Item Parameter Drift in Computer Adaptive Testing Due to Lack of Content Knowledge within Sub-populations

BY

BEYZA AKSU DÜNYA
B.S., Istanbul University, 2008
M.Ed., Boston College, 2012

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, 2017

Chicago, Illinois

Defense Committee:
Everett V. Smith Jr., Chair and Advisor
Kimberly Lawless
Yue Yin
John Stahl, Pearson Vue
Nicole Makas Risk, American Medical Technologists

This dissertation is dedicated to my husband, Hamza Dünya. You provided me endless emotional support, generously shared every task of being a graduate parent with me, and took care of our baby boy with love when I was working on my dissertation. Thank you for all of your love, patience, and support.

To my little son, Yağız A. Dünya, you shone in my darkest days and brightened my life with your presence. You are my sunshine; you are my hope. I love you.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my academic advisor and committee chair, Dr. Everett Smith, for his continuous support, encouragement, and professional expertise during my doctoral studies and research at University of Illinois at Chicago. Without his scholarly guidance, this dissertation would not have been possible.

I would like to present my sincerest appreciation to my committee member, Dr. John Stahl, for his dedicated guidance, understanding, and patience as well as his professional knowledge on CAT. He helped me generously, regardless time and place, no matter if it was during business hours or the weekend. He made time from his busy schedule to meet me and discussed my research problems to make my research valuable and significant.

I would like to present a big thank to my fellow, my colleague, my committee member Dr. Nicole Risk, for listening me, understanding me, and guiding me not only during my dissertation, but also through my entire doctoral journey. Her dedicated guidance will be forever remembered.

I also would like to thank the other members of my committee, Professor Kimberly Lawless and Professor Yue Yin, for providing their valued expertise, suggestions, advice, and support throughout my dissertation. Their deep insights and constructive feedback on my dissertation proposal added a lot to my study.

I also would like to express my warmest thanks to Dr. Kirk Becker for all the technical and software support, his programming guidance, and his willingness to help me for completing this study.

ACKNOWLEDGEMENTS (continued)

I thankfully acknowledge the Republic of Turkey, Ministry of National Education for their sponsorship which made my graduate studies possible.

Finally, I would like to thank my husband, Hamza Dünya, for being with me, understanding me, calming me down in my most frustrated moments, and providing me endless emotional support all the time. Without his belief with me, I would not be where I am right now.

BAD

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1PL	One-parameter Logistic Model
2PL	Two-parameter Logistic Model
3PL	Three-parameter Logistic Model
CAT	Computer Adaptive Tests
CBT	Computer-Based Tests
CI	Confidence Interval
CTT	Classical Test Theory
DIF	Differential Item Functioning
DTF	Differential Test Functioning
FN	False Negative
HLM	Hierarchical Linear Model
ICC	Item Characteristic Curve
IPD	Item Parameter Drift
IRT	Item Response Theory
MAD	Mean Absolute Difference
MCAT	Multidimensional Computer Adaptive Testing
MLE	Maximum Likelihood Estimation
MNSQ	Mean-Square
MRM	Mixed Rasch Models
PBT	Paper-Based Tests
PPD	Person Parameter Drift

RMSE	Root Mean-Squared Error
SEM	Standard Error of Measurement
TCC	Test Characteristic Curve

SUMMARY

I conducted this study to investigate potential item parameter drift (IPD) impact on person ability estimation and classification of examinees to achievement levels in computer adaptive testing (CAT). The overarching goal was to evaluate the impact of IPD that occurs due to differential access to content knowledge attributable to possible curricular, instructional, infrastructural, and practice differences in CAT. I conducted a series of simulations using two hypothetical item banks. I specified number of items in the test, mean of the item difficulties, and examine abilities based on an operational CAT exam. To address my research questions, I manipulated three factors; (a) percentage of IPD items in the test, (b) percentage of examinees that had poorer access to content knowledge, and (c) targeting of the item pool to the examinees. To serve as baseline, I simulated two non-IPD conditions for both item pools and compared the results from IPD conditions with these baseline conditions. I evaluated IPD impact on ability estimation precision and classification accuracy using multiple indicators including bias, root mean square error (RMSE), mean absolute difference (MAD), number and percentages of false decisions and their significance, correlations between estimated and true ability parameters, and Spearman's rank order correlations. In addition, I evaluated person fit indices of misclassified examinees to assess the effectiveness of these indices in detecting misfit that can be observed as a result of IPD in real life tests.

The results revealed that IPD exposed to a sub-group of examinees can affect classification accuracy of those examinees substantially, but IPD impact on average ability estimation was minor. Given the fact that an examinee misclassified into a lower achievement level due to unequal opportunity to learn and perform may result in fairness issues in educational

testing, the findings indicate potential reasons of reduced access to content knowledge, such as curricular, instructional, infrastructural, and practice differences in CAT, may derail the examination process. The study provides useful information to the states and districts planning to implement or are currently implementing CAT as part of their assessments by emphasizing every examinee should be given equal opportunity to learn the content and demonstrate their true ability in the exam.

I. INTRODUCTION

A. Background

Starting with the use of computers, applications of computer technology to assessment were facilitated in terms of its convenience of administration, test security, and efficiency. In conjunction with the important developments in measurement theory (Item Response Theory (IRT); Birnbaum, 1968; Lord, 1952; Rasch, 1960), testing programs started moving from traditional paper-and-pencil tests to computer adaptive tests (CAT) (e.g., American Society of Clinical Pathology, National Council of State Boards of Nursing). A computer adaptive test (CAT) is a form of computer-based test administration in which every single examinee takes a customized test (Gershon & Bergstrom, 1995). An examinee's ability is refined after each response is made, and items are selected sequentially based on the most recent ability estimate (Wainer et al., 2010). More specifically, if an examinee is directed a question that is too difficult for his/her ability level, the following question asked would be easier. Thus, we can know the most about an examinee's ability because we can select appropriate questions that are at the same level as the examinee's proficiency.

There are four main components to develop a CAT system: (a) an item pool or bank from which to select items, (b) item selection criteria, (c) a method for scoring the test, (d) a stopping rule to terminate the test (Green, Bock, Humphries, Linn, & Reckase, 1984). The collection of test items that are ready to use during the test is stored in a computer media and called an item bank or item pool. These items are either a set of items generated by computer software with specified item parameters relying on an IRT model or a set of existing items that have been calibrated using an IRT model (Reckase, 2009).

Another important issue while developing a CAT is determining the item selection method for selecting a test item from the bank. The first item selected in the test is typically slightly below the average ability of the population of all test takers. This average of the population is usually determined based on comprehensive pretesting of the examinee population. Starting with an item that is slightly below the average ability of the population reduces the chance that the first item appearing on the test will be too difficult for examinees. The criteria for choosing the next item ranges from administering items that maximize the reliability of each examinee's score to determining items based on more complex test assembly approaches. Nevertheless, the main item selection criteria in a CAT is to minimize the measurement error associated with examinee score while maximizing test information function (Luecht & Sireci, 2011).

CAT relies on IRT techniques to score examinees. In IRT, the metric used to indicate the difficulty of an item is the same with the metric that is used to identify ability of an examinee. Therefore, an average difficulty item has the same value on the scale with an average ability examinee (Luecht & Sireci, 2011).

Lastly, stopping rules are used to terminate a CAT session. In some situations, each examinee is administered a pre-determined number of items regardless of the measurement error. This is called fixed-length CAT (e.g., College Board, 1993; Northwest Evaluation Association, 2005). In other cases, a variable-length testing approach is used, in which the CAT session ends when some pre-specified level of measurement precision is met (Luecht & Sireci, 2011).

