a fair number of items, especially in the a- (discrimination) and c- (guessing) parameters. The authors also examined the impact of the new item calibrations on ability measure estimation through simulation. The new item calibration due to change in the a- and b- (difficulty) parameters showed a moderate impact on theta estimates for the most proficient English examinees. Therefore, the authors recommended to replace the original item calibrations with the new calibrations.

Hagge, Woo, and Dickison (2011) found similar results in their study. They investigated the impact of item drift on examinee ability estimation in a variable length CAT using the Rasch model. Their sample included a large item pool from a high-stakes licensure exam for two test administrations, with over 18,000 examinees for the first administration and close to 53,000 examinees for the second administation. They examined how robust examinee ability estimates were in the presence of IPD and to what extent the pass-fail decisions were impacted when drift occurred. They looked at several drift conditions that varied the percentage of items in the bank with drift (5%, 10%, and 20%), the magnitude of the drift (.5 logits, .75 logits, and 1.0 logits), and the direction of drift (easier, harder, and both). The study was a fully crossed design. They evaluated the differences between recalibrated and original examinee ability estimates and pass-fail decision consistency. The results indicated that as the percentage of drift items and the magnitude of drift increased, so did the differences in the theta estimates. The largest difference was .40 logits when 20% of the items had drifts of 1.0 logits. This finding was consistent in the pass-fail decision consistency as well, where consistency was greater than 95% for all conditions, except when 20% of items in the bank had drifts of .75 or 1.0 logits. These findings confirmed results from other studies that suggest examinee ability estimates are robust to item drift in large operational pools, especially for conditions that may represent normal drift. The decision consistency was still high, even under extreme conditions (i.e., 20% of the items had drifts of 1.0 logits).

In a similar line of research, a handful of studies focusing on compromised CAT items (a previously identified type of IPD) have examined their effect on item parameter estimation and ability estimation. The results of these investigations have consistently found a significant amount of positive bias in the ability estimates (Jurich, Goodman, & Becker, 2010; Yi, Zhang, & Chang, 2008; Guo, Tay, & Drasgow, 2009). Yi, Zhang, and Chang (2008) investigated the effects of cheating on ability estimation using a CAT. Under various CAT selection criteria, the authors compared the error in ability estimates that resulted from compromised items. The results indicated that there was severe positive bias in the ability estimates. For low-ability students, the mean differences between estimated and true abilities ranged from 0.89 to 3.88 logits, and on average the increase in ability measure was over 1 SD. The authors also noted that compromised items had less influence on examinees with higher initial true ability. Gou et al. (2009) found similar results in their study examining the resistance of CAT to small-scale cheating. Their findings showed that the presence of compromised items led to a drastic overestimation of ability for the low-ability students. The overestimation in scores impacted the test's reliability, resulting in an exam that was unable to discriminate among examinees.

Using IRT observed-score equating, Jurich et al. (2010) looked at how the presence of compromised items in a bank impacted the pass-fail decisions for a CAT. They evaluated three different types of scaling methods that they used to equate the base form to the new form: mean-sigma, Stocking-Lord, and fixed anchor. The authors compared the recovery of the correct pass-fail examinee classifications after analyzing the data using the three methods. As in the previous studies, the results showed a significant inflation of examinee pass rates for each method. However, unlike the studies reported previously, both cheaters and honest examinees who took the new form benefited from the compromised items (i.e., received inflated ability measures). Based on this finding, the authors theorized that when anchor items are compromised, the scaling methods incorrectly adjust for differences in ability. This incorrect adjustment then benefits all examinees.

E. Summary of Literature and Proposed Study

The presence of drift in an item calls into question the accuracy of ability estimates and pass-fail decisions made from the test scores. IPD research is limited to date, especially research investigating the impact of IPD on ability estimation. A few studies have demonstrated that the effect of IPD on theta estimation is minimal (Chan et al., 1999; Wells et al., 2002). However, other studies have shown negative impacts on ability estimation and pass-fail consistency (Jones & Smith, 2006). Most of the studies of IPD have examined drift in fixed-length or fixed-item testing, and procedures used in these studies are not necessarily appropriate for CAT. As the literature review points out, research on IPD in CAT is sparse, and research regarding measurement precision in CAT is limited even further. Only a handful of studies have explored the impact of IPD on examinee ability estimation and classification accuracy in CAT. Some studies have shown promising findings that indicate that CATs are fairly robust to IPD (i.e., there are minimal effects on theta estimation); however, there is some research that has found IPD to negatively impact measurement and influence pass-fail decisions.

The goal of my research is to expand the literature on how various aspects of IPD might impact measurement precision and test efficiency in CAT. I investigated the impact of various amounts and magnitudes of IPD in a CAT item bank on measurement precision, pass-fail classifications, and test efficiency. Additionally, I studied various CAT item bank sizes to determine how bank size impacts IPD. To set up my simulations I used operational certification exam specifications and then analyzed the data using the Rasch model. Based on the results I observed, I offered recommendations to testing organizations that use CAT about how to deal with IPD and maintain their item banks.

III. METHOD

A. Research Questions

1. What amount of drift can be present in the item bank before examinee ability estimates, pass-fail decisions, and the overall efficiency of the exam become compromised?
2. What magnitude of drift has the greatest impact on examinee ability estimates, pass-fail decisions, and overall exam efficiency?
3. Do the effects of IPD on examinee ability, pass-fail decisions, and overall exam efficiency vary by the size of the item pool?

B. Overview

I performed a simulation study to investigate the effects of various aspects of IPD on examinee ability estimation, classification accuracy, and efficiency in a CAT exam. The study was a fully crossed design with a total of three variables manipulated across the conditions. These included two factors of drift: number of items and magnitude of IPD. My third variable was the size of the item pool. I evaluated the various conditions on a number of criteria and determined the extent to which the IPD jeopardized the measurement precision, confidence in pass-fail decisions, and efficiency of the exam.

I focused on b-drift or changes in item difficulty only, since b-drift is more common than the other types of item parameter drift, and researchers have concluded that a-drift is hard to detect, hard to define and has a minimal impact on theta estimates (Donoghue & Isham, 1998; Song & Arce-Ferrer, 2009).

I created my test is structured to mimic a high-stakes certification exam with similar content restrictions and item difficulty parameters. In the next section of the chapter, I describe the item pool properties, simulation parameters, and evaluation criteria. For the purpose of deriving stable parameter estimates, I generated 100 replications for each condition and then averaged the results over the replications.

C. Independent Variables

1. Amount of drift

In order to evaluate differences in item bank size, I used a set number of items with IPD in each item bank rather than a set percentage of the item bank that had IPD. These numbers of were 100, 75, and 50. I randomly selected the items from the item bank to have IPD. For each simulation, I used the item file with the corresponding number of IPD items (100, 75, or 50) based on the corresponding condition. I then modified the modelled probabilities of a correct response for these IPD items to match the magnitude of drift for the specified condition across all the simulations. For example, if the item was in the 0.5 logit-shift condition, then the item's difficulty became easier or harder by 0.5 logits. To reflect the change in the item's difficulty parameter, the examinee's probability of correctly answering that item increased according to his or her ability level.

I chose to use 100, 75, and 50 items because they represent typical amounts of item banks that are compromised by IPD that researchers have studied: 10%, 15% and 20% (Hagge et al., 2011; Wells et al., 2002). For a medium-sized item bank of 500 items, 10%, 15% and 20% translate to 100, 75 and 50 items with drift.

2. Magnitude of drift

I also used three different magnitudes of drift. The difficulty values of the selected items shifted by 0.5 logits, 0.75 logits, and 1.0 logits. I selected these magnitude values of drift based on the standard error of an item bank (.25) and previous research. Studies evaluating drift or displacement values often recommend setting a threshold value of at least 0.6 to ensure that the researcher is not just evaluating normal error rates (Han & Guo, 2011; O'Neill, 2013).

For each of these magnitudes, I simulated items to become both easier and harder over time. When examinees encountered the drift items, the simulation modified their responses to reflect the magnitude of drift specified by the condition.

To modify the probability of a correct response to an item with IPD, the simulator subtracted the adjusted item's difficulty (i.e., the item's drifted difficulty) from the true item difficulty. The simulator used the adjusted difficulty value and the measure of the true ability of the examinee to calculate the probability of a correct response. The respective item files specified the amount of drift in the bank (100, 75, or 50 items) and the magnitude of the drift (0.5, 0.75, and 1). For items that drifted easier the simulator subtracted the magnitude of drift from the item difficulty (e.g., when the magnitude of drift is 0.5 logits an item difficulty of 2.0 logits becomes 1.5 logits). For items that drift harder the simulator added the magnitude of drift to the item difficulty.

3. Size of item pool

I simulated item pools of varying sizes to examine the impact of IPD when item pools were small (300 items), medium or average (500 items), and large (1,000 items). My choice of item pool sizes reflects both sizes of item banks used in practice and previous research findings. An item bank of 500 items is typical for many certification organizations such as the one I modelled my test after. Researchers have also concluded that a desirable minimum bank size falls in the 400-500 item range, but that a larger item bank (e.g., 800-1000 items) is more efficient (Bergstrom & Stahl, 1992; Lunz & Stahl, 1993). However a smaller item bank of only 200 or 300 items can still be efficient with minimal item exposure when item banks are well targeted to examinee ability (Lunz & Stahl, 1993).

D. **Test Properties**

1. Item pool

I obtained item parameters for this study by duplicating item parameter distributions used on the high-stakes certification exam. I simulated a series of variable length exams based on test constraints defined for the high-stakes exam modeled in this study (Table 1). The table outlines the proportion of items and the minimum and maximum number of items that the simulator could administer from each content domain. I used these criteria to simulate the CAT item pool, which contained 300, 500, or 1,000 items. Note that the distribution of item difficulties mirrored the operational exam (M = .006 and SD = 1.079), for the test as a whole and for each content area. However, the number of items at each item difficulty level varied from the operational exam, because I simulated the item difficulties to three different item pool sizes. The exam covered six content areas and used only dichotomously scored multiple-choice (MC) items.

Table 2

Test Properties and Item Characteristics

Content	**Proportion of Items**	**Range of Item Difficulty**	**Item Difficulty M(SD)**	**Minimum/Maximum Number of Items**
A	12%	-2.91 – 3.06	-.058(1.147)	9/18
B	20%	-3.75 – 2.4	-.274(1.124)	15/30
C	20%	-2.5 – 2.13	.225(.973)	15/30
D	25%	-3.2 – 2.79	-.097(1.085)	19/38
E	14%	-2.54 – 2.89	.167(1.013)	11/21
F	9%	-3.08 – 2.03	-.048(1.128)	7/14
Total	100%	-3.75 – 3.06	.006(1.079)	75/150

2. Examinee distribution

The samples for all simulations consisted of 500 hypothetical examinees. A sample size of 500 is typical in the certification and licensure field (Kim, Barton, & Choi, 2010). I used

Excel to generate the measures of the examinees' true ability based on the ability distribution of the examinee population who took the high-stakes certification exam. Their mean was .926, and their standard deviation was .726. The measures of true ability ranged from about −1.65 to 3.61 logits. A CAT simulator generated the examinees' responses for each replication (Becker, 2013). The data consisted of each examinee's responses to the items administered (i.e., a response string of 01 data). The simulator created this string by first randomly selecting an examinee ability level from the population distribution. The simulator then used the Rasch model for dichotomously scored items and specified item parameters to compute the probability of correctly answering each item for this ability level. The simulator then compared the probability of answering each item correctly to a random number from a uniform distribution with a 0–1 range, U(0,1). If the uniform number was less than the probability of a correct response, the simulator scored the response as correct. If the uniform number was greater than the probability of a correct response, the simulator scored the response as incorrect.

E. CAT Simulation

The program that conducted the CAT simulations used a simulator that Becker (2013) created and modified to meet the specifications for this study.

1. Ability estimation

The simulator used the exam's cut score for each examinee's initial ability estimate. After each response, the simulator estimated the examinee's ability using unconditional maximum likelihood estimation (UCON). UCON is a variation of the joint maximum likelihood estimation method (JMLE) in which the likelihood function is conditioned on the number-right score (Wright & Panchapakesan, 1969). JMLE is an iterative, two-stage procedure where "joint" means that it estimates item parameters and ability parameters simultaneously. When the observed raw score for the item and ability parameters match the expected raw score within a specified tolerance level, then item and ability parameters are estimated (Wright & Panchapakesan, 1969).

2. Item selection

I selected the properties of the initial item according to the content constraints of the exam. The simulator chose the first item from the content area with the largest proportion of items (or the highest weight). In this case, content area D had the most items on the the test (see Table 2); thus, the simulator most likely chose the first item from content area D. If no items in this content area met the statistical criterion (i.e., no items had difficulty values the same as the cut score for the first item, or examinee ability for subsequent items), then the simulator considered other content areas. When none of the content areas contained items that met the criterion, the range of item difficulties from which the simulator could select an item was expanded. The simulator selected subsequent items by comparing the current content distribution of the test to the content outline in Table 2. The content area on the exam with the least amount of proposed items determined the content area from which the simulator choose the next item, and the target item difficulty (i.e., the target probability of a correct response, based on the examinee ability estimate) dictated the actual item selected. If no items from the desired content area met the target item difficulty, then the simulator considered any unadministered items from that content area. When the availability of unadministered items from all content areas was exhausted, then the test ended for the examinee.