Numerous studies compared the performance of CAT with computer-based tests (CBT) and paper-based tests (PBT). CATs have been found to be superior to their counterparts in terms of their administrative features and psychometric properties (Wang & Kolen, 2001). Several

important advantages associated with CAT include but are not limited to (a) improved measurement precision with lower measurement error, (b) convenience while administering, scoring, and reporting, and (c) enhanced security and fairness (Georgiadou, Triantafillou, & Economides, 2007; Wainer et al., 2010). Along with the advantages that CAT provides, some concerns about the performance of CAT compared to CBT/PBT have been well-researched since the 1970s. Early concerns associated with feasibility limitations of CAT such as computer errors, power failures, and hardware quality have been greatly reduced or eliminated with the advances in technology (Gershon, 2005; Rudner, 1998). Cost-related limitations, particularly at the stage of item bank development and maintenance, and psychometric considerations are discussed as disadvantages associated with CAT in the literature (Gershon, 2005; Rudner, 1998).

Despite all the advantages that CAT provides, validity of inferences in CAT depends on the extent to which test scores accurately reflect examinee knowledge/skills (Goldstein, 1983). The accuracy of scores can be impacted by item parameter drift (IPD). Goldstein (1983) defined IPD as the differential change of item parameters over testing occasions. IPD is also interpreted as a form of differential item functioning (DIF) where differential functioning of items occurs across examinee groups associated with different testing occasions or time points (Goldstein, 1983). When operational test items drawn from item pools are used repeatedly in CATs, item parameters have been found to drift or change over time (Goldstein, 1983). When item parameters change across multiple administrations or time points more than would be expected due to measurement error alone, it cannot be assumed that parameter values are invariant across testing occasions.

As examinee ability estimation is a function of item parameters under IRT framework, we expect examinee ability estimates will change if item parameters change (Wells, Subkoviak,

& Serlin, 2002). Therefore, failure to detect the presence of IPD may lead the violation of a major IRT assumption, measurement invariance, which says that examinees with the same latent trait should have the same probability of getting an item correct (Goldstein, 1983). This violation may diminish the effectiveness of items and impact precision of person ability estimates, and poses a threat for testing programs that need to maintain a stable scale over time. Failing to identify IPD may also create a disadvantage for individuals or groups of examinees and jeopardize conclusions about trends in large-scale assessments. Moreover, drifting items may weaken the linking across assessments and undermine comparability of scores across different forms or different time points of an assessment (Donoghue & Isham, 1998).

IPD is observed as a central concern in IRT applications, but it may also be associated with fluctuations in *p*-values and point-biserial estimates under classical test theory (CTT) framework. Moreover, IPD may occur on both adaptive tests and fixed-item tests (Clark, 2013). Studies have investigated various aspects of IPD in both fixed-item tests and CATs using different IRT models. Their investigations included factors that cause IPD (Bergstrom, Stahl, & Netzky, 2001; Bock, Muraki, & Pfeiffenberger, 1988; Chan, Drasgow, & Sawin, 1999; Goldstein, 1983; Mislevy, 1982) and methods for detection of IPD in fixed-item tests (Donoghue & Isham, 1998) and in CATs (Han, 2003; Masters, Muckle, & Bontempo, 2009).

Although several researchers have investigated the factors that cause drift and how to detect drift, few have examined the effects of IPD, particularly in a CAT environment. Wells, Subkoviak, and Serlin (2002) examined the effect of IPD on ability estimates and found that IPD has little effect on ability estimates under two simulated conditions. However, they noted that the effect of IPD may be detrimental under different conditions in which the percentage of drifting items, magnitude of drift, test length and testing occasions change. Han, Wells, and Sireci (2012)

examined the impact of different combinations of IPD on test equating, examinee scores, and classification determinations and found that IPD had substantial impact on precision of score estimates and classification accuracy under some of the study conditions. Babcock and Albano (2012) studied scale stability in the presence of IPD and how deviance from item parameter and scale stability affect item parameter recovery and classification accuracy. Their findings suggested that the scale maintained acceptable parameter recovery under the conditions where there was equal item drift in both directions or under conditions of small to moderate periodic changes in the latent trait; however, they added that substantial item drift or major changes in the ability can dramatically diminish scale stability. Despite the potential impact of IPD on ability estimates, pass-fail decisions, classifications, and scale stability, available research investigating the impact of IPD is very limited and has revealed inconsistent findings. This warrants assessing the outcomes of drift under various contexts, including K-12.

CAT in K-12 for interim, formative, and summative assessments are relatively new and its benefits and drawbacks are still being discussed. According to the most recent literature, a handful of states including Delaware, Utah, Hawaii, and Oregon are using CAT on a widespread basis in K-12 (Davis, 2012). In addition, at least 20 states indicated their plans to use CAT versions of their tests in K-12 starting from 2014-15 (Davis, 2012). Despite the benefits of CAT, such as pinpointing a student's proficiency level more precisely with shorter tests, there are some concerns attached to its use in K-12. One major concern about CAT use relates to curricular, instructional, infrastructural, and practice differences across schools, districts, and states (Ash, 2008; Han & Guo, 2011; Kingsbury & Wise, 2011; Risk, 2015). Such factors lead to differences in opportunities to learn the test material among examinees. As Kingston (as cited in Ash, 2008, p. 4) states,

> The adaptive testing model assumes that everyone has taken courses or learned subjects in the same way. Subjects in which curriculum varies greatly from place to place presents a particularly difficult challenge for computer-adaptive tests, which are often created on a national or state level.

He urged test developers to carefully assess whether the test is the same in a rural area as it is in the center of an urban area, since IRT assumes a universal definition of hard and easy (as cited in Ash, 2008, p. 4). When examinees are not supported to perform to their full potential, test scores' validity and fairness are impacted and consequential decisions linked to their test scores do not reflect their actual knowledge, ability, and skills (NRC, 2007). Limited number of studies examined IPD impact in CAT due to curricular and practice differences in cases where items drifted easier in CAT. They analyzed IPD impact on decisions associated with type II errors such as an examinee falsely passing (Guo, 2009; Han & Guo, 2011). Those decisions may be serious in certification and licensure exams where people are licensed in fields such as medicine, and pharmacy. On the other hand, while a student falsely classified into a higher achievement level due to content/item familiarity and practice may not create a serious validity threat in most educational tests, a student falsely failed or classified into a lower level due to differential access to content knowledge that can attributed to the factors such as curricular differences, poor instruction, infrastructural issues, and lack of adequate practice may demolish test fairness. Such test result errors may even lead to social justice issues, depending on the exam's stakes. Therefore, it is important to study potential IPD impact when a group of test

takers are affected by differences in curriculum, instruction, resources/infrastructure, or exam practice.

B. Purpose of the Study

The purpose of this simulation study was to evaluate the impact of IPD, which occurs due to lack of exam content knowledge on person parameter estimation and classification accuracy in a CAT, when factors such as percentage of drifting items, percentage of examinees that had poorer access to content knowledge, and item pool targeting, vary. Various reasons may cause a lack of content knowledge in a K-12 context. These reasons may include insufficient school infrastructure, poor instruction, curricular differences, and lack of sufficient practice. The main goal of the study was to evaluate the impact of changes on examinees' ability and examinee-item interaction because of lack of content knowledge across examinees from different strata (i.e. schools and districts) as the IPD source in CAT. Specifically, for this study, items were allowed to drift only for examinees that had poorer access to exam content knowledge.

C. Study Approach and Research Questions

The vast majority of possible IPD sources are changes in the persons interacting with items (J. Stahl, personal communication, September 9, 2014). Those sources include differential opportunities to practice and learn, cheating, and curricular updates (Han & Guo, 2011; NRC, 2007; Risk, 2015). As such, I created a simulation scenario where Person Parameter Drift (PPD) existed for a group of affected persons. This scenario assumed individuals who attended certain schools were affected by lack of particular content knowledge, which can be attributable to various sources in real life testing such as poor instruction, ill-equipped schools, changes in curriculum, and less practice on a content area. Thus, students who attended those schools were disadvantaged on all the items in that particular content area.

I explored three drift factors: (a) the percentage of the content area that includes drift items, (b) the percentage of PPD examinees in the examinee sample, and (c) targeting of the item pool to assess the impact of IPD on ability estimates and classification accuracy. The content area percentage with IPD items varied by three levels: 20%, 40%, and 60% of the item pool. The percentage of PPD examinees in the sample varied by four levels: 20%, 30%, 40%, and 50% of the examinee sample. Lastly, the targeting of the item pool varied by two: well-targeted item pool with mean item difficulty very close to the mean ability of examinees and easy item pool with mean item difficulty that is 1 logit below the mean ability of examinees. The items in the selected content area drifted within a range from 0.50 to 1.00 logits from their original values. All items drifted in one direction; they became harder, since poorer access to content knowledge makes these particular items harder to examinees in real life testing situations. This was a fully crossed design study: 3 (content percentage) x 4 (PPD percentage) x 2 (item pools). I examined both main and interaction effects. To evaluate the impact, I compared the results of a baseline CAT for each item pool (with no drift) with the modified CAT (with IPD), according to the scenario.

The overarching research question was, "To what extent are person measure estimates and classification accuracy impacted when IPD exists in CAT?" To answer this question, these particular questions were addressed:

1. What is the impact of IPD on person measure estimates when only a sub-group of people are affected by the drift?
2. What is the impact of IPD on classification of examinees to achievement levels when only a sub-group of people are affected by the drift?

3. Are the effects of IPD on person measure estimates and classification of examinees consistent across three factors of drift: proportion of IPD items in the pool, proportion of PPD examinees in the sample, and item pool targeting?
4. Does item pool targeting change the effect of IPD on low, medium, and high ability PPD examinees' classification to achievement levels differently?
5. Holding all else constant, does the change in the impact of IPD over the levels of one factor depend on the level of another factor?

D. Significance of the Study

Educational measurement research has repeatedly cited factors that may lead differential access to learning opportunity including curricular, instructional, infrastructural, and practice differences as important sources of IPD (Ash, 2008; Han & Guo, 2011; Kingsbury & Wise, 2011; NRC, 2007; Risk, 2015). Unlike other factors that may cause IPD, such as security breaches or cultural and historical events, the impact of such factors is not always apparent and differs across examinees (Han & Guo, 2011). When the effects of such factors are not diagnosed properly, certain items may become harder to examinees who lacked the opportunity to learn exam content. Such situations may provide an unfair advantage or disadvantage to some examinees, impact their test performance, the ability estimates, and in turn, impact decisions made from test scores. In addition to the potential consequences for individual examinees, drifting items may eventually threaten pretest items as new pretest items are calibrated using person measure estimates in CAT. The vast majority of existing simulation studies assumed that every examinee in a sample is affected by IPD equally, despite the fact that IPD may impact examinees in varying degrees in operational tests. This research addressed an issue not

sufficiently explored before: IPD impact in CAT resulting in person ability changes and vary across examinees.

This will contribute to the IPD literature, particularly in education where CAT is used for various purposes, such as interim and summative evaluations, part of college admission decisions, and statewide progress decisions. The findings could also provide useful information to testing organizations and test developers about the amount of drift resulting from examinees' lack of content knowledge, as well as the potential consequences of such drift on the ability estimates and classification results. The implications of this study extend beyond research in the education field to all certification and licensure programs using CAT; all of which can be affected by examinees' differential access to exam content knowledge.

II. LITERATURE REVIEW

Research has repeatedly found that item parameters may differ across examinee subgroups and test administrations in both fixed-item tests and computer adaptive tests (CAT) (Bock, Muraki, & Pfeiffenberger, 1988; Clark, 2013; Goldstein, 1983). Parameter change across subgroups of examinees is defined as differential item functioning (DIF) (Pine, 1977) while parameter change over time is defined as item parameter drift (IPD) (Bock, Muraki, & Pfeiffenberger, 1988; Goldstein, 1983). Although researchers studied DIF extensively, the literature on IPD is relatively small. In this chapter, I will present and evaluate relevant research on IPD. The chapter is organized as three main sections: (a) a brief information about IRT and Rasch models, (b) an extensive definition of CAT, and (c) an overview of existing IPD research. IPD is defined as change in item parameters over subsequent testing occasions and the number of item parameters varies depending on the IRT model. Hence, IPD is defined differently based on the IRT model being used. Before evaluating the research on IPD, I will provide information about IRT and Rasch models. Then, I will explain CAT in detail, including the history and development of CAT, processes and components that CAT involves, and advantages and disadvantages of CAT. I will then elaborate on available research on IPD in both fixed-item tests and CAT, including the sources of IPD, detection of IPD, and the impact and consequences of IPD.

A. Item Response Theory and Rasch Model

All analyses that were conducted in this study relied on the Rasch model. Modern adaptive testing uses Rasch or IRT models as their mathematical foundations (Gershon, 2005). Hence, an overview of the Rasch model and IRT is needed. IRT is an item-level, latent trait measurement model (Lord, 1980). In IRT, item characteristic curves (ICC) are used to estimate

the probability of a correct response of a person on an item, given a latent ability. Examinees with more latent ability have higher probabilities for giving a correct answer to items than examinees with lower latent ability (Hambleton, 1989). There are multiple IRT models, each of which employ a logistic function that differs according to the number of item parameters that describe and name the particular model (as cited in McCoy, 2010, p. 6). The most commonly used models for CATs are the Rasch model and the 2- and 3-parameter logistic IRT models (2PL and 3PL models, respectively) (Gershon, 2005; Hambleton, Swaminathan, & Rogers, 1991). In these models, the probability of answering an item correctly depends on item parameters and an examinee's latent ability level. In the 3PL IRT model, item parameters are the item discrimination (*a*), guessing (*c*), and the difficulty (*b*) parameters. The item parameters within the 2PL model include item discrimination (*a*) and the difficulty (*b*) parameters. In the Rasch model, the only item parameter considered is the difficulty (*b*) parameter.