3. Constraints (content balancing/exposure control)

The simulator adaptively selected items based on content balancing and exposure control specifications. Content balancing constraints included the six content areas and the proportion of items associated with each content area (see Table 1). The target percentages for the content areas and number of items per content area matched those in the test blueprint of the high-stakes exam. To control for item exposure, I used a *randomesque procedure*. The randomesque procedure selects the 10 best available items for the corresponding theta value. The simulator randomly selects the item for administration from among these 10 items. Each test included a minimum of 75 items and a maximum of 150 items, which is the criterion of the high-stakes certification exam, mimicked in this research.

4. Stopping rule

I used a confidence interval criterion as the exam's stopping rule. That is, the simulator compared the current estimated ability and standard error against the location of the cut score (i.e., passing standard). If the current ability estimate had a 95% chance of being higher or lower than the cut score and the simulator had administered the minimum number of 75 items, the exam terminated. Additionally, the maximum item rule also ended the exam. The exam terminated for those cases where the simulator administered 150 items before it reached the specified 95% confidence interval. In such cases, the simulator compared the examinee's final ability estimate at that time to the passing score to determine whether the examinee passed or failed the exam. If the ability estimate was above the cut score, the examinee received a passing score. If it was below the cut score, the examinee received a failing score.

5. Direction of drift

The direction of drift did not vary across the different conditions. The IPD items had an uneven split (i.e., 75% of the selected items drifted easier, and 25% drifted harder). I made this choice was based in part on previous research findings that showed that an even split of drift washes out (Babcock & Albano, 2012; Hagge et al., 2011; Stahl & Muckle, 2007). Additionally, in practice there are higher incidences of easier shifts in item difficulty, mainly because of overexposure and cheating. Another reason that easier drift is more typical is that the possible sources of an easy shift outweigh the sources of a harder shift (see Table 1).

F. **Evaluation**

I compared study conditions on measures of estimation precision, classification accuracy, and efficiency (see Table 3). For each condition, I calculated the evaluation criteria and averaged them over the 100 replications.

1. Precision

To quantify errors in theta recovery and evaluate the precision of the estimated ability level from the CAT simulations, I used both conditional and unconditional statistics. To assess the overall measurement precision for each condition, I calculated descriptive statistics (mean, SD, and range), the correlations between known and estimated thetas, SEs, bias values, RMSEs, and absolute average differences (AAD). These are commonly used statistics for evaluating estimation precision in simulation-based studies (Guyer & Thompson, 2011; Harwell, Stone, Hsu, & Kirisci, 1996; Lee & Dodd, 2012; Moyer, Galindo, & Dodd, 2012; Wang & Chen, 2005). The bias value is a measure of the systematic deviation of an examinees CAT estimated ability from the examinees true ability. The lower the bias value, the closer the CAT estimated ability is to the examinee's true ability. The RMSE value is a measure of absolute accuracy in parameter recovery, taking into account the bias value and the variability of the sample parameter. Similarly, the lower the RMSE value, the more accurately the CAT has estimated the examinee's true ability. AAD is another measure of average bias among true and estimated ability, where lower values indicate more accurate estimates. (See Table 3 for mathematical formulas.)

2. Classification accuracy

I assessed classification accuracy using the established cut score by identifying the number of false-positives (pass) and false-negatives (fail), as well as the total percentage of misclassifications for each condition. Other researchers have used these same criteria to evaluate classification accuracy (Hagge et al., 2011; Lunz et al., 1993; Stahl et al., 2002). I established the exam's cut score based on the examinee ability distribution, targeted at the pass rate of the modeled high-stakes exam, 70%. This resulted in a cut score at a logit value of 0.59.

3. Efficiency

I evaluated the effect of IPD on exam efficiency using criteria common in simulation studies: total test length and item exposure rates (Lee & Dodd, 2012; Moyer et al., 2012). Since longer tests are less efficient, I examined the number of items administered upon termination of the CAT in each condition. Item exposure is also an indicator of test efficiency, because overexposed items present issues with test security and underexposed items are a waste of resources (Moyer et al., 2012).

a. Test length

To determine the average test length for each condition, I calculated the mean number of test items administered to the examinees. When the difficulty of an item equals the ability level of the examinee ($b = \theta$), the standard error of the estimated value for theta is minimized, and item information is maximized. The overall test information is a cumulative function of item information. In turn, test information determines test precision (i.e., the larger the test information function, the more precise the test). Easier items contribute less information than optimal items (i.e., items well targeted to the examinee ability distribution). Thus, as items become easier, additional items are needed to achieve the desired level of precision. Therefore, as the total number of items administered increases, the efficiency of the exam decreases.

b. Item exposure

The simulator calculated item exposure rates, which are simply the percentages of examinees who were administered each item. I compared the frequency distributions, means, standard deviations, and maximums of these exposure rates across conditions.

Table 3
Evaluation Criteria

Measure	**Measure Index**	**Description**	**Formula**
Precision	Bias	Systematic deviation of estimated ability from true ability.	$\frac{\sum_{j=1}^{n}(\hat{\boldsymbol{\theta}}_{\imath} - \boldsymbol{\theta}_i)}{n}$
	Root Mean Square Error (RMSE)	A measure of absolute accuracy in parameter recovery.	$\sqrt{\frac{\sum_{j=1}^{n}(\hat{\boldsymbol{\theta}}_{\imath} - \boldsymbol{\theta}_i)^2}{n}}$
	Absolute Average Difference (AAD)	Another measure of average bias between true and estimated ability.	$\frac{\sum_{i=1}^{n}\lvert\hat{\boldsymbol{\theta}}_{\imath} - \boldsymbol{\theta}_i\rvert}{n}$
Classification Accuracy	False Positives (FP)	The number of examinees who receive a passing score, but should have failed.	N/A
	False Negatives (FN)	The number of examinees who receive a failing score, but should have passed.	N/A
	Total Percentage of Misclassification	The total percentage of both false positives and false negatives resulting from each condition.	$\frac{FP + FN}{n}$
Test Efficiency	Test Length	The average number of test items administered to each examinee.	$\frac{\sum_{i=1}^{n} \boldsymbol{K_i}}{n}$
	Average Item Exposure Rate	The sum of the item exposure rates divided by the total number of examinees.	$\frac{\sum \boldsymbol{m_k} * 100}{n}$

Note. $\hat{\theta}_i$, θ~i~ represents the estimated and known thetas for examinee *i*, *n* is the total number of examinees in each condition, K_i is the total number of items administered to examinee *i,* and m_k represents the number of times item *k* was administered across all *i* examinees. There is no current industry rule of thumb for satisfactory values for any of the criteria listed.

G. Analysis Approach

1. Response model

I used the Rasch dichotomous model to obtain item and examinee parameters for the certification exam (Rasch, 1960). The dichotomous model is appropriate when examinees' responses to an item are scored as either correct or incorrect (Wolfe & Smith, 2007). The model form is given as

$$P_{nix} = \frac{\exp(\beta_n - \delta_{ix})}{1 + \exp(\beta_n - \delta_{ix})}$$

where *P* is the probability of examinee *n* scoring *x* on item *i*, *x* is the item response coded 1 (correct) and 0 (incorrect), δ is the difficulty parameter for item *i,* and β is ability parameter for examinee *n.*

2. Baseline Condition

The first condition created a baseline recovery rate using the established cut score. The baseline used the content constraints for the operational exam and did not reflect any of the test parameters for the experimental conditions used in this study. From these simulated data sets, I averaged and documented the initial rates for estimation precision, classification accuracy, and the efficiency criteria over 100 replications. I then compared the baseline to the various study conditions in order to determine the extent of change in the evaluation criteria each condition produced.

3. Research questions

To address the three research questions, I used a crossed 3(amount) × 3(magnitude) × 3(item pool) factorial design with a total of 27 experimental conditions so that I could examine both main and interaction effects. I evaluated examinee ability measures, pass-fail decisions, and exam efficiency, comparing the values obtained in each experimental condition to those obtained in the baseline condition. I used the bias, RMSE and AAD statistics to evaluate examinee ability measures. To examine classification accuracy in the experimental conditions, I compared the number of false-positive and false-negative occurrences. I also compared the total percentages of misclassification for each condition to the initial percentages for the baseline condition. For example, I classified an examinee as a false-negative if the examinee scored above the cut score in the baseline condition but below the cut score in the experimental condition. Finally, to evaluate test efficiency, I compared test lengths and item exposure rates from each of the experimental conditions to the values obtained from the baseline condition.

IV. RESULTS

In this chapter, I present the findings from the analyses I carried out to answer my research questions. The research questions are as follows:

1. What amount of drift can be present in the item bank before examinee ability estimates, pass-fail decisions, and the overall efficiency of the exam become compromised?
2. What magnitude of drift has the greatest impact on examinee ability estimates, pass-fail decisions, and overall exam efficiency?
3. Do the effects of IPD on examinee ability, pass-fail decisions, and overall exam efficiency vary by the size of the item pool?

I addressed all three research questions in terms of measurement precision, classification accuracy, and exam efficiency. Therefore in order to streamline the results, I present them in the order of the evaluation criteria I described in the Method section (pp. 53–55). I conclude this chapter with the results from a supplemental analysis I conducted in response to findings from the original analysis.

A. Precision

For the most part, the lowest values of RMSE, bias, and AAD and the highest correlation between estimated and true examinee ability occurred for the baseline conditions across the three sizes of item pools (see Table 4). As I introduced drift into the item pool, measurement precision increased as the number of items with drift and the magnitude of drift increased. The differences among the measures of precision are discussed in the following sections.

Table 4
Measures of Precision

Pool Size	IPD Items	Magnitude	RMSE	Bias	AAD	Corr.
1,000 items	**Baseline**	**n/a**	**0.001274**	**0.003259**	**0.029107**	**0.9988**
	100	1.0	0.00644	0.039656	0.071457	0.9985
		0.75	0.003432	0.018766	0.052338	0.9984
		0.5	0.002287	0.006362	0.040281	0.9985
	75	1.0	0.004307	0.03445	0.05796	0.9986
		0.75	0.002719	0.019802	0.045217	0.9985
		0.5	0.002085	0.011471	0.037784	0.9984
	50	1.0	0.003286	0.024269	0.050836	0.9986
		0.75	0.002162	0.011456	0.039365	0.9986
		0.5	0.001877	0.005815	0.035923	0.9985
500 items	**Baseline**	**n/a**	**0.001459**	**0.009281**	**0.030958**	**0.9988**
	100	1.0	0.007965	0.021594	0.071318	0.9981
		0.75	0.005832	0.004849	0.068362	0.9981
		0.5	0.003554	-0.01411	0.047497	0.998
	75	1.0	0.00625	0.009174	0.069565	0.9982
		0.75	0.003936	-0.00083	0.054201	0.9986
		0.5	0.00253	-0.01033	0.040159	0.9985
	50	1.0	0.002606	0.005761	0.042846	0.9987
		0.75	0.002162	-0.00405	0.037807	0.9987
		0.5	0.001947	-0.00723	0.036029	0.9986
300 items	**Baseline**	**n/a**	**0.001528**	**0.01504**	**0.031805**	**0.9989**
	100	1.0	0.015214	0.038958	0.110505	0.9976
		0.75	0.005801	0.001838	0.065025	0.9989
		0.5	0.002776	-0.01383	0.04174	0.9989
	75	1.0	0.009119	0.030063	0.086203	0.9981
		0.75	0.004578	0.014562	0.059692	0.9984
		0.5	0.002066	-0.00775	0.037355	0.9989
	50	1.0	0.005351	0.027075	0.065962	0.9979
		0.75	0.002932	0.008388	0.046993	0.9983
		0.5	0.001986	-0.00136	0.037677	0.9984

1. Correlation

The correlations between estimated and true ability were consistently high across all conditions (see Table 4). Correlations decreased when IPD was present, but differences were negligible. Correlations were lower with higher magnitudes of drift regardless of the number of IPD items. In most cases, correlations were slightly higher when there were fewer IPD items with drift in the bank (i.e., correlations were highest for all three 50 IPD item conditions compared to the 75 and 100 IPD item conditions). However, these differences were extremely small.

2. RMSE

The baseline condition had the smallest RMSE value across all conditions and all three item pools. Likewise, within each number of IPD items (100, 75, and 50), conditions with drift of 1.0 logits produced higher RMSE values, followed by 0.75 and 0.5 logits respectively. However, RMSE values were not higher for all three magnitudes of drift when there were more IPD items in the bank. For example, RMSE values were higher when there were 75 IPD items with 1.0 logits of drift compared to when there were 100 IPD items with 0.75 or 0.5 logits of drift. Similarly, when there were 50 items with IPD of 1.0 logits of drift, RMSE was higher than when there were 75 IPD items with 0.75 or 0.5 logits of drift. This pattern was consistent across all three item pools, as illustrated in Figures 1–3.