Within Rasch measurement models, there are three different models, to include the dichotomous model (Rasch, 1960), the rating scale model, and the partial credit model (Andrich, 1978; Wright & Masters, 1982) function well with adaptive tests and have been used extensively (Gershon, 2005). In the Rasch model, the latent ability described by the items on a test is a continuous logit metric from negative to positive infinity. All possible test item difficulties and examinee ability measures are placed on this logit continuum. This single continuum usually implies there is a point where the difference between the estimate of item difficulty and examinee ability is as close to zero as possible. This idea is crucial in CAT (Gershon, 2005). Most of the CAT algorithms select items to alter difficulty of the test to the current ability estimate of an examinee by choosing an item at the point where the difference between examinee ability and item difficulty is zero. By targeting item difficulty to examinee ability, CAT

maximizes the statistical information from each item. When the information is maximized, the standard error of measurement (SEM) is minimized. Therefore, by adaptive administration of items, CAT improves measurement precision and reduces test length (Gershon, 2005). If the item difficulty measure is higher than the examinee ability measure, the Rasch model predicts that the examinee will have less than a 50% probability of correctly answering the item; if the item difficulty measure is lower than the examinee ability measure, the Rasch model predicts that the examinee will have a greater than 50% probability of correctly answering that item (Bergstrom & Lunz, 1999; Wright & Stone, 1979). CAT algorithms that use 2PL and 3PL IRT models incorporate the additional discrimination and guessing parameters when selecting items and estimating ability. This study focuses on the Rasch model; for more information regarding 2PL and 3PL models see Wainer et al. (2010).

B. Computer Adaptive Testing

1. Historical background of adaptive testing and computer adaptive testing

One of the earliest examples of adaptive testing is the Stanford-Binet intelligence test (Reckase, 1989). Binet started a test with item subsets that matched what he believed was the examinee's likely ability level. If the examinee succeeded on the item subsets, he continued to give harder item subsets until the examinee failed regularly. If the examinee failed the initially administered item subset, he gave relatively easier item subsets until the examinee succeeded regularly (Linacre, 2000). Other early examples of adaptive testing include two-stage testing (Angoff & Huddleston, 1958) and pyramidal testing (Krathwohl & Huyser, 1956). The basic principle of these tests was to order items based on their *p*-values and to develop fixed paths through the items. Then, tests were matched to test takers via these paths (Reckase, 1989). The development of the Rasch and IRT models led to the creation of large-scale calibrated item

banks in which items were placed on the same scale (Choppin, 1985). These calibrated item banks were used to create individualized tests. Each individualized test was then equated to every other test that was drawn from the bank (Wright & Bell, 1984).

CAT was implemented on mainframe systems dating back to the 1960s, limiting its use to the military and private training companies who could afford to purchase expensive hardware (Gershon, 2005). The Armed Services Vocational Aptitude Battery (ASVAB) was one of the earliest examples of CAT (McBride & Sympson, 1985; Moreno, Wetzel, McBride, & Weiss, 1984). This test was developed by the Navy Personnel Research and Development Center (NPRDC) with the purpose of better selecting personnel and classifying them in the Armed Services by increasing accuracy of test scores, reducing test compromise, and reducing testing time. The NPRDC tested and implemented several generations of CAT-ASVAB from 1979 to 1992. It remained in operational use until 1996 when it was replaced with a new generation system (Gershon, 2005). Other early examples of CAT include the Educational Testing Service Wide-Range Vocabulary test (Kreitzberg & Jones, 1980; Lord, 1977), College Board Advanced Placement tests (Ward, 1988), Assessment System's CAT software (Vale, 1981), the Psychological Corporation's CAT version of the Differential Aptitude Battery (McBride, 1988), and the U.S. Army's Computerized Adaptive Screening Test (CAST; Sands & Gade, 1983*).* The rapid developments and availability of efficient, accessible, and powerful technologies have led to increased use of CAT in various settings including military aptitude, higher education admissions, certification in medicine and technology, nursing licensure, and interim school- and district-based programs (Linacre, 2000; Way et al., 2010). Recently, it has begun to be used in summative statewide programs in K-12. Some examples of widely known high-stakes exams administered in adaptive format are the Graduate Management Admission Test® (GMAT®) and

the National Council Licensure Examination *(*NCLEX®*)*. The NCLEX® was started to be implemented in adaptive format in 1994 and the GMAT® was transformed to CAT in 1997 (Guernsey, 2000).

2. The process of computer adaptive testing

CAT is widely used for both estimating examinees' ability level and for estimating examinees' attitude on a particular latent trait (Gershon, 2005). Therefore, in the following sections, the terms "ability" and "trait level" will be used interchangeably to refer to the person measure. Likewise, the terms "item level" and "item severity" will be used interchangeably with item difficulty. In this section, fundamental procedures that CAT involves will be explained.

a. Developing an item pool

The first step in the CAT process is to create an item pool. An item pool or item bank is an accumulation of test items that may be administered during the test. An item pool also includes item parameters and details about the items' development, use and re-calibration processes (Linacre, 2000). The items are "coded by subject area, instructional level, instructional objective measured, and various pertinent item characteristics" (Gronlund, 1998, p. 130). As CATs tend to be much shorter than traditional tests, each item is more critical. This makes item pool development in CAT arduous work (Wainer et al., 2010). Flaugher (2000, p. 38) states that,

> The better quality of the item pool, the better job the adaptive algorithm can do. The best and most sophisticated adaptive program cannot function if it is held in check by a limited pool of items or items of poor quality.

Wise (1997) suggested two criteria to assess the quality of an item pool: (a) there should be sufficient number of items in the pool that can used as informative items during a test session, and (b) the items in the pool must yield sufficient information at the ability level which the test

developer has the greatest interest. In addition, item parameters must remain unchanged to maintain the integrity of CAT. Davey and Nering (2002) supported these guidelines by stating that in an ideal item pool, there should be an adequate number of items to create multiple test forms for examinees with various ability levels.

b. Item selection criteria

Next, item selection criteria much be established. While estimating examinees' ability and matching that estimate to the next best item from the item pool, there are several key issues that need to be determined: how to select the first item, how to select the next item to be administered after seeing the examinee's response to the current one, how to maintain content balancing, and how to control overexposure of the items. I will address how CAT handles each of these issues below.

1) How to select the first item

Selection of the first item in CAT depends on the goal of the test. If the test is a criterion-referenced test where the goal is to decide if an examinee falls beyond a threshold of knowledge or skills, the first item is usually selected at or around the threshold (pass score) of the test. If the test is a trait-based, norm-referenced test where the purpose is to find out a particular ability or trait level of an examinee in relation to the other's ability or trait level, the first item administered is usually selected from slightly below the mean ability of the population being tested. This mean ability is usually determined based on extensive pretesting of the examinee population (Luecht & Sireci, 2011).

2) How to select the next item(s)

Selection of the second and subsequent items are dictated by examinee's answers to the previous item(s). The adaptive algorithm of CAT for item selection

usually depends on the measurement model used (Lord & Novick, 1968; Rasch, 1960; Rasch, 1980; Wright & Stone, 1979). One of the most commonly cited item selection approaches in the psychometric literature is the Maximum Information (MI) method (Lord, 1977; 1981). The MI method suggests that the unused item that provides the most information I_i $(\hat{\theta})$ at the respondent's currently estimated trait level (θ on the Rasch dimension) is chosen as the next item (Barrada, Olea, Ponsada, & Abad, 2009; Lange, 2006). Yet, this method may lead to a highly unbalanced distribution of content and exposure of items, with some items presented to almost all examinees and many that are never administered (Barrada, Olea, Ponsada, & Abad, 2009). Thus, item selection strategies with exposure control mechanisms and content balancing strategies have been developed.

3) Item overexposure

The adaptive algorithm for item selection usually depends on the item response model employed as well as considerations such as balancing content and controlling item overexposure. Without such constraints, CAT algorithms select the statistically "best" items. This may result in some items being more likely to be seen than others, particularly the ones in the beginning of the adaptive process; however, with a good item exposure mechanism, one can increase the use of the least popular items and decrease the exposure of most popular items without decreasing measurement precision (Pastor, Dodd, & Chang, 2002). Therefore, item selection strategies that limit item exposure were developed in order to assure that some items are not overexposed (Rudner, 1998). The frequently cited item selection approaches with item overexposure control mechanisms include: Maximum Information and Stocking and Lewis (aka, conditional SH) item exposure control procedure (Stocking & Lewis, 1998) and the Sympson-Hetter (SH) unconditional item exposure control procedure (Hetter &

Sympson, 1997; Sympson & Hetter, 1985). In the unconditional procedure by Sympson and Hetter (1985; 1997), the probability of administering an item from a pool is controlled through a formula that defines the probability of administering an item, P(A), as the product of the probability of selecting an item P(S) and the conditional probability of administering an item when it is selected P(A|S). It aims to constrain P(A) for each item to a specified target exposure rate, *r,* by manipulating the conditional probability of that item P(A|S), that can also be considered as the item exposure control parameter, K_I. In the conditional model proposed by Stocking and Lewis (1998), the exposure control parameter controls not only the number of times that an item is selected, but also the ability level of examinee who sees the item. It ensures that an item is seen by individuals from various abilities. Other examples of conditional item selection approach included Davey-Parshall method (Davey & Parshall, 1995) which sets an exposure control parameter conditional on the items that have already been administered, and a tri-conditional method, proposed by Nering, Davey, and Thompson (1998), which combines SH, conditional SH, and the Davey-Parshall to maximize the benefits of each technique (Pastor, Dodd, & Chang, 2002).

4) <u>Content balancing</u>

Content balancing is another issue that can be addressed via item selection algorithms. One of the typical requirements of CAT is measuring one construct at a time. Yet, one construct or dimension may be composed of several sub-areas that are perceived as more/less important than others to the test developer. In these situations, CAT can select the next item to maximize information while also conforming to the specified content schema (Gershon, 1995; 2005). This can be achieved with content balancing in CAT.

c. Scoring the test

The CAT algorithm is an iterative process (Rudner, 1998). It begins by assuming an initial estimate of examinee ability (location) and then selecting an optimal test item according to an item selection rule described previously. A score is given based on the examinee response to the item and the examinee's ability estimate is updated based on the responses to all administered items (Reckase, 2009). Two general classes of methods are typically used to update estimate of location: maximum likelihood estimation (MLE) and Bayesian estimation (Reckase, 2009). In MLE, examinee ability is updated by using the difficulty of the items already administered and the response to the most recent item (Gershon, 2005). The method searches for the θ that provides the highest value for the equation based on the item parameters and examinee scores for the administered items (Reckase, 2009). The next item administered is the one that provides the maximum information; the item that has difficulty parameter closest to the examinee's most recent ability estimate. MLE yields unstable estimates for short tests under the 2PL and 3PL IRT models and is, therefore, not preferable for CAT when these models are used (Gershon, 2005; Reckase, 2009). Bayesian estimation algorithms are based on the normal distribution or a known typical distribution of examinees taking a particular exam (Segall, 1996). The algorithm begins by administering items as if the true ability estimate of an examinee is around the mean of the population distribution and updates each "prior" distribution with the new information based on the examinee response. The Bayesian estimation procedure is usually used with 2PL and 3PL IRT models in CAT (Gershon, 2005).

d. Stopping rule

The method for determining when to finish the recurring procedure of CAT is called a "stopping rule." There are several different methods that are used to terminate a

CAT session. In some cases, all examinees are administered a pre-determined fixed number of items regardless the measurement error associated with the scores. This procedure is called fixed-length CAT (e.g., College Board, 1993; Northwest Evaluation Association, 2005). In other cases, the variable-length CAT approach is used. In this approach, not every examinee answers the same number of items. The stopping rule might end the exam when the ability estimation has reached an acceptable level of precision or end when an acceptable level of confidence has been met to make a decision about the examinee (Reckase, 2009; Luecht & Sireci, 2011).

Test developers need to determine the appropriate approach to terminating the test based on the purpose of exam. The fixed-length stopping rule is typically used in norm-referenced contexts (e.g., achievement testing). A test session stops when the measurement error associated with the score falls below a previously designated level (Lord, 1980). This criterion provides a minimum standard level of reliability for each examinee's estimated score. The variable-length stopping rule is usually employed in criterion-referenced contexts where a specific threshold is used to make a decision on examinees (e.g., licensure and certification testing). A session ends when an examinee's score is clearly above or below that specific threshold (Luecht & Sireci, 2011).

Other stopping rules have been developed more recently, such as the predicted standard error reduction stopping rule (PSER) that uses the predicted posterior variance to identify the change in standard error that would occur as a result of administrating additional items (Choi, Grady, & Dodd, 2011). This rule has been proposed to function primarily for short CATs. Other stopping rules for ending the CAT session may to be developed in the future.

3. Advantages of computer adaptive testing

Four basic standards of testing include reliability, validity, fairness and feasibility (Gershon, 2005). CAT utilizes the advances of technology and measurement theory to develop and deliver tests that align with these standards.

CAT provides test developers the opportunity to develop more reliable tests that increase measurement precision. CAT algorithms select items from an item bank appropriate to the current examinee's ability estimate and maximize the information from each item. When test information is maximized, the SEM is minimized; when information is minimized, the exam can be shorter without loss of reliability (Gershon, 2005; Lunz & Bergstrom, 2011). Therefore, CAT can substantially reduce test length without losing measurement precision compared to non-adaptive tests (Weiss, 1983; Weiss & Kingsbury, 1984).

There are many definitions of test validity. Kelley (1927) defined validity as "a test that measures what it purports to measure." Cronbach and Meehl (1955) emphasized one important aspect of validity as construct validity, meaning that the inferences made from the test scores are meaningful and useful. A threat to construct validity arises when some unrelated sub-dimensions contaminate the measurement. This is called construct-irrelevant variance (Messick, 1989). CAT is administered in significantly less time (by less than 50%) than non-adaptive tests as fewer items are needed to reach acceptable precision. Shorter tests can reduce fatigue, anxiety and burden and can enhance validity when these factors that may introduce construct-irrelevant variance are reduced (Gershon & Bergstrom, 1991; Huff & Sireci, 2001; Rudner, 1998).

Content validity is another crucial aspect of validity addresses when the test's content reflects critical skills and knowledge (Cronbach & Meehl, 1955). CAT can enhance content validity with a sophisticated item selection algorithm by ensuring that content is balanced for

each test-taker (Gershon, 2005). CAT also avoids administering irrelevant items that are not appropriately targeted to examinees' ability (e.g., too easy to hard). CAT algorithms can enhance validity by eliminating types of items that may provoke unwanted answering behaviors such as guessing, careless mistakes, and response patterns (Linacre, 2000).

A well-developed item bank in CAT can also promote fairness. Human intervention is removed from the selection of test forms to delivering and scoring the test. Further, each examinee has the same opportunity to demonstrate ability or achievement as any other examinee (Gershon, 2005; Rudner, 1998). Adaptive test algorithms that allow administering different sets of items to examinees reduce threats, such as cheating, and improves test security. This improved security and elimination of human intervention enhance fairness with CAT (Gershon, 2005; Lunz & Bergstrom, 1991). In CTT, item difficulty depends on the subpopulation taking the test and varies based on the distribution of examinee ability in the sample and characteristics of the items administered. Therefore, an examinee's performance is dependent upon the ability distribution of examinees that take the test; however, when a test is drawn from an IRT or Rasch calibrated item bank, the estimate of an examinee's ability is supposed to be statistically equivalent regardless whether he/she is administered difficult or easy items. Thus, the use of Rasch and IRT models is another factor that increases fairness with CAT (Lunz & Bergstrom, 2011).