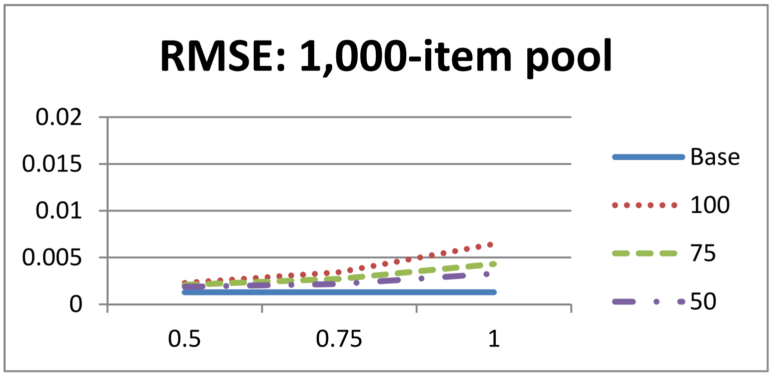

Figure 1. RMSE values for the large item pool with 1,000 items.

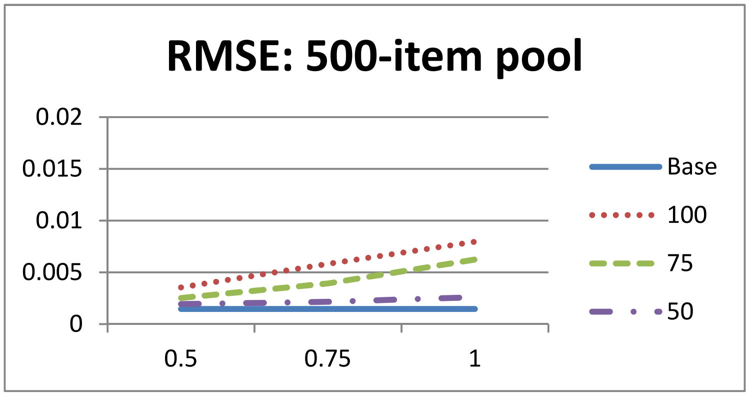

Figure 2. RMSE values for the medium item pool with 500 items.

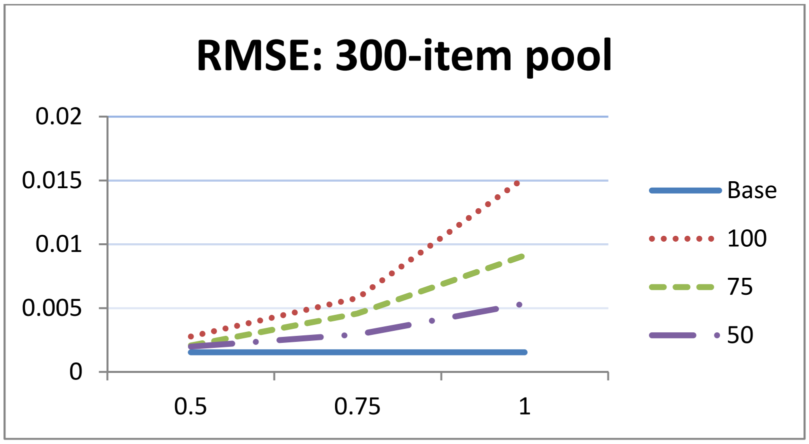

Figure 3. RMSE values for the small item pool with 300 items.

3. AAD

The pattern of AAD values mimics that of the RMSE values (see Figures 4–6). AAD values were the lowest for the baseline conditions, followed by conditions with 50 IPD items and a drift magnitude of 0.5 logits. Similarly, the conditions with 100 IPD items and drift magnitudes of 1.0 logits had the highest AAD values, and AAD values were similarly large in conditions with drift magnitudes of 1.0 logits compared to 0.75 or 0.5 logits, regardless of the number of IPD items in the bank. This pattern was consistent across all three item pools.

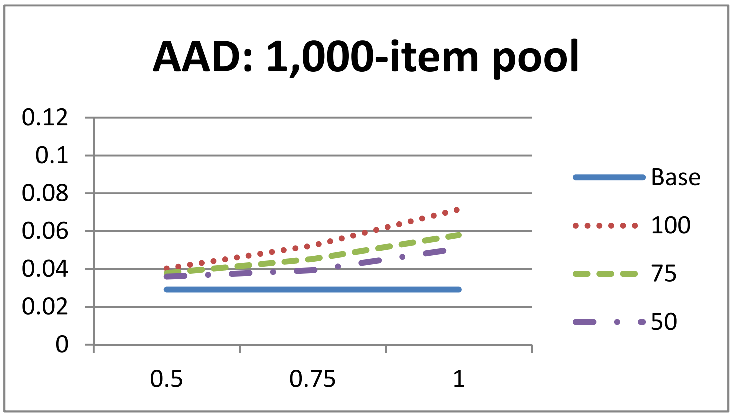

Figure 4. AAD values for the large item pool with 1,000 items.

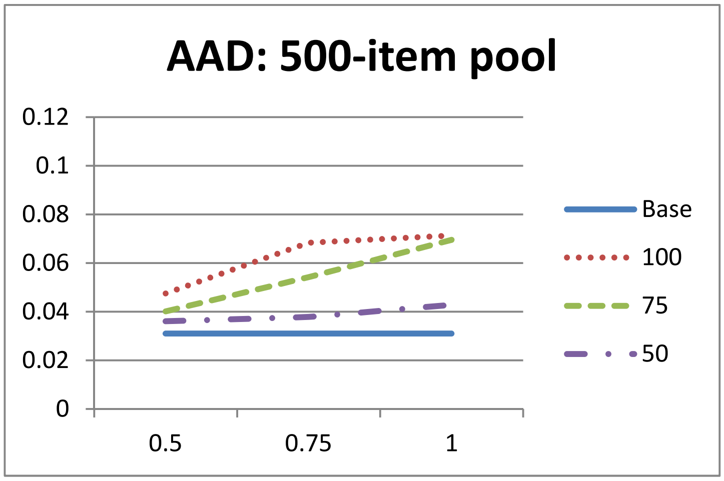

Figure 5. AAD values for the medium item pool with 500 items.

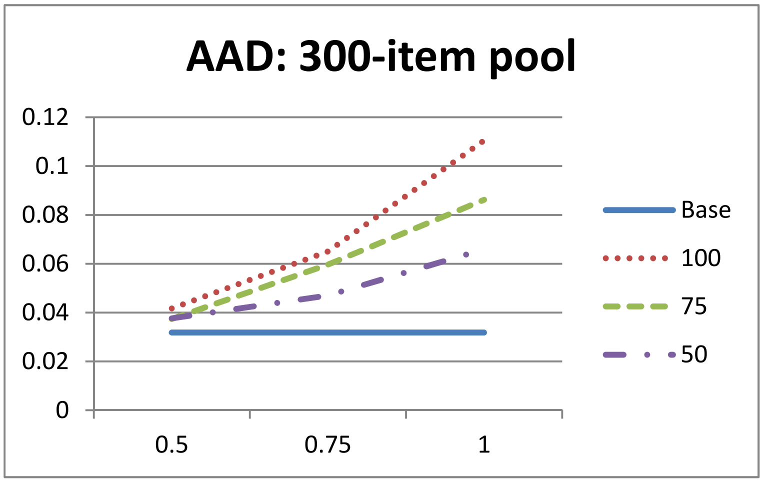

Figure 6. AAD values for the small item pool with 300 items.

4. <u>Bias</u>

Like the results for RMSE and AAD, bias was smallest for the baseline condition that had the item pool with 1,000 items (see Figure 7). By contrast, the condition of 100 items with IPD and a drift magnitude of 1.0 logits yielded the largest amount of bias. The least amount of bias occurred when there were 50 items with IPD and a drift magnitude of 0.5 logits. However, bias tended to be higher when the magnitude of drift was 1.0 compared to 0.75 and 0.5, regardless of the number of IPD items in the bank.

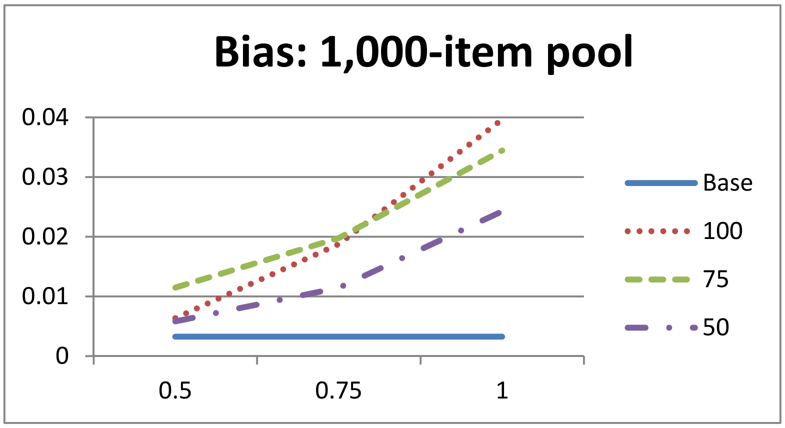

Figure 7. Bias values for the large item pool with 1,000 items.

However, this pattern did not occur for the item pools with 300 and 500 items. Across these pool sizes, six bias values for the IPD conditions were below those for the baseline condition (see Figures 8–9). Additionally, five of those bias values below the baseline for the 500-item pool and three of those bias values for 300-item pool were negative. All eight of the conditions with negative bias values occurred in the two smaller item pools, where there were a larger percentage of IPD items. Six of the eight conditions had drift magnitudes of 0.5 logits and the other two had drift magnitudes of 0.75 logits.

Logically then, we would expect that the condition with 50 IPD items and drift of only 0.5 logits would produce the least amount of bias. This was true for the pool of 300 items, but not for the pool of 500 items. In the 500-item pool condition, the condition with 50 IPD items with 0.5 drift magnitude had a lower bias value than the baseline condition, but it was the fifth smallest. The condition showing the least amount of bias in the 500-item pool had 75 IPD items with a drift magnitude of 0.75 logits. These unexpected differences in bias values for the 300- and 500-item pools and the 1,000-item pool might be an artifact of item pool size (i.e., not due to IPD). Despite these discrepant results, the differences in bias values between the baseline condition and the IPD conditions showing the least amount of bias were negligible; there was only a difference in bias of .008 in the item pool with 500 items and .014 in the item pool with 300 items. Bias values were large for both item pools when the magnitude of drift was 1.0 logits. This was true when there were 50, 75, and 100 IPD items in the pools.

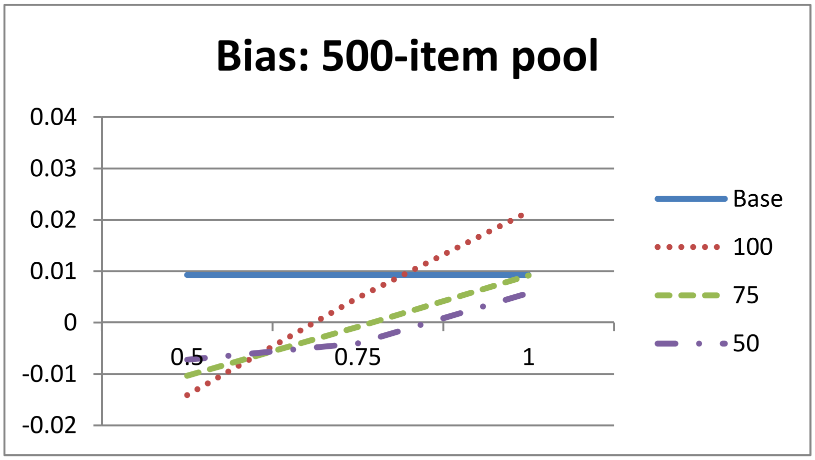

Figure 8. Bias values for the medium item pool with 500 items.

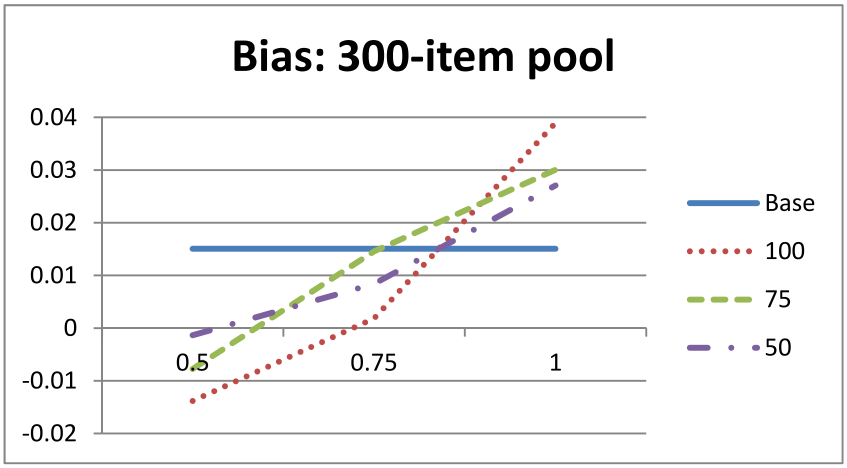

Figure 9. Bias values for the small item pool with 300 items.

B. <u>Classification Accuracy</u>

Compared to the measures of precision, the measures of classification accuracy showed an unclear pattern of results (see Table 5). The total misclassification percentages were relatively small for all conditions, and all were well within measurement error (i.e., no misclassifications were outside the 95% or even the 90% confidence interval). In this next section, I will discuss the differences among the measures of classification accuracy.

Table 5

Measures of Classification Accuracy

Pool Size	**IPD Items**	**Magnitude**	**FP**	**FN**	**Total Percentage of Misclassification**
1,000 items	**Baseline**	**n/a**	**3**	**4**	**1.4%**
	100	1.0	4	2	1.2%
		0.75	2	2	0.8%
		0.5	1	2	0.6%
	75	1.0	8	3	2.2%
		0.75	1	3	0.8%
		0.5	6	3	1.8%
	50	1.0	1	0	0.2%
		0.75	5	2	1.4%
		0.5	3	4	1.4%
500 items	**Baseline**	**n/a**	**2**	**3**	**1.0%**
	100	1.0	15	0	3.0%
		0.75	0	6	1.2%
		0.5	0	6	1.2%
	75	1.0	1	6	1.4%
		0.75	1	6	1.4%
		0.5	0	8	1.6%
	50	1.0	3	2	1.0%
		0.75	3	5	1.6%
		0.5	0	4	0.8%
300 items	**Baseline**	**n/a**	**3**	**2**	**1.0%**
	100	1.0	6	3	1.8%
		0.75	0	4	0.8%
		0.5	0	8	1.6%
	75	1.0	5	2	1.4%
		0.75	2	6	1.6%
		0.5	0	9	1.8%
	50	1.0	3	1	0.8%
		0.75	2	3	1.0%
		0.5	0	5	1.0%

1. Total misclassification

For all three item pools, findings were inconsistent across both the number of IPD items present in the item pool and the magnitude of drift. Additionally, there were IPD conditions for each item pool size that showed a smaller misclassification percentage than in the baseline condition. These conditions, however, were not consistent across the item pools. For example, in the item pool with 1000 items, the lowest percentage of misclassification occurred when there were 50 IPD items with a drift magnitude of 1.0 logits; but for the item pool with 500 items, the lowest percentage occurred when there were 50 IPD items with a drift magnitude of 0.5 logits (see Table 5). The largest percentage of misclassification across all conditions and item pool sizes was only 3%, a very encouraging finding. This occurred in the item pool with 500 items when there were 100 IPD items with a drift magnitude of 1.0 logits. The highest percentages of total misclassifications occurred in conditions when drift had a magnitude of 1.0 logits. Again, this was a consistent finding across the three item pool sizes. The reverse was not true, however. Conditions with drift magnitudes of only 0.5 logits did not consistently yield the lowest percentages of total misclassifications.

2. FP and FN

A pattern does appear when we compare the number of FP classifications to the number of FN classifications. On the whole, more FP classifications occurred when the magnitude of drift was 1.0 logits, whereas more FN classifications occurred when the magnitude of drift was only 0.5 logits. These findings were similar across the three sizes of item pools. A pattern did not appear when comparing the magnitude of drift was 0.75 logits. For the item pools with 300 and 500 items, more FN classifications occurred when the magnitude of drift was 0.75 logits. The item pool with 1000 items had nearly the same number of FP and FN classifications when the drift magnitude was 0.75 logits. Again, these differences might be more of an artifact of item pool size rather than conditions of IPD.

C. Test Efficiency

The measures of test efficiency showed the greatest inconsistency of all the evaluation criteria (see Table 6). For all three item pool sizes, the baseline conditions had the highest average test lengths, average exposure rates, and in most cases, the highest maximum test lengths and maximum exposure rates as well. This indicates that the exams administered with IPD were more efficient.

Table 6

Measures of Test Efficiency

			Test Length		**Exposure Rate**	
Pool Size	**IPD Items**	**Magnitude**	**Mean**	**Maximum**	**Mean**	**Maximum**
1,000 items	**Baseline**	**n/a**	**99.91**	**145.04**	**10.0%**	**71.2%**
	100	1.0	98.62	144.08	9.9%	71.3%
		0.75	98.64	144.71	9.9%	70.8%
		0.5	98.81	143.44	9.9%	71.1%
	75	1.0	99.19	145.25	9.9%	71.1%
		0.75	98.89	144.7	9.9%	70.8%
		0.5	98.72	143.56	9.9%	71.0%
	50	1.0	99.23	145.84	9.9%	70.8%
		0.75	99.34	144.88	9.9%	70.9%
		0.5	99.15	144	9.9%	70.7%
500 items	**Baseline**	**n/a**	**100.09**	**147.56**	**20.0%**	**83.4%**
	100	1.0	86.36	141.45	17.3%	83.8%
		0.75	97.85	144.61	19.6%	82.9%
		0.5	98.25	143.55	19.7%	85.5%
	75	1.0	97.52	142.75	19.5%	81.5%
		0.75	97.52	142.75	19.8%	83.7%
		0.5	99.38	145.79	19.9%	82.3%
	50	1.0	99.44	144.08	19.9%	83.2%
		0.75	99.76	145.96	20.0%	83.0%
		0.5	99.6	147.32	19.9%	82.7%
300 items	**Baseline**	**n/a**	**100.33**	**147.12**	**33.4%**	**88.7%**
	100	1.0	96.98	143.32	32.3%	87.7%
		0.75	98.79	143.44	32.9%	88.3%
		0.5	99.75	144.23	33.3%	87.6%
	75	1.0	97.29	141.65	32.4%	87.3%
		0.75	98.44	143.21	32.8%	88.4%
		0.5	100.12	146.85	33.4%	88.4%
	50	1.0	97.56	144.87	32.5%	88.8%
		0.75	98.44	144.92	32.8%	89.4%
		0.5	99.258	145.01	24.1%	88.9%

1. Test length

The results for test length demonstrate that test length increases as the magnitude of drift decreases. This finding is consistent across the conditions, another finding against expectations. When we compare maximum test lengths, we see that those values show this same pattern of increasing values with decreasing drift magnitude for the item pools with 300 and 500 items, but not for the item pool with 1000 items. For this item pool, the opposite occurred, with the lowest maximum values occurring when the drift magnitude was 0.5 logits, and the highest maximum values occurring when the drift magnitude was 1.0 logits. These findings were more in line with what we would expect to see. Again this could be an artifact of item pool size and not IPD.

2. Item exposure

Exposure rates exhibit a less discernible pattern than test lengths. The average exposure rates were highest for the baseline conditions, and only one or two IPD conditions had higher maximum exposure rates than the baseline conditions. Again, this finding was consistent across all three item pool sizes.

For the item pool with 1,000 items, there were no differences in mean exposure rates for the nine IPD conditions (see Table 6), and there were virtually no differences in mean exposure rates for these conditions in comparison to the baseline condition (0.1%). The highest maximum exposure rate value when there were 100 IPD items in the pool with a drift magnitude of 1.0 logits, and the lowest maximum exposure rate occurred when there were 50 IPD items in the pool with a drift magnitude of 0.5 logits. These findings were more in line with expectations. Additionally, mean exposure rates were highest when drift was 1.0 whether there were 50, 75, or 100 IPD items. However, mean exposure rates were not lowest in conditions with drift magnitude of 0.5 logits. Once again, these differences were all slight.

Differences in exposure rates were more apparent for conditions in the item pools with 300 and 500 items, but again there was no clear pattern (see Table 6). For each item pool size, at least one IPD condition had the same average exposure rate as the baseline condition, but the exposure rates for the other seven IPD conditions were all lower. The lowest exposure rates in the 300- and 500-item pools did not occur in the same conditions. Results were similar for the maximum exposure rates. The baseline conditions did not yield the highest values, and the majority of the IPD conditions had values that were the same or smaller than those in the baseline condition. Likewise, the conditions with the highest and lowest maximum values varied by item pool size and the number of IPD items present in the pool.

D. Supplemental Analysis

The surprising finding from overall test efficiency that the baseline condition was the least efficient of all the conditions across all three item pool sizes led to a supplemental analysis. I theorized that the lack of test efficiency could be due to the fact that items in my item pool were not targeted to my examinee sample. To test this hypothesis, I ran two additional series of simulations using the medium-sized item pool with 500 items. One series used an entirely new sample of examinees with a retargeted ability distribution that matched the mean and SD of the item pool's overall item difficulty. The other series of simulations used the same sample as the original analysis, but the simulations simply modified the probability of a correct response from 50% to 60%.[2]

1. Test efficiency

My hypothesis was that the mistargeting of examinees and items resulted in an inefficient exam. Therefore, I first compared the results from the baseline conditions from each series of simulations (i.e., original analysis, retargeted sample, 60% targeted probability). When strictly comparing the baseline conditions, the results from the supplemental analyses revealed that test efficiency increased for all three item pool sizes with the retargeted sample only (see Table 7). Mean test lengths for all three item pool sizes were lower for the retargeted sample, and the mean and maximum exposure rates also was substantially lower. These measures of test efficiency were higher, however, for the simulated conditions in which the target probability of a correct response was 60%.

[2] Since this was a supplemental analysis, these simulations only included 10 replications as opposed to the 100 replications I performed on the original work. This decision to use 10 replications was based on the recommendations of researchers who claimed that when running simulations on work with a non-empirical sample, 10 replications are sufficient (Harwell, Stone, Hsu & Kirisci, 1996).

Table 7
Test Efficiency Comparison for Baseline Conditions across the Three Sets of Analyses

		Test Length		**Exposure Rate**	
Analysis	**IPD Items**	**Mean**	**Maximum**	**Mean**	**Maximum**
Original	300	100.33	147.12	33.4%	88.7%
	500	100.09	147.56	20.0%	83.4%
	1,000	99.91	145.04	10.0%	71.2%
Retargeted Sample	300	91.09	150	30.4%	58.7%
	500	90.63	150	18.1%	46.2%
	1,000	90.17	150	9.0%	35.3%
60% Probability	300	101.52	150	33.84%	68.56%
	500	102.29	150	20.46%	49.80%
	1,000	101.13	150	10.11%	27.12%

I next evaluated all of the conditions from the two supplemental series of simulations. Surprisingly, when comparing the baseline conditions to the IPD conditions, the results from both simulation series (i.e., retargeted and 60% probability) were not significantly different from the results obtained in the original analysis (see Table 8). The longest average test length and the highest average exposure rates occurred in the baseline condition. In the next section of this chapter, I discuss differences in the results I obtained from these two series of simulations.

Table 8

Test Efficiency Comparison for the 500-Item Pool across the Three Sets of Analyses

			Test Length		**Exposure Rate**	
Analysis	**IPD Items**	**Magnitude**	**Mean**	**Maximum**	**Mean**	**Maximum**
Original	**Baseline**	**n/a**	**100.09**	**147.56**	**20.0%**	**83.4%**
	100	1.0	86.36	141.45	17.3%	83.8%
		0.75	97.85	144.61	19.6%	82.9%
		0.5	98.25	143.55	19.7%	85.5%
	75	1.0	97.52	142.75	19.5%	81.5%
		0.75	97.52	142.75	19.8%	83.7%
		0.5	99.38	145.79	19.9%	82.3%
	50	1.0	99.44	144.08	19.9%	83.2%
		0.75	99.76	145.96	20.0%	83.0%
		0.5	99.6	147.32	19.9%	82.7%
Retargeted Sample	**Baseline**	**n/a**	**90.63**	**150**	**18.1%**	**46.2%**
	100	1.0	88.24	150	17.65%	49.30%
		0.75	88.95	150	17.79%	50.00%
		0.5	88.89	150	17.78%	53.74%
	75	1.0	89.06	150	17.81%	44.10%
		0.75	89.26	150	17.85%	51.42%
		0.5	89.53	150	17.91%	48.76%
	50	1.0	89.78	150	17.96%	56.46%
		0.75	89.46	150	18.01%	45.06%
		0.5	90.25	150	18.05%	57.36%
60% Probability	**Baseline**	**n/a**	**102.29**	**150**	**20.5%**	**49.8%**
	100	1.0	99.58	150	19.92%	54.02%
		0.75	101.94	150	20.39%	50.64%
		0.5	103.51	150	20.70%	50.14%
	75	1.0	100.67	150	20.13%	47.18%
		0.75	102.06	150	20.41%	52.52%
		0.5	102.47	150	20.49%	51.36%
	50	1.0	100.43	150	20.09%	63.26%
		0.75	101.94	150	20.39%	50.64%
		0.5	103.64	150	20.67%	59.26%

a. 60% targeted probability

The measures of test efficiency from the 60% targeted probability analysis mirrored the measures obtained in the original analysis. The longest tests resulted when the drift magnitude was 0.5 logits, and the shortest tests resulted when the drift was 1.0 logits. These findings held regardless of how many IPD items were present in the bank.

The exposure rate results were also similar to those from the original analysis (i.e., inconsistency was apparent). There were six conditions that had mean exposure rates that were smaller than those in the baseline conditions, and only two conditions had larger mean exposure rates. Both of these were from conditions in which the magnitude of drift was only 0.5 logits. The maximum exposure rates across conditions were also inconsistent. There was only one condition with a rate lower than the rate in the baseline condition, but it had a 1.0 drift magnitude. The other two 1.0 conditions had higher maximum exposure rates, and the two highest maximum rates occurred in conditions with only 50 IPD items. But again, like the results from the original analysis, all of these differences were minimal.

b. Retargeted sample

In the analysis with the retargeted sample, the longest test lengths were in the baseline conditions, and all conditions hit the maximum number of 150 test items. This series of simulations also yielded tests that were longest when the drift magnitude was smallest at 0.5 logits. The mean exposure rate was also still highest for the baseline conditions, and there were two maximum exposure rates for IPD conditions that were lower than the rates for the baseline condition. Again, as with the original analysis, no real pattern emerged for the exposure rates.