CAT also provides examinees important advantages regarding test scheduling and prompt score reporting (Gershon, 2005; Rudner, 1998). In past times, PBTs were provided on certain occasions and if examinees missed it, they had to wait for the next administration. In many cases, CAT provides examinees increased general accessibility to high-stakes tests (Gershon, 2005).

Tests may be given on demand and typically the test scores are immediately available (Rudner, 1998).

Another advantage of CAT is the potential use of graphical images and multimedia presentation within items, which enables measuring concepts not possible with text-only formats. In addition, a mistyped or miskeyed item would be less likely to impact decisions of high-stakes tests with CAT, as such an item would only affect a subgroup of test takers (Linacre, 2000).

Another important advantage of CAT is cost feasibility, particularly for already established testing programs. The main costs associated with CAT include test content development, administration, scoring, and reporting. The cost associated with developing an item bank for CAT varies depending on the exam's purpose. For example, in high-stakes, norm-referenced tests administered to thousands of examinees, a large number of items is needed to ensure test security and to cover a wide range of ability; however, in low-stakes and/or self-assessment tests, a small item bank with less than 100 items can be sufficient (Gershon, 2005).

Needing to write a new test for every administration in non-adaptive tests is replaced by bank maintenance tasks in CAT. Once item pools have been developed, item parameters are calibrated for every item in the bank. The cost associated with this task varies depending on whether the testing program uses previously administered items or newly developed items. In the first case, calculation of item parameters only requires re-analyzing the old datasets. Conversely, for all newly-developed items in a bank, piloting the items on hundreds of examinees may be costly. However, once the CAT program is established, organizations can experience reduced costs as they only need to maintain the item bank instead of writing completely new items (Gershon, 2005). The costs associated with the test delivery and administration depend on the security level of the test. While low-stakes CAT exams can be administered over internet, high-

stakes tests are administered in secured settings such as test administration centers. The costs for scoring and reporting tests are greatly reduced in CAT as there are no bubble sheets to collect and scan. Reporting is done electronically or on paper at the time of testing for many organizations, and it eliminates the need to generate print reports (Gershon, 2005).

4. Limitations of computer adaptive testing

Despite important advantages that CAT provides to examinees, test developers, and organizations, there are some limitations and practical considerations. These limitations and considerations can be grouped into three categories: (a) feasibility limitations, (b) cost-related limitations, and (c) psychometric limitations.

Although most of the feasibility limitations are reduced with improvements in technology and computer literacy of people, there is some research addressing them. Rudner (1998) examined the effectiveness of CAT for large-scale adaptive tests such as the SAT and GMAT® and stated that CATs are only feasible if the facility has enough hardware for a large number of examinees. Davis (2012) addressed this concern particularly in statewide K-12 testing by indicating that states have to have enough devices and bandwidth to administer CAT. Bugbee and Bernt (1990) compared the performance of paper-pencil and computerized tests and found that examinees are constrained in computerized tests, as they cannot use common strategies such as underlying text and scratching out eliminated choices. On the other hand, a few studies comparing performances of the two test formats demonstrated the absence of inability of CAT to capture measures originally assessed by using paper pencil tests (Lunz & Bergstrom, 1991; National Council State Boards of Nursing, 1991).

The CAT administration, scoring, and reporting costs may be smaller than non-adaptive tests for previously established testing programs; however, developing a large enough item bank

to cover a range of abilities with content specifications and to ensure overall bank security with exposure control mechanisms, and calibrating item parameters for newly-developed items by piloting on large number of examinees may be expensive at the beginning (Gershon, 2005). In addition, computer hardware that can handle complicated item selection scoring algorithms, and calculations of estimate ability may be costly (Luecht & Sireci, 2011). For example, Delaware had to allocate substantial funds to buy additional servers, redesign training for teachers who would be test administrators, and distribute netbooks to prepare schools to transition to CAT for statewide assessments (Davis, 2012).

One psychometric limitation discussed in the literature is that IRT and Rasch models are not applicable to all item types (Linacre, 2000; Rudner, 1998). CAT works well with multiple-choice items or one-word response questions, but experts have contradicting opinions about whether it works well with longer answers or essays (Davis, 2012). In addition, CAT may not be appropriate for all subject areas and skills (Rudner, 1998). It works efficiently in subject areas where content is measured using discrete questions that can be dichotomously scored as right or wrong; however, when questions are associated with a common stimulus or problem scenario, the questions' content may become confounded with how the stimulus or scenario is presented (Way et al., 2010).

Another concern raised by researchers is that CAT provides less control over the tests by psychometricians and content matter experts, as they are not able to review a test form before administration (Luecht & Sireci, 2011). The examinees' review of their answers in CAT is another issue examined by researchers. Despite examinees complaining about not being able to review their answers to previous items, some researchers argue that reviewing items may result in biased estimates (Wainer, 1993). For example, an examinee could fail initial items on purpose;

the CAT algorithm would estimate low ability and administer easy sets of items. The examinee could return previous items and change the responses and possibly get 100% of the items correct. This strategy may lead to the examinee's ability being scored at the highest level (Rudner, 1998). Therefore, reviewing previous answers may lead to upwardly biased scores for some examinees in CAT.

Another potential constraint of CAT composes the focus of this research: IPD due to factors that are not easy to detect, such as curriculum and practice changes. If the item parameter changes from its initial calibration, the accuracy of ability estimates can be compromised (Han & Guo, 2011). As the quality of a CAT depends on the quality of the item bank that the items are selected from, the effectiveness of CAT can be limited due to factors that lead to hard-to-detect changes in item parameters and item banks.

D. Item Parameter Drift

In this section, I will review available research on item parameter drift (IPD), factors that cause IPD, methods used for the detection of IPD, and the impact and consequences of IPD on measurement. One theoretical aspect that makes IRT and Rasch models useful for many psychometric data analyses is parameter invariance; that is, the equivalence of item and person parameters belong to different populations and measurement applications (Rupp & Zumbo, 2006). Yet, in practice, item parameters may not stay invariant. When an item functions differently for the examinees with same ability level, differential item functioning occurs (DIF; Holland & Wainer, 1993). When item parameters change over time, IPD occurs (Goldstein, 1983). More precisely, IPD is defined as differential change in the item parameters over subsequent time points or testing occasions (Wells, Subkoviak, & Serlin, 2002). Some researchers think of IPD as one form of DIF (Bock, Muraki, & Pfeiffenberger, 1988; Goldstein,

1983). In both of them, an item functions differentially on two or more sets of data. DIF examines if items behave the same in the examinee subgroups (e.g., race, grade level) whereas IPD examines whether items function same across testing occasions (Donoghue & Isham, 1998).

There are number of ways of observing IPD in operational items, depending on the IRT model being used. Under the Rasch model, item difficulty values may fluctuate over administrations with items becoming more difficult or easier over time. In 2PL and 3PL IRT models, item discrimination values may also vary over administrations with items becoming more or less discriminating. Although parameter drift is usually associated with IRT and Rasch models and applications, drift in item parameters may also be observed in *p*-values and point-biserial estimates under CTT framework. In addition, IPD may be observed in both fixed and adaptive test formats (Clark, 2013). Therefore, various aspects of IPD have been researched for the cases when items are repeatedly used, regardless of the test form or IRT model that is being utilized.