2. Precision and classification accuracy

I also reexamined the measures of precision and classification accuracy. These results yielded a clearer pattern of results than those from the original analysis but lead to the same conclusions. I present the results in the next section of this chapter.

a. Precision

(1). Baseline comparison

For the most part, when comparing the measures of precision for the baseline conditions across the three sets of analyses, the most precise estimates were obtained in the original analysis (see Table 9). This was true across all three item pool sizes. The original analysis yielded lower RMSE and AAD values and higher correlations than the two supplemental analyses. However, this was not true for the bias values. Both supplemental analyses, yielded lower bias values for all three item pool sizes in comparison to the bias values from the original analysis. The 60% targeted probability analysis produced bias values that were lower than those from the analysis with the retargeted sample for the item pools with 500 and 1,000 items. For the 300-item pool, the retargeted sample analysis produced the lowest bias value. Once again, these differences were all quite small.

Table 9
Precision and Classification Accuracy Comparison between Baseline Conditions for the Three Sets of Analyses

		Precision				Classification Accuracy		
Analysis	**IPD Items**	**RMSE**	**Bias**	**AAD**	**Corr.**	**FP**	**FN**	**Total Misclassification**
Original	300	0.0015	0.015	0.03181	0.9989	3	2	1.0%
	500	0.0015	0.0093	0.03096	0.9988	2	3	1.0%
	1,000	0.0013	0.0033	0.02911	0.9988	3	4	1.4%
Retargeted Sample	300	0.0055	0.0035	0.05844	0.9978	1	3	0.8%
	500	0.0054	0.0065	0.05896	0.9978	3	6	1.8%
	1,000	0.0058	0.0004	0.06004	0.9977	2	7	1.8%
60% Probability	300	0.0066	0.0088	0.06473	0.994	7	12	3.80%
	500	0.0056	-0.0023	0.06049	0.9947	10	9	3.80%
	1,000	0.007	-0.005	0.0685	0.9937	10	8	3.60%

(2). 60% targeted probability

Similar to the test efficiency results, the precision results from the 60% targeted probability analysis mirrored the results from the original analysis, where items were targeted at a probability of a 50% correct response (see Table 10). The most precise estimate for all four measures across all the conditions occurred in the baseline condition. The condition with the lowest precision or highest values for RMSE, AAD and bias was the condition with 100 IPD items and a 1.0 drift magnitude. The condition with 50 IPD items and a 0.5 drift magnitude had the best measures of precision. Additionally, the measures of precision were lower as the magnitude of drift increased. This finding was consistent across all conditions for three amounts of IPD items.

(3). Retargeted sample

The precision results for the retargeted sample mirrored the results from the original analysis and from the 60% targeted correct response analysis (see Table 10). The most precise estimates were obtained in the baseline condition for all measures. The condition with only 50 IPD items and a drift magnitude of 0.5 logits had the highest precision (i.e., lowest values for the measures of precision). The condition with 100 IPD items and a drift magnitude of 1.0 logits had the lowest precision (i.e., highest values for the measures of precision). In this analysis, the measures of precision were lower as the magnitude of drift increased, with the highest measures of precision (i.e., lowest precision) resulting under conditions with a drift magnitude of 1.0 logits, regardless of the number of IPD items.

Table 10
Precision and Classification Accuracy Comparison for the 500 Item Pool across the Three Sets of Analyses

			Precision				**Classification Accuracy**				
Analysis	**IPD Items**	**Magnitude**	**RMSE**	**Bias**	**AAD**	**Corr.**	**FP**	**FN**	**Total Misclassification**	**Outside the 95% CI**	**Outside the 90% CI**
Original	**Baseline**	**n/a**	**0.0015**	**0.00928**	**0.031**	**0.9988**	**2**	**3**	**1.0%**	**0**	**0**
	100	1.0	0.008	0.02159	0.0713	0.9981	15	0	3.0%	0	0
		0.75	0.0058	0.00485	0.0684	0.9981	0	6	1.2%	0	0
		0.5	0.0036	-0.0141	0.0475	0.998	0	6	1.2%	0	0
	75	1.0	0.0063	0.00917	0.0696	0.9982	1	6	1.4%	0	0
		0.75	0.0039	-0.0008	0.0542	0.9986	1	6	1.4%	0	0
		0.5	0.0025	-0.0103	0.0402	0.9985	0	8	1.6%	0	0
	50	1.0	0.0026	0.00576	0.0428	0.9987	3	2	1.0%	0	0
		0.75	0.0022	-0.004	0.0378	0.9987	3	5	1.6%	0	0
		0.5	0.0019	-0.0072	0.036	0.9986	0	4	0.8%	0	0
Retargeted Sample	**Baseline**	**n/a**	**0.0054**	**0.00647**	**0.059**	**0.9978**	**3**	**6**	**1.8%**	**0**	**0**
	100	1.0	0.0351	-0.0927	0.1582	0.9973	0	3	0.60%	0	0
		0.75	0.0153	-0.058	0.1025	0.9979	1	8	1.80%	0	0
		0.5	0.011	-0.0564	0.0839	0.9974	0	9	1.80%	0	0
	75	1.0	0.0197	-0.0623	0.1167	0.9978	1	4	1.00%	0	1
		0.75	0.0122	-0.051	0.0903	0.9974	1	6	1.40%	0	0
		0.5	0.0085	-0.0347	0.0745	0.9975	0	8	1.60%	0	1
	50	1.0	0.0097	-0.0376	0.0804	0.9977	3	10	2.60%	0	1
		0.75	0.0068	-0.0301	0.0662	0.9981	3	6	1.80%	0	0
		0.5	0.0061	-0.0314	0.0623	0.998	1	10	2.20%	0	0
60% Probability	**Baseline**	**n/a**	**0.0056**	**-0.0023**	**0.0605**	**0.9947**	**10**	**9**	**3.80%**	**0**	**0**
	100	1.0	0.038	-0.124	0.1565	0.9926	1	43	8.80%	1	31
		0.75	0.0208	-0.0899	0.1192	0.9938	1	35	7.20%	0	21
		0.5	0.0126	-0.0627	0.0907	0.9937	1	25	5.20%	0	8
	75	1.0	0.0231	-0.0819	0.1231	0.9937	1	33	6.80%	0	14
		0.75	0.0159	-0.0768	0.1017	0.9939	0	28	5.60%	0	8
		0.5	0.0088	-0.0383	0.0752	0.9944	4	15	3.80%	0	1
	50	1.0	0.0129	-0.0599	0.0905	0.994	0	26	5.20%	0	8
		0.75	0.0095	-0.0448	0.078	0.9944	0	25	5.00%	0	7
		103.64	0.0069	-0.0403	0.0661	0.9954	2	11	2.60%	0	2

b. Classification accuracy

(1). Baseline comparison

When comparing indices of classification accuracy for the baseline conditions across the three sets of analyses, the slight improvements found in test efficiency and precision were not replicated (see Table 9 again). The findings from the original analysis showed the highest classification accuracy. The one exception was a slightly lower total percentage of misclassifications for the 300-item pool size with the retargeted sample. The 500- and 1000- item pools in the retargeted sample, and all three item pool sizes in the 60% targeted probability analysis had higher total misclassifications than the original analysis. The pattern of the number of FPs to FNs also varied in the two supplemental analyses as compared to the original. More FPs occurred for all three item pool sizes in the retargeted sample analysis, and an increase in FPs was only found in the 300- item pool for the 60% targeted probability analysis. Despite the increase found in the number of misclassifications in the two supplemental analyses, none of them occurred outside the 95% CI.

(2). 60% targeted probability

The 60% targeted probability analysis resulted in more misclassifications than the original analysis and revealed a different pattern of results (see Table 10 again). In this set of simulations, the total percentage of misclassification was similar to the measures of precision, where the highest percentage occurred when there was a drift magnitude of 1.0 logits, and the lowest occurred where there was a drift magnitude of 0.5 logits. Additionally, the largest misclassification percentage occurred in the 100 IPD items with 1.0 logit drift condition, and the smallest misclassification percentage occurred in the 50 IPD items with 0.5 logit drift condition. These findings were more in line with expectations when compared to the lack of a clear pattern in the original analysis.

The most notable difference was the larger amount of FN classifications. Unlike the original analysis, where more FP classifications occurred when the drift was 1.0 logits, more FN classifications than FP classifications occurred regardless of drift magnitude. Despite the higher number of misclassifications, there was only one outside the 95% CI. However, there were a number of misclassifications outside the 90% CI, ranging from 1 to 31 depending on the condition. Additionally, within each set of IPD items, both the highest number of misclassifications as well as the most misclassifications outside the 90% CI occurred when drift was 1.0 logits. This held true regardless of the number of IPD items, where misclassifications were higher with 50 IPD items and 1.0 logit drift than with 100 IPD items and 0.5 logit drift.

(3). Retargeted sample

The results for classification accuracy for the retargeted sample were more similar to the original analysis, in that they lacked a clear pattern (see Table 10 again). Both analyses had an equal number of misclassifications, and the baseline condition did not yield the lowest percentage of misclassifications. However, unlike the original analysis where only one condition had fewer misclassifications, four IPD conditions in the retargeted sample analysis had a lower percentage of total misclassifications. Furthermore, both the highest and lowest percentages came from conditions with drift magnitudes of 1.0 logits. The type of misclassification was more similar to the results of the 60% probability analysis, where there were consistently more FNs compared to FPs in each condition. Again, there were no misclassifications outside the 95% CI, but there were a few outside the 90% CI. These, however, were minimal compared to the 60% probability analysis and are not of great concern.

V. DISCUSSION

In this chapter, I summarize the key findings from my study and point out the practical implications of those findings. I first discuss the results in terms of the evaluation criteria and then address the research questions. I also discuss the practical implications of the findings for IPD and CAT in relation to each of the questions and make recommendations to testing organizations employing CAT based on the findings. In addition, I point out several strengths and limitations of this study and discuss the study's significance for the certification testing industry. Finally, I conclude the chapter with suggestions of other relevant research questions related to IPD that future researchers might find interesting to explore.

A. Evaluation Criteria

1. Precision

The simulations revealed promising results for measurement precision. The differences between the conditions in the measures of precision were minimal, and the impact of drift was as expected. As I introduced drift into the item pool, the values for most measures of precision, except correlations, increased both as the number of items with drift and the magnitude of drift increased. These were expected results, because when drift is present in the item pool, the precision of ability estimates diminishes. Looking at the results for the measures of RMSE, AAD and bias, the values were not higher for all three magnitudes of drift when there were more IPD items in the bank. The values of all three measures tended to be larger when items had 1.0 logits of drift, regardless of the number of IPD items in the bank. This finding suggests that the magnitude of drift actually has a greater impact on the precision of scores than the number of items with IPD in the item pool.

There were two unexpected results for measurement precision. First, some correlations were higher in conditions with drift. This indicates that when drift was present in the item pools there were smaller differences between the examinees estimated and true ability levels. However, these differences were extremely small and previous research by Witt et al. (2003) reported similar results.

The second finding against expectations was the negative bias values that occurred in the 500- and 300-item pools. The occurrence of negative bias values is slightly puzzling because of the distribution of the direction of IPD items. The direction of drift for the IPD items was an uneven split where more of the items (i.e., 75%) drifted easier. This means that we should see more inflated examinee scores as a result of the easier items. However, for these conditions with negative values, examinees estimated ability levels were less than their true ability levels. This might have something to do with which examinees (e.g, high vs. low ability) receive IPD items, how many IPD items they receive, or the location of the examinees in relation to the cut score. If more examinees received harder drifting items and answered them incorrectly then their estimated ability levels would likely be lower. Or if only examinees at the extreme ends of the distribution received the easier drifting items an inflation of estimated ability might not be as likely. Despite the negative bias values, the differences between true and estimated ability levels in all eight conditions were miniscule and practically insignificant. These findings coincide with results from the study by Gou and Wang (2003). They also found negative bias values in a few conditions, but they were all small and trivial in a practical sense.

Overall, the findings and conclusions for the measures of precision are similar to those of previous studies that looked at the effects of IPD using both Rasch and 2PL models in FIT and CAT. Larger amounts of drift resulted in decreased measurement precision, but these researchers found drift to have a minimal impact on overall examinee ability estimates (Jones & Smith, 2006; McCoy, 2009; Rupp & Zumbo, 2003; Stahl et al., 2002; Wells et al., 2002; Witt et al., 2003). A few studies have reported that IPD had a significant negative impact on measurement precision (Abad et al., 2010; Babcock & Albano, 2012; Skorupski, 2006). However, two of these were FIT studies (Babcock & Albano, 2012; Skorupski, 2006), and only the study by Babcock and Albano used the Rasch model. The CAT study by Abad et al. employed the use of the 3PL model.

2. Classification accuracy

The findings for classification accuracy were inconsistent for both the number of IPD items present in the item pool and the magnitude of the item drift. This was true across all three item pool sizes. The largest percentage of misclassifications across all conditions and item pool sizes was only 3%. Additionally, no misclassifications occurred outside the 95% or even the 90% confidence interval in any drift condition. This lack of pattern combined with the fact that all the misclassifications are within the Type I error rate (i.e., alpha .05), suggests that any fluctuations found in the FP and FN classifications are simply measurement error. This finding is different from the findings of previous research on IPD in FIT, which reported a handful of misclassifications outside the 95% confidence interval (Jones & Smith, 2006; Stahl et al., 2002; Witt et al., 2003). There were few misclassifications, and they were considered non-significant. However, Jones and Smith (2006) reported misclassifications beyond the 5% expected by chance. The finding of not one misclassification outside the 90% or 95% CI in this research might suggest that CAT is even more robust to drift than FIT.

Looking at the types of misclassification, more FP classifications occurred when the drift magnitude was 1.0 logits. Certification organizations typically consider FP classifications more detrimental than FN classifications, because a candidate receives a passing status when he or she is ill-equipped to perform the functions that the test is assessing. Thus, these results indicate that item drift of higher magnitudes might have a greater negative impact than item drift of lower magnitudes. This finding is not surprising, as we would expect estimates and resultant conclusions that are more inaccurate to arise when an item exhibits more drift. Another finding in support of this conclusion was that the highest percentage of total misclassifications among the conditions occurred when drift had a magnitude of 1.0 logits. Again, this was a consistent finding across all three item pool sizes.

The number of IPD items administered to each examinee or the type of examinee (low, moderate or high ability) who responded to IPD items might explain the inconsistency in the classification accuracy findings. If the exam only administered IPD items to examinees who were at the extreme ends of the ability distribution, then the IPD would not have affected those examinees with ability levels around the cut score, and we would not expect changes in the classification accuracy. How the IPD items were distributed would also likely influence the number of FP and FN classifications that occurred. Unfortunately I can only speculate, because the simulator program does not track this information.

Overall, the results and conclusions for classification accuracy are similar to the results and conclusions for the measures of precision. The differences in classification accuracy between the baseline and the IPD conditions were minor and did not affect the classifications of the examinees beyond that of normal measurement error (i.e., alpha .05). This suggests that there was virtually no impact on the examinee's classification accuracy, regardless of the magnitude of item drift or the number of IPD items present in an item pool. These results were also consistent across the different item pool sizes. A handful of studies of FIT (Babcock & Albano, 2012; Song & Arce-Ferrer, 2009) and CAT (Jurich et al., 2010; Guo et al., 2009) found an impact on pass-fail decisions, but only one of the FIT studies used the Rasch model (Babcock & Albano, 2012). Previous IPD research using the Rasch model in both FIT (Jones & Smith, 2006; Stahl et a.l, 2002; Witt et al., 2003) and CAT (Hagge et al., 2011; McCoy, 2009) supported the findings from my research that little classification accuracy error occurs in the presence of IPD.

3. Test efficiency

For all three item pool sizes, the baseline conditions had the highest average test lengths, average exposure rates, and in most cases, the highest maximum test lengths and maximum exposure rates as well. This finding indicates that the exams administered with IPD were more efficient. This finding is not as expected, because the presence of IPD impacts measurement precision and thus in theory should impact the efficiency of a test.

The pattern of results for test length demonstrates that test lengths increase as the magnitude of drift decreases. This finding was consistent across the three amounts of IPD items present in the pool, another finding against expectations, as one would assume that the larger the drift, the more inaccurate the ability estimate, thus resulting in a longer test. This could be due to the combination of the distribution of IPD items and the 95% CI stopping rule. Since the IPD items had an uneven split of 75% easier and 25% harder, more items became easier and thus the overall item bank became easier. The 95% CI stopping rule ends the exam when an examinee's ability estimate has a 95% chance of being above or below the cut score. Examinees correctly answering IPD items that became easier might be reaching the 95% CI threshold faster and ending their exams sooner. Therefore, combining this rule with an easier item bank might result in a shorter test for a large number of examinees, bringing down the average test length.

The values of maximum test lengths mimic the pattern of increasing values with decreasing drift magnitude for two of the three item pools. The opposite occurred for the 1,000-item pool, which is more in line with expectations. However, this result could be an artifact of the item pool size as opposed to a factor of IPD, as larger item pools tend to yield more efficient tests (Weiss, 1982).

While unexpected, the differences in the test efficiency measures for the baseline and the IPD conditions were all minimal and insignificant. Similar to the results for the measures of precision and classification accuracy, these findings suggest that we can expect little impact on test efficiency when IPD is present in the item pool.

B. Supplemental Analysis

The findings from the original analysis on overall test efficiency were somewhat surprising; the baseline condition was the least efficient of all the conditions across all three item pool sizes. One possible explanation for this could be the off-targeting of the examinee ability distribution to the item difficulties present in the item pool. I based the item difficulty and examinee ability distributions for this research on empirical distributions from an operational exam. This mismatch of ability to item difficulty is common in certification testing, because certification is a test of minimal competence, and the average examinee ability tends to be higher than the average item difficulty (Babcock & Albano, 2012). Thus, when the test specifications target items at a probability of a 50% correct response, a number of items from the item pool are not used. This will likely yield a less efficient exam than when there is a better match between examinee ability and item difficulty.

Therefore, I conducted a supplemental analysis to test this reasoning. Test efficiency did seem to improve when comparing the baseline condition results between the supplemental analyses and the baseline condition results from the original analysis. However, this was not the case when looking at the differences between the baseline conditions and the IPD conditions. The results from both sets of supplemental analyses did not prove to be much different from the results of the original analysis. Test efficiency measures from both the 60% targeted probability analysis and the retargeted sample analysis mirrored the measures from the original analysis, where the baseline conditions still had the longest average test lengths, and no discernible pattern emerged with the exposure rates. These findings support my earlier conclusion that the shorter test lengths might be due to a combination of the distribution of IPD items and the 95% CI stopping rule.

In terms of precision and classification accuracy, the original analysis produced more precise measures of precision than the two supplemental analyses. The original analysis produced lower values for all measures of precision and higher classification accuracy. However, the differences in the measures from the two supplemental analyses were small and not significant. Additionally, the pattern of results was similar to the pattern from the original analysis and actually had a more discernible pattern. This might suggest that the off-targeting slightly convolutes the interpretation of the effects of IPD, but ultimately these findings indicate that the targeting of examinees does not change the effect that IPD has on the precision of scores and classification of examinees in CAT.

Overall, these negligible differences between the results from the IPD conditions and from the baseline condition for all measures of precision, classification accuracy, and test efficiency support the same conclusions reached from the original analysis. The findings again suggest that the IPD present in the item bank does not significantly impact the precision of scores, classification accuracy of examinees, or test efficiency, and they provide more evidence as to the robustness of CAT to even large amounts of IPD.

C. Research Questions

1. Research Question 1: What amount of drift can be present in the item bank before examinee ability estimates, pass-fail decisions, and the overall efficiency of the exam become compromised?

Based on the results of this study, a large amount of drift or number of items with drift can be present in the item bank without affecting the ability estimates, pass-fail decisions, or the overall efficiency of the exam. Other research findings that show drift to have minimal impact on measurement precision (Hagge et al., 2011; Wells et al., 2002; Witt et al., 2003), classification accuracy (Hagge et al., 2011; Jones & Smith, 2006; Witt et al., 2003), and passing rates (Huang & Shyu, 2003) also support this conclusion. The caveat to this, however, is that the magnitude of drift is important to note. The findings from this study suggest that it is not the number of IPD items, but rather the magnitude of the IPD determines the degree of impact of drift on examinee ability estimates, pass-fail decisions, and the overall efficiency of the exam.

2. Research Question 2: What magnitude of drift has the greatest impact on examinee ability estimates, pass-fail decisions, and overall exam efficiency?

Higher magnitudes of drift in this study had more of a negative effect on measurement precision, classification accuracy and test efficiency, a finding which is similar to previous research findings (Hagge et al., 2011; Rupp & Zumbo, 2003; Song & Arce-Ferrer, 2009). The results from this study indicate that items with a drift magnitude of 1.0 logits have the greatest impact on examinee ability estimates, pass-fail decisions, and overall exam efficiency regardless of the number of items with drift in the bank. This finding suggests that monitoring the degree of drift in items is more important than determining the number of items that have drifted over time. Although the results showed minimal impact overall, testing organizations employing CAT should pay closer attention to the magnitude of drift in their items. Identifying and recalibrating or removing only items with a high degree of drift might be a more efficient strategy to extend the life of an item bank rather than identifying and recalibrating or removing all items that have drifted.

3. Research Question 3: Do the effects of IPD on examinee ability, pass-fail decisions, and overall exam efficiency vary by the size of the item pool?

Overall, better measurement precision, classification accuracy, and test efficiency were observed as the item pool size increased. However, even when the item pool consisted of only 300 items the impact on all evaluation criteria was minimal. This suggests that CAT is robust to IPD even when the item pool is small. No previous researchers that I am aware have investigated item pool size and its impact on IPD in terms of measurement precision, classification accuracy, or exam efficiency; therefore, this finding offers some new information. Testing organizations employing CAT can feel confident that IPD will not significantly impact examinee ability estimates, pass-fail decisions, and overall exam efficiency, even when using a smaller item pool.

D. Strengths and Limitations

1. Strengths

A major strength of my study is that I based my data and simulations on a real certification exam. I not only used the item parameters and examinee distribution, but I also employed the same exam structure and all of the CAT test properties. Other researchers conducting CAT studies have used empirical item parameters (Witt et al., 2003; Hagge et al., 2011) or simulated data (Skorupski, 2006; Song & Arce-Ferrer, 2009; Stahl et al., 2002). Additionally, my selection of IPD items was random and based on the exam specifications. Previous research on both FIT (Bock et al., 1988) and CAT (McCoy, 2009) has found drift to be systematic across various subtests and content areas. Therefore, by administering the IPD items according to the examinee's current ability estimate and interspersing them throughout all eight content areas, my simulations mimic how IPD items would present in a live testing situation.

Another strength of this work is my use of a set number of IPD items in the simulations rather than a percentage of items in the item bank. In the past researchers have tried to determine whether certain percentages of drift can compromise the item bank (e.g., 5%, 10% and 20%). Since my study actually evaluated specific numbers of IPD items (100, 75, and 50) across the three sizes of item pools, it incorporated eight different percentages of IPD items ranging from 5 to 33%. This is a far more extensive investigation of amount of IPD items than most previous research, which only tested two or three percentage levels.

Previous research supports all of the evaluation criteria and components of drift manipulated in my study. The literature establishes the legitimacy of each factor I used in my drift conditions, as well as the many criteria I used to evaluate the outcomes of my research. Therefore, I can have confidence in my study design and the results. Overall, the major advantage of my work is that it takes into account the practical constraints and conditions of testing organizations and evaluates IPD in more real world testing settings.

2. Limitations

Despite the strengths, this study has a number of limitations. There are several limitations related to the factors of drift that I chose to use. My study does not consider any magnitudes of drift higher than 1.0 logits. It is conceivable and likely in many cases that item drift could be of greater magnitude. I also chose not to vary the direction of drift in my study. Although I incorporated items that drifted both easier and harder, I focused on a single condition where 25% of the items drifted harder and 75% of the items drifted easier. Future work may incorporate different proportions of items that drift easy and hard. Additionally, although I noted the random draw of IPD items across all content areas as a strength, it can also be seen as a limitation. Two previous studies found that IPD affected one content area more than another (Chan et al., 1999; Sykes & Fitzpatrick, 1992). These were FIT studies that looked exclusively at detecting IPD and not the effects of IPD on test performance, but both studies concluded that the differences in item difficulty values were due to modifications in curriculum or instruction. Therefore, it is possible that the effects of IPD on test performance might be greater if IPD were more present in one particular content area than in others.

Another limitation of this study was the use of a mistargeted examinee distribution. Because I based the simulations on an empirical examinee ability distribution, the mean ability level of the examinees did not ideally target the mean item difficulty level in the pool. Although the supplemental analysis tried to compensate for this difference, I only ran the simulations over 10 replications as opposed to the full 100 replications used in the original work. This decreases the accuracy of the results and limits the ability to draw useful conclusions.

This study also makes the assumption that IPD impacts all examinees in the same way. However, this is not likely the case in a practical setting. I made this assumption based on the assumption of invariance, which assumes that examinees of a given ability have the same probability of answering an item correctly. Rupp and Zumbo (2006) characterize IPD as LOI. So, what does it mean if IPD is present? There isn't always an explanation for why an item drifts, yet the occurrence of IPD persists. What is IPD if you can't explain why it occurs? Items do not necessarily change but people do. Certain groups of examinees may "drift" and not the items. Most reasons for item drift are person centered (e.g., cheaters, changes in curriculum, differential instruction, etc.). Thus, the effect might be more person parameter drift (PPD), where certain people drift in their ability level. Future research might explore PPD and evaluate whether IPD impacts various outcomes for certain groups rather than the entire population.

Lastly, it is important to raise the issue of the generalizability of the findings. As with any study, the generalizations of these results are somewhat limited. Simulations by nature are advantageous, because one becomes aware of the finding before the finding happens empirically (e.g., if the finding were to indicate a negative impact on examinee scores, the issue could be corrected before it affected real examinees). However, the findings are limited to the exact studied conditions. Because I conducted my simulations based on the exam's specific set of criteria, the study does not account for other testing conditions. Therefore, the results are reflective of only the exact set of simulated criteria. In addition, this study's use of criterion-referenced data limits the use of the results to criterion-referenced testing only. One of the goals of this research was to evaluate not only the precision of scores in the face of IPD but also the recovery of the pass-fail classifications. Testing organizations and institutions using exams based on norm-referenced tests, such as those in education, are not able to apply the findings from this study.