1. Why is Item Parameter Drift a Concern?

The presence of IPD may lead to a violation of one of the major assumptions of IRT and Rasch models—that examinees with the same ability or latent trait level have the same probability of answering an item correctly (Goldstein, 1983). Under the IRT and Rasch framework, an examinee's ability estimation is a function of item parameters. Therefore, ability estimates for examinees are expected to change if the item parameters change (Wells, Subkoviak, & Serlin, 2002). Failing to monitor this change can lead to inaccurate score calculations, comparisons, and misclassifications of examinees (Babcock & Albano, 2012). In large-scale assessments, IPD presence may create a disadvantage for individual examinees and jeopardize any conclusions about trends (Donoghue & Isham, 1998).

Testing programs usually employ common item equating methods in order to ensure a stable scale across time points and testing administrations. In common item equating, examinees take tests that include linking items placed to all forms as well as non-linking items specific to each form. The item parameter estimates for these linking items are assumed as invariant after their first use; however, these parameters may shift over time (Babcock & Albano, 2012). Failing to address IPD in linking items weakens linking across assessments and test administrations and undermines the comparability of scores (Donoghue & Isham, 1998). IPD threats measurement applications that need to maintain a stable scale using linking items (Wells, Subkoviak, & Serlin, 2002). IPD may also create a significant threat to the fairness and validity of score interpretations, particularly for the assessments used for monitoring change over time (Han, Wells, & Sireci, 2012).

2. Factors that influence item parameter drift

In practice, item parameter estimates will not be invariant; they may be expected to fluctuate over time (Kingsbury & Wise, 2011). Estimates may vary due to a number of random and systematic changes that have been researched fairly extensively in the past. Random sources such as sampling fluctuation or measurement error can result in changes in parameters (Swaminathan & Gifford, 1983). IPD can also be attributed to some systematic changes that explain the differences in item parameters over time. Shifts in item parameter estimates as a result of random error should not be ignored only because they are random. Yet, the amount of random error can be reduced to an acceptable level by using a greater number of common items (Huff & Hambleton, 2001). On the other side, systemic shifts in item parameters can be controlled by identifying the factors leading to them (Taherbhai & Seo, 2013). These factors are similar in adaptive and fixed-item tests and will be presented in this section.

One important source of IPD frequently cited in the IPD literature is curriculum or instruction changes (Bock, Muraki, & Pfeiffenberger, 1988; Goldstein, 1983; Han & Guo, 2011; Mislevy, 1982). When a change is made to content standards or when a new curriculum is adopted, the content of certain items may receive more or less emphasis, leading to changes in parameter estimates over time. For example, a study by Mislevy (1982) analyzed the impact of curriculum changes on fourth grade science items designed to measure metric system conversion. As teachers started to spend more time on teaching the metric system and less time on teaching the American system due to the progress of metrification, science items that used the metric system became easier and items that used the American system became harder. In a similar study, Bock, Muraki, and Pfeiffenberger (1988) examined drift on the College Board Physics Achievement Test and revealed that 10 items became differentially harder while 11 items became differentially easier over a 10-year period. The change in item parameters was attributed to the change in the emphasis of physics curricula of American secondary schools over time. Sykes and Fitzpatrick (1992) analyzed the invariance of item difficulty estimates for 285 items used for a licensure test administered over a five- year period. They detected directional drift in the items, where the items became more difficult over time. The drift source was associated with curricular emphasis changes. A similar study by DeMars (2004b) analyzed item parameter drift of items in the areas of information literacy and global issues. The study revealed that information literacy items displayed a greater amount of drift due to the change in the content area over time.

IPD can also be related to the characteristics of items such as content, skills, or knowledge that the item assesses. In a study by Chan, Drasgow, and Sawin (1999), item parameter stability of the Armed Services Vocational Aptitude Battery (ASVAB) was studied

over a period of 16 years. The differences in the item parameter changes were attributed to the content the items assessed. For example, the items requiring more content-specific knowledge were found to be more susceptible to drift than those assessing general principles and skills. Similarly, Bock, Muraki, and Pfeiffenberger (1988) hypothesized one factor contributing to IPD is the content the items assess may be covered more actively in the mainstream media. People may become more knowledgeable that which is emphasized in the mainstream media, causing items containing this content to become less challenging over time.

Factors related to test security, such as developing and training test-wise strategies and item over-exposure, may also result in parameter drift. Developing test-wise strategies (e.g., teaching to the test, taking a review course, or cheating behavior) can cause changes in item parameter estimates over time. Examinees may develop skills to select the correct answer despite lacking sufficient content knowledge (Clark, 2013). In a study by Bergstrom, Stahl, and Netzky (2001), a national certification exam was analyzed for drift and some items drifting more than 0.5 logits displayed evidence of being taught in a review course.

Item overexposure is another cause of IPD. This can happen in adaptive tests where the same items are used from an item bank or in fixed-item tests where the same form is used over multiple administrations. In CAT, pretest items are administered to a predetermined number of examinees to calibrate item parameters. When too many examinees see the item in its pretesting phase, the item becomes over-exposed before it is used as an operational item. If too many examinees see an active item, it may also be over-exposed (Bergstrom, Stahl, & Netzky, 2001). Exposed items are likely to become less difficult for later examinees who have specific or prior knowledge about the item, that is observed in the decrease of the item difficulty over time (Bock, Muraki, & Pfeiffenberger, 1988).

Another cause of IPD is due to construct-irrelevant changes made to the item between administrations. This includes changes in item position on the form, scoring of the item, mode of presentation, changes in the position of the picture relative to the text, and formatting (Clark, 2013; Taherbhai & Seo, 2013). Researchers indicated that changes in item positions and arrangements between administrations negatively affect stability of item difficulties and classification of examinees into achievement levels (Meyers, Murphy, Goodman, & Turhan, 2012; Whitely & Dawis, 1976; Yen, 1980).

3. Detection of item parameter drift

A number of techniques have been developed to identify IPD. Some of these techniques operate on fixed-item tests, while many of them were developed to detect IPD in CATs. The main reason for the distinction between procedures is the range restriction of ability measures in CAT as opposed to fixed-item tests. When an item is calibrated based on CAT data, the inflection point of item characteristic curves (ICC) is estimated from a small range of abilities as CAT is targeted to ability; however, in a fixed-item form, a broader sample is used to look at differences between ICCs (J. Stahl, personal communication, January 16, 2015). Therefore, I will explain the IPD detection methods in fixed and adaptive test formats separately.

a. Detection of item parameter drift in fixed-item tests

Some researchers used IRT models to detect IPD in fixed-item tests by examining whether the invariance of item parameters holds across time points. Their approach included comparing the parameters from two or more administrations to determine whether the same IRT model fits across multiple administrations. Bock, Muraki, and Pfeiffenberger (1988) analyzed an IPD detection approach for monitoring and updating an IRT scale on the College Board Physics and English Achievement Test. Using the 3PL IRT model with a time dependent model, in

which item parameters and the parameter trends are estimated at the same time, they evaluated the stability of item parameters. They adopted analysis of variance (ANOVA) to assess two-way interactions between items and testing occasions over a period of ten years. The two-way ANOVA (items * year groups) results indicated a significant IPD occurred over time among the Physics items but not among the English items. In addition, IPD had more impact on the item difficulty parameters compared to the slope and discrimination parameters in both content areas.