E. Implications for Practitioners

The previous sections discussed several results concerning the impact of IPD. In this section I highlight the results most relevant to testing organizations and offer some guidance on how organizations might manage items with IPD present in their item banks. The results of this study indicated one important consideration for practice, the magnitude of IPD. For all the criteria examined, the magnitude of drift appeared to have the largest impact on measurement precision, where the values for RMSE, AAD and bias were higher when the drift magnitude was 1.0 logits. Thus in an applied setting, practitioners might choose to focus on the high drift items (i.e., items with a drift magnitude of ±1.0 or more logits).

With respect to the maintenance of item banks, practitioners should make every effort to identify items with drift, perhaps using one of the methods mentioned in the literature review. However, the choice of method should take into account the measurement model and the parameters of the exam. Once the IPD items are identified and removed from the operational pool, practitioners need to evaluate the items to decide what approach to take to correct them. Practitioners can choose to delete the item entirely from the item pool, they can recalibrate the item with the current sample of examinees; or they can rework the content of the item stem or answer choices and recalibrate the item as a pretest item. Simply recalibrating the item with the current sample of examinees could present an issue with restricted range of ability estimates. In practice, it is likely that the current examinee sample is not as diverse as the sample used to calibrate the item originally. Therefore, the new item calibration might not be accurate. An alternative option to help mitigate this effect would be to reseed the item into your experimental pool to get a more heterogeneous sample of examinees.

Numerous researchers have concluded that IPD has little impact on examinee ability estimates, but there is still a lot unknown about the effects of IPD, especially in CAT. Although the results of this study suggest that even large amounts of IPD are not a threat in CAT, by no means should testing organizations use these findings as an indication not to monitor IPD. The presence of IPD introduces trait-irrelevant differences; therefore it continues to present a threat to measurement precision and accurate classification of examinees. The results of my study pertain specifically to the parameters that I have tested and practitioners cannot generalize them to all testing situations. Testing organizations that fail to identify and handle items with IPD risk disadvantaging their examinees and jeopardizing the validity of their exams.

F. Suggestions for Future Research

Although this study provided answers to several questions regarding the effects of IPD in CAT, there are still many areas and IPD conditions to address in future studies to fully understand how IPD can impact measurement and to reach more generalizable conclusions. First, my study focused on a narrow set of IPD criteria to manipulate. Future work might include various other amounts, magnitudes, or directions of drift. Another area for future research is to examine the effects of other sample sizes. I chose to use a consistent sample of 500 examinees across all the conditions. I based this decision on the rationale that 500 is a typical number of examinees for certification testing; however, other researchers might evaluate whether the impact of drift changes with varying numbers of examinees. Similarly, I looked at the impact of IPD across three sizes of item pools (300, 500, and 1,000). Other researchers might study different item pool sizes. In addition, I modeled a specific CAT certification exam in my research and employed the content restrictions and test properties of that exam. However, it is also important to understand how various content restrictions and CAT criteria, such as differences in cut scores, item selection, and stopping rules might change or impact the effect of IPD.

One possible direction for future study is to examine drift in the initial administered items of an exam. How does IPD impact estimation precision when there is drift present in the first few items administered? CAT exams rely heavily on the initial ability estimates of examinees in order to gain information concerning which subsequent items to administer. If the exam inaccurately estimates ability at the beginning of the exam, it is plausible that this could impact the final ability estimates and test efficiency. Thus, researchers might investigate the effect of IPD through a series of conditions with varying amounts, magnitudes, and directions of drift in the items administered up front.

Another possible avenue for research would involve looking at the actual number of administered drift items versus a probabilistic draw of items. My research randomly selected a set of items to exhibit drift within the bank, and then administered items based on examinee ability level and the selection criteria of the exam. Therefore, the number of drift items varied for each examinee, where some might have encountered many items with drift and others might have encountered none. This method of probabilistic draw mimics how an operational bank would distribute drift items to examinees. However, future research could examine the impact of IPD if every examinee got a set number or percentage of drift items throughout the test—such as exposing each examinee to 5, 10 or 15 drift items despite their estimated ability levels. This is not as realistic, but such research could provide more interpretable results and a definitive pattern as to the impact of IPD on the precision of scores and accurate classification of examinees.

Finally, researchers might be interested in looking at the effect of IPD at varying levels of item difficulty and examinee ability. Would changing the targeting of examinee ability to item difficulty alter how IPD impacts measurement precision and classification accuracy? For example, the results of this study would likely have been impacted had the cut score of the exam been moved well above the mean of the examinee distribution. A larger number of misclassifications might have occurred in this context. This type of research aligns with previous research on FIT that found differences in the effect of IPD on both measurement precision and classification accuracy when looking at different levels of item difficulty and examinee ability (Skorupski, 2006; Song & Arce-Ferrer, 2009). To my knowledge, this line of work has not been investigated in CAT. As such, future research in this area is needed to understand how IPD interacts with varying ability levels and item difficulties for CAT.

G. Conclusion

The findings from this study have several practical implications. Specifically, this work helps to show the robustness of CAT in terms of measurement precision, classification accuracy, and test efficiency in the face of IPD. The findings from this study may help testing organizations make decisions regarding how they evaluate IPD in their item banks. The most relevant finding is that the magnitude of drift is a more important consideration than the number of drift items when assessing IPD in an item bank. Results from the numerous conditions consistently indicated that regardless of the size of the item pool or the number of IPD items in the item bank, drift of 1.0 logits had the most negative impact on measurement precision. Therefore, I recommend that testing organizations focus their resources on identifying and correcting items with high magnitudes of drift (i.e., remove or recalibrate the items), as these items are likely to have the most impact on the precision of scores and accuracy of pass-fail decisions. Additionally, this research suggests that the size of the item pool is not a factor affecting the impact of IPD in CAT. Testing organizations using small item pools can therefore be less concerned about item pool size contributing to the impact of IPD on their examination results. Finally, this study helped to generate many interesting research questions for researchers to explore in the future. Although the results from this study and other CAT-related research has shown minimal impact of IPD, there is a need for additional research as there are still many unanswered questions. In the context of the parameters of this study we can conclude thus far that CAT is fairly robust to IPD. However, in other testing scenarios where the ability of the examinees, difficulty of the items, or the cutscore of the exam are varied this might not be the case. If we modify these variables, does IPD continue to show a minimal impact on measurement precision, classification accuracy, and test efficiency? The jury is still out on these types of situations.

REFERENCES

Abad, F. J., Olea, J., Aguado, D., Ponsoda, V., & Barrada, J. R. (2010). Deterioro de parámetros de los ítems en tests adaptativos informatizados: estudio con eCAT [Item parameter drift in computerized adaptive testing: Study with eCAT]. *Psicothema, 22,* 340-7.

Babcock, B., & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement*, *36*(7), 565-580.

Baker, F. B. (2001). *The basics of item response theory*. College Park, MD: ERIC clearinghouse on assessment and evaluation. Retrieved from http://ericae.net/irt/baker

Becker, K. (2013). CAT Simulator.

Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow & J. B. Olsen (Eds.), *Innovations in computerized assessment* (pp. 67–91). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

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

Bock, D. B., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*(4), 275-285.

Chan, K. Y., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery. *Journal of Applied Psychology, 84*(4), 610-619.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement, 25*(1), 31-45.

Davey, T., & Pitoniak, M. J. (2006). Designing computer-adaptive tests. In S. M. Downing & T. M.

Haladyna (Eds.), *Handbook of test development* (pp. 543–573). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

DeMars, C. E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education, 17*(3), 265-300.

Deng, H., & Melican, G. (2009, April). *An investigation of scale drift in computer adaptive test*. Paper presented at the Annual Meeting of National Council on Measurement in Education, San Diego, CA.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement, 22*(1), 33-51.

Doris, J., & Sarason, S. (1955). Test anxiety and blame assignment in a failure situation. *The Journal of Abnormal and Social Psychology*, *50*(3), 335.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–515). Westport, CT: Praeger Publishers.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Flaugher, R. (2000). Item Pools. In Wainer, H. (Ed.) *Computerized adaptive testing: A Primer*. Mahwah, NJ: Erlbaum.

Gershon, R. C. (1992). Test anxiety and item order: New concerns for item response theory. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 1, pp. 175–194). Norwood, NJ: Ablex.

Gershon, R. C. (1996). The effect of individual differences variables on the assessment of ability for computerized adaptive testing (doctoral dissertation). Dissertation Abstracts International: Section B: *The Sciences & Engineering*, *57*, 4085.

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement*, *6*(1), 109–127. Retrieved from http://proxy.cc.uic.edu/docview/620672971?accountid=14552

Glas, C. A. W. (2000). Item calibration and parameter drift. In W. J. van der linden & C. A. W. Glas (Eds.). *Computerized Adaptive Testing*: *Theory and practice* (pp.183-199). Norwell MA: Kluwer Academic.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20*(4), 369-377.

Guernsey, L. (2000, August 6). An ever-changing course: Taking admissions tests on computer. *New York Times*: Education Life, pp. 32–33.

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.

Guo, F., & Wang, L. (2003, April). *Online calibration and scale stability of a CAT program*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Guyer, R., & Thompson, N.A. (2011). *Item response theory parameter recovery using Xcalibre 4.1.* St. Paul MN: Assessment Systems Corporation.

Hagge, S., Woo, A., & Dickison, P. (2011, October). *Impact of Item Drift on Candidate Ability Estimation*. Paper presented at the annual conference of the International Association for Computerized Adaptive Testing, Pacific Grove, CA.

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

Han, K. T., & Guo, F. (2011). *Potential impact of item parameter drift due to practice and curriculum change on item calibration in computerized adaptive testing* (R-11-02). Graduate Management Admission Council Research Report.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied psychological measurement*, *20*(2), 101-125.

Hatfield, J.P., & Nhouyvanisvong, A. (2005, April). *Parameter drift in a high-stakes computer adaptive licensure examination: An analysis of anchor items.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Huang, C., & Shyu, C. (2003, April). *The impact of item parameter drift on equating*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Jones, P. E., & Smith, R. W. (2006, April) *Item Parameter Drift in Certification Exams and Its Impact on Pass-Fail Decision Making.* Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA.

Jurich, D. P., DeMars, C. E., & Goodman, J. T. (2012). Investigating the impact of compromised anchor items on IRT equating under the nonequivalent anchor test design. *Applied Psychological Measurement*, *36*(4), 291-308.

Jurich, D. P., Goodman, J. T., & Becker, K. A. (2010, May). *Assessment of various equating methods: Impact on the pass-fail status of cheaters and non-cheaters.* Poster presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Kim, D., Barton, K., & Choi, S. (2010, May). *Sample size impact on screening methods in the Rasch model.* Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Kingsbury, G. G., & Wise, S. L. (2011). Creating a K-12 Adaptive Test: Examining the Stability of Item Parameter Estimates and Measurement Scales. *Journal of Applied Testing Technology, 12,* ___. Retrieved from http://www.testpublishers.org/journal-of applied-testing-technology

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*, 147–154.

Klein, S. P. (1981). *The effect of time limits, item sequence and question format on the California bar examination.* A report prepared for the Committee of Bar Examiners of the State of California and the National Conference of Bar Examiners.

Kolen, M., & Brennan, R. (1995). *Test equating methods and practices*. New York: Springer-Verlag.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research*, *55*(3), 387–413.

Lee, H., & Dodd, B. G. (2012). Comparison of exposure controls, item pool characteristics, and population distributions for CAT using the partial credit model. *Educational and Psychological Measurement*, *72*(1), 159-175.

Li, X. (2008). An investigation of the item parameter drift in the examination for the certificate of proficiency in English (ECPE). *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 6*, 1–28.

Linacre, J. M. (2013). Winsteps® Rasch measurement computer program. Beaverton, Oregon: Winsteps.com

Linacre, J. M. (2000). Computer-adaptive testing: A methodology whose time has come. In S. Chae, U. Kang, E. Jeon, & J. M. Linacre (Eds.), *Development of Computerized Middle School Achievement Tests*. MESA Research Memorandum, (69).

Livingston, S. A. (2004). *Equating test scores (without IRT).* Princeton, NJ: ETS. Available online at http://www.ets.org/Media/Research/pdf/LIVINGSTON.pdf

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

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

Luecht, R. M., & Sireci, S. (2012). *A review of models for computer-based testing* (Research Report No. 2011-12). Retrieved from College Board website: http://research.collegeboard.org/publications/content/2012/05/review-models-computer-based-testing

Lunz, M. E., Stahl, J. A., & Bergstrom, B. A. (1993, April). *Targeting, test length, test precision, and decision accuracy for Computerized Adaptive Tests*. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based value-added accountability. *Journal of Educational and Behavioral Statistics, 31*, 35-62.

Martineau, J. A. (2004). *The Effects of Construct Shift on Growth and Accountability Models*. (Unpublished doctoral dissertation). Michigan State University, East Lansing, MI.

Masters, J., Muckle, T., & Bontempo, B. (2009, April). *Comparing methods to recalibrate drifting items in computerized adaptive testing*. Paper presented at the annual conference of the American Educational Research Association, San Diego, CA.

McCoy, K. M. (2009). *The impact of item parameter drift on examinee ability measures in a computer adaptive environment.* (Unpublished doctoral dissertation). University of Illinois at Chicago, Chicago, IL.

Meng, H., Steinkamp, S., & Matthews-Lopez, J. (2010). *An investigation of item parameter drift in computer adaptive testing.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Meyers, J., Miller, G. E., & Way, W. D. (2006, April). *Item position and item difficulty change in an IRT-based common item equating design*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Moyer, E. L., Galindo, J. L., & Dodd, B. G. (2012). Balancing flexible constraints and measurement precision in computerized adaptive testing. *Educational and Psychological Measurement.* doi: 10.1177/0013164411431838

Munz, D. C., & Smouse, A. D. (1968). Interaction effects of item-difficulty sequence and achievement-anxiety reaction on academic performance. *Journal of Educational Psychology*, *59*(5), 370.

Neely, D. L., Springston, F. J., & McCann, S. J. (1994). Does item order affect performance on multiple-choice exams? *Teaching of Psychology*, *21*(1), 44–45.

Olea, J., Revuelta, J., Ximenez, M. C., & Abad, F. J. (2000). Psychometric and psychological effects of review on computerized fixed and adaptive tests. *Psicología*, *21*, 157–173.

O'Neill, T. (2013). How much item drift is too much? *Rasch Measurement Transactions, 27*(3),1423-1424.

Plake, B. S. (1980). Item arrangement and knowledge of arrangement on test scores. *Journal of Experimental Education*, *49*, 56–58.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danish Institute for Educational Research.

Reckase, M. D. (2011). Computerized adaptive assessment (CAA): The way forward. In *The road ahead for state assessments, policy analysis for California education and Rennie Center for Education Research & Policy.* (pp.1-11). Cambridge, MA: Rennie Center for Education Research & Policy.

Rupp, A. A., & Zumbo, B. D. (2003). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. *The Alberta Journal of Educational Research, 49*(3), 264-276.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and psychological measurement*, *66*(1), 63-84.

Segall, D. O., & Moreno, K. E. (1999). Development of the computerized adaptive testing version of the Armed Services Vocational Aptitude Battery. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Skorupski, W. P. (2006, April). *The effects of item parameter drift on equating test scores.* Paper presented at the annual meeting of the National Council on Measurement, San Francisco, CA.

Song, T., & Arce-Ferrer, A. (2009, April). *Comparing IPD detection approaches in common-item nonequivalent group equating design.* Paper presented at the annual conference of the National Council on Measurement, San Diego, CA.

Stahl, J., Bergstrom, B., & Shneyderman, O. (2002, April). *Impact of item drift on test-taker measurement.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Stahl, J. A., & Muckle, T. (2007, April). *Investigating displacement in the Winsteps Rasch calibration application.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Stone, C. A., & Lane, S. (1991). Use of restricted item response theory models for examining the stability of item parameter estimates over time. *Applied Measurement in Education, 4(*2),125-141.

Sykes, R. C., & Fitzpatrick, A. R. (1992). The stability of IRT b values. *Journal of Educational Measurement, 29*(3), 201-211.

Sykes, R. C., & Ito, K. (1993, April). *Item parameter drift in IRT-based licensure examinations.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, *12*(1), 15–20. doi:http://dx.doi.org/10.1111/j.1745-3992.1993.tb00519.x

Wainer, H., Dorans, N. J., Green, B. F., Steinberg, L., Flaugher, R., Mislevy, R. J. ... Thissen, D. (2010). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum AssociatesPublishers. Retrieved from http://proxy.cc.uic.edu/docview/617855820?accountid=14552

Wang, W. C., & Chen, C. T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the WINSTEPS program for the family of Rasch models. *Educational and Psychological Measurement*, *65*(3), 376-404.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*(4), 473–492.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–-375.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26*(1), 77-87.

Whitely, S. E., & Dawis, R. V. (1976). The influence of test context on item difficulty. *Educational and Psychological Measurement*, *36*(2), 329–337.

Wise, S. L., & Plake, B. S. (1989). Research on the effects of administering tests via computers. *Educational measurement: Issues and practice*, *8*(3), 5–10.

Wise, S. L. Roos, L. L., Plake, B. S., & Nebelsick-Gullett, L. J. (1994). The relationship between examinee anxiety and preference for self-adapted testing. *Applied Measurement in Education*, *7*(1), 81–91.

Witt, E. A., Stahl, J. A., Bergstrom, B. A., & Muckle, T. (2003, April). *Impact of item drift with non-normal distributions*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Wolfe, E.W., & Smith Jr., E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part I—Instrument development tools, 202-242. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Rasch Measurement: Advanced and Specialized Applications*. Maple Grove, MN: JAM Press.

Wollack, J. A., Sung, H. J., & Kang, T. (2005) *Longitudinal effects of item parameter drift.* Paper presented at Annual Meeting of the National Council on Measurement in Education, Montreal, CA.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological measurement, 29*(1), 23-48.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement*, *17*(4), 297–311.

Yi, Q., Zhang, J., & Chang, H. (2008). Severity of organized item theft in computerized adaptive testing: A simulation study. *Applied Psychological Measurement, 32*(7), 543–558.

Zwick, R. (2006). Higher education admissions testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 647–679). Westport, CT: Praeger Publishers.

VITA

Nicole Makas Risk
University of Illinois at Chicago
Email: ncolwe2@uic.edu

EDUCATION

2015	**University of Illinois at Chicago** Ph.D. in Educational Psychology Emphasis: Measurement, Evaluation, Statistics, and Assessment (MESA) Dissertation: "The Impact of Item Parameter Drift in Computer Adaptive Testing (CAT**)**" Advisor: Everett Smith, Ph.D. GPA: 4.0
2009	**New York University** Master of Arts in Educational Psychology Emphasis: Psychological Measurement and Evaluation Thesis: "Examining the fidelity of *INSIGHTS* into Children's Temperament intervention program" Advisor: Sandee McClowry, Ph.D. GPA: 3.93
2007	**California State University Fullerton** Bachelor of Arts in Psychology, *Dean's list*

STUDY ABROAD

Summer 2009	**New York University, Pretoria, South Africa** Educational and Social Reform Research Project: "The Social and Educational Injustices of the Matric Exam"

CERTIFICATION TESTING EXPERIENCE

2013 – Present	**Psychometrician** American Medical Technologists, Chicago, IL
2012 – 2013	**Computer Adaptive Test Consultant** American Medical Technologists, Chicago, IL Oversee and manage the conversion of two fixed item tests to computer adaptive versions. Prepare and analyze data for simulations. Interpret simulation results and provide feedback and reports to Board of Directors and subject-matter expert committees regarding item performance and test development.
06/2012 – 08/2012	**Summer Psychometric Intern** National Council for State Boards of Nursing, Chicago, IL

Completed a research project using data from the NCLEX exam that investigated the ability to decompose item difficulty from characteristics such as: item type, content area and cognitive level using the linear logistic test model (LLTM). Assisted in day-to-day testing operations including standard setting and item writing workshops. Prepared standard setting Technical Report.

01/2012 – 06/2012 **Measurement Intern**
American Medical Technologists, Chicago, IL

Conducted psychometric analyses related to test construction: validity, reliability, equating, and scoring. Analyzed examinee and test item performance data from computer-based test administrations using classical and Rasch methodologies. Interpreted results and prepared technical reports measuring item performance for subject-matter expert committees. Performed data manipulations on large-scale data files of test results.

RESEARCH EXPERIENCE

2014 – Present **Computer Adaptive Test Coursework Development**
University of Illinois at Chicago, Chicago, IL

2010 – 2013 **Research Assistant**
Institute for Government and Public Affairs,
University of Illinois at Chicago, Chicago, IL
Child Care and Preschool Quality: Domain-Specific Measures and their Policy Implications (PI: Rachel A. Gordon, Ph.D.)

Investigate psychometric properties of various child care measures used in evaluating child care quality. Manage expert content review of quality measures. Collect and analyze large-scale data sets for the project.

2010 – 2013 **Research Assistant**
University of Illinois at Chicago, Chicago, IL
Measurement, Evaluation, Statistics, and Assessment Laboratory (MESA Lab)

Perform statistical analyses of data for faculty on non-grant funded research. Review and assist with the writing of the Method sections of papers and grant applications. Provide advice on developing surveys, rating scales, and tests. Assist with setting up appropriate data files for ease in data entry. Assist with interpreting statistical output. Identify appropriate statistical tests for answering various research questions. Assist with developing research designs for both faculty and student research projects.

2008 - 2010 **Research Assistant**
New York University, New York City, NY
INSIGHTS into Children's Temperament
(PI: Sandee McClowry, Ph. D.)

Conducted preliminary research and literature reviews for various areas of interest related to the project. Performed statistical analyses of classroom observation data related to the Teacher-School-Age Temperament Inventory. Employed qualitative and quantitative research methods to examine the fidelity of the implementation of the intervention. Conducted literature reviews for P.I.'s new book *Temperament Based Classroom Management.*

TEACHING EXPERIENCE

Spring 2015 **Adjunct Faculty**
University of Illinois at Chicago, Department of Educational Psychology
Essentials of Quantitative Inquiry in Education

Fall 2014 **Teaching Assistant**
University of Illinois at Chicago, Department of Educational Psychology
Essentials of Quantitative Inquiry in Education

Spring 2014 **Teaching Assistant**
University of Illinois at Chicago, Department of Educational Psychology
Essentials of Quantitative Inquiry in Education

Fall 2013 **Teaching Assistant**
University of Illinois at Chicago, Department of Educational Psychology
Essentials of Quantitative Inquiry in Education

Summer 2011 **Teaching Assistant**
University of Illinois at Chicago, Department of Educational Psychology
Essentials of Quantitative Inquiry in Education

PUBLICATIONS

Peer-Reviewed Journals

Colwell, N., Gordon, R. A., Fujimoto, K., Kaestner, R., & Korenman, S. (2013). New evidence on the validity of the Arnett Caregiver Interaction Scale: Results for preschoolers in the Early Childhood Longitudinal Study-Birth Cohort. *Early Childhood Research Quarterly, 28,* 218-233.

Colwell, N. (2013). Test anxiety, computer adaptive testing, and the common core. *Journal of Education and Training Studies.*

Book Chapters

Collins, A., Colwell, N., & McClowry, S. G. (2011). Maintaining fidelity of the intervention. In B. M. Melnyk & D. Morrison-Beedy (Eds.), *Designing, conducting, analyzing, and funding intervention research: A practical guide for success.* New York, NY: Springer.

MANUSCRIPTS UNDER REVIEW (DRAFTS AVAILABLE)

For Peer-Reviewed Journals

Gordon, R. A., Hofer, K. G., Fujimoto, K., Colwell, N., Kaestner, R., & Korenman, S. (2012). *New Evidence About the Validity of the ECERS-R for Evaluations of Preschool Programs Aimed at Improving School Readiness*. Manuscript submitted to Early Education and Development.

MANUSCRIPTS IN PREPARATION (DRAFTS AVAILABLE)

For Peer-Reviewed Journals

Colwell, N. (2013). *Item parameter drift in computer adaptive testing and its effects on classification accuracy.* Manuscript in Preparation.

Gordon, R., Colwell, N., Fujimoto, K., Kaestner, R., & Korenman, S. (2013) *Domain-Specific Quality Measures for Early Childhood Programs: New Evidence from the Study of Early Child Care and Youth Development*. Manuscript in preparation.

Gordon, R. A., Fujimoto, K., Colwell, N., Abner, K. S., Kaestner, R., Wakschlag, L. S., & Korenman, S. (2012). *Measuring socio-emotional development in a large-scale survey*. Manuscript in preparation.

PRESENTATIONS

Aksu, B. & Colwell, N. (2014, April). *Monitoring Rater Facet in a Highland Dance Championship*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Colwell, N. (2013, April). *Item Parameter Drift in Computer Adaptive Testing and its Effects on Classification Accuracy*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Gordon, R., Colwell, N., Fujimoto, K., Kaestner, R., & Korenman, S. (2013, April) *Domain-Specific Quality Measures for Early Childhood Programs: New Evidence from the Study of Early Child Care and Youth Development*. Presented at the Bi-Annual Meeting of the Society for Research in Child Development, Seattle, WA.

Colwell, N., Gordon, R. A., Fujimoto, K., Kaestner, R., & Korenman, S. (2012, November). *New Evidence on the Validity of the Arnett Caregiver Interaction Scale: Results from the Early Childhood Longitudinal Study-Birth Cohort*. Paper presented at the annual meeting of the National Council on Family Relations, Phoenix, AZ.

Colwell, N., Gordon, R. A., & Fujimoto, K. (2012, April). *New evidence on the validity of the Classroom Assessment Scoring System*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, BC.

PROFESSIONAL AFFILIATIONS

American Educational Research Association
National Council on Measurement in Education
Institute for Credentialing Excellence

STATISTICAL SOFTWARE EXPERIENCE

Stata, Winsteps, SAS, SPSS, R, Facets, ITEMAN, HLM, MPlus, Conquest, BILOG, Excel