Cook, Eignor, and Taft (1988) examined the presence of IPD using the 3PL IRT model on two forms of a biology test from three different time points. They assessed the effect of recent instruction on the item parameter estimates by giving the old and the new forms of the test in two consecutive fall semesters to examinees who had not taken biology for at least one year and the new form in spring to another group of examinees who had recently taken a biology course. Their results from both the IRT and CTT analyses showed item parameter estimate instability for fall and spring administrations, despite stable estimates from two forms administered in fall, indicating that item difficulty estimates differed depending on the time of the administration. They concluded the biology instruction proximity to testing may have impacted item performance and resulted in drift.

Stone and Lane (1991) investigated the IPD impact by employing a model testing strategy using the 2PL IRT model. They examined item parameter estimate stability of 19 items in a math test between two testing occasions. They compared two models (restricted and non-restricted model), using difference chi-square statistics in order to test the parameter invariance. In the restricted model, the item difficulty and discrimination parameters were constrained to be equal across groups. Using a chi-square likelihood ratio test, they examined the difference between the two models. Significant chi-square statistics indicated that the unrestricted model

provided a better model fit. They also found that while some item parameters varied, the majority of item parameters were invariant over time.

In another study, Sykes and Ito (1993) examined the impact of IPD on equating in two licensure exams of related health-care professions. They employed analysis of covariance (ANCOVA) to examine shifts in item difficulties and how these shifts were related with some variables (e.g., content domain) and item position that may affect stability of item parameters. The item difficulty estimations were obtained using the Rasch model over an 8-year period. They investigated IPD by comparing differences between mean item difficulty pairs for each item pool over time for the two exams, as well as the reason, magnitude, and effect these differences had on pass-fail rates. Their results showed there were not substantial differences in item difficulty pairs between item positions across the two banks which indicated non-significant impact of item position on stability of difficulty parameters; however, they detected substantial systematic changes in item difficulties for pairs of items as a function of time within the mean item bank difficulties for both exams. That result indicated significant item bank drift over time.

IPD is frequently viewed as one form of DIF and some researchers employ commonly used DIF methods to detect IPD. Donoghue and Isham (1998) examined drift in items across two occasions using DIF methods. They simulated dichotomous data according to the 3PL IRT model, which has three parameters: discrimination (*a*), difficulty (*b*), and guessing (*c*). In the study, one-third of items had *a* parameter drifting only, one-third of items had *b* parameter drifting only, and the one-third of the items had both parameters drifted. They compared the three main types of measures to detect drift: (a) IRT-based, (b) Mantel-Haenszel-based, and (c) Chi-square based DIF detection methods. Across the 12 different methods used, Lord's chi-squared measure was found to be the most effective to detect drifting items. The next most

effective measures to detect drift were Raju's exact unsigned interval, the BILOG/PARSCALE chi-square statistics, and Kim and Cohen's closed-interval signed-area measure. The Mantel-Haenszel measures also functioned well, but displayed low correlations with other drift measures. Overall, they noted that the measures that rely on a common *c* parameter estimate resulted more accurate IPD detection than did measures that rely on separate *c* parameter estimates for both of the two occasions. They were cautious about generalizing their findings to various amounts of IPD, as it looks unrealistic to expect that items will always exhibit uniform drift.

IRT-based DIF detection methods examining the item characteristic curves (ICC) and test characteristic curves (TCC) have also been used to identify IPD by providing visual representation of changes in parameter estimates over time. For example, Chan, Drasgow, and Sawin (1999) examined the time effect on psychometric properties of items from the ASVAB across five different time points over a period of 16 years. They analyzed 200 items from eight subtests by plotting ICCs and TCCs to determine if the item parameter estimates and tests changed substantially over time. Their results indicated that of the selected 200 items, 25 (12.5%) items displayed significant shift in difficulty estimates. Although some subtests showed differential test functioning (DTF) over time, the effect sizes were not large and DTF was eliminated with the removal of a few items. They concluded that time had an effect on the psychometric properties of both psychological items and tests.

Wollack, Cohen, and Wells (2003) examined the effect of removing speeded examinees, the examinees for whom the test was overly speeded, on the stability of score scale, using college-level placement testing program data across an 11-year time span (1990-2000). The test was an 80-minute test of grammar and reading with different test forms published each year.

Each form used the same overlapping items, creating a link across the forms. Bolt, Cohen, and Wollack's (2001) IRT mixture model was used to identify the speeded examinees. Prior to equating earlier forms with the most recent form in 2000, they analyzed if the common items' difficulty estimates had shifted from their anchored values. This analysis was conducted using iterative linking through the TCC method (Stocking & Lord, 1983) with Lord's chi-square statistics (Lord, 1980). While items from the non-speeded group provided better score scale stability than use of the speeded group, on average 67% and 74% of items, respectively, were flagged as drifting.

Similar to the DIF statistics, some researchers used statistical programs and software to calculate statistics and obtain values for parameter drift. DeMars (2004a; 2004b) used both a multiple-group DIF detection method and statistical programs to detect IPD when examining drift over multiple test administrations compare their effectiveness. The methods he used were (a) BILOG-MG, the computer program which analyzes 2PL and 3PL IRT models to examine changes in item difficulty over time, (b) CUSUM (cumulative sum of standardized differences) statistic to detect changes in item difficulty or discrimination over time, and (c) the modified KPC approach, a multiple-group DIF detection method using linear contrasts of the discrimination and difficulty parameters. Using the 3PL IRT model, he simulated data composed of 100 items, where 10 items exhibited drift in varying degrees and direction across five time points. The drift conditions were linear, uneven, and sudden changes in parameters across these time points. The results showed that BILOG-MG and the modified KPC effectively detected IPD, but falsely identified non-drifting items as showing IPD in acceptable error rates. The CUSUM procedure did not function as effectively as the other two procedures in detecting IPD in item difficulty and discrimination parameters.

The WINSTEPS (Linacre, 2005) program provides displacement statistics to detect and monitor IPD. The displacement statistics is defined as the logit difference between an anchored value and the value for the difficulty that would be obtained if the parameter was freely estimated, all else being held constant (Linacre, 2005). Stahl and Muckle (2007) conducted a study to determine whether displacement statistics identify drift or if they contain statistical artifact. Statistical artifact results from the way displacement statistics are calculated and can have a major influence on the usefulness of the interpretation usefulness based on this statistic. Using simulations, they replicated certain conditions of IPD to assess their impact on the interpretation of results based on the displacement statistics. Sample sizes of persons and items were selected to be reflective of typical sample sizes in the certification and licensure field. In 81 unique conditions, including nine candidate/item size combinations, three different percentages of drift items and three drift directions were simulated using the Promissor simulator (Becker, 2013). Results indicated that situations where the degree of drift is symmetrically distributed in both easier and harder directions, the artifact impact on the displacement statistics was minor; however, as the drift becomes more asymmetrically distributed, artifact impact became more noticeable and resulted in non-drifting items being flagged as significantly drifting.

The robust z-statistics (z_R) method has been widely used with Rasch, 3PL and two parameter partial credit (2PPC) models in order to detect items that have unstable item difficulty estimates in linking/equating. Huynh and Meyer (2010) applied the robust z-statistics method to the 3PL and 2PCC IRT models to determine whether drift is present for either the difficulty or discrimination parameter estimates. Using a large-scale state assessment data for 5th grade students, they demonstrated the usefulness of the robust z-statistics in detecting unstable items with 3PL binary items and 2PPC partial credit items. The authors noted that their observation

only applies to the data at hand and should be investigated further on different data sets and situations.

Wells, Hambleton, Kirkpatrick, and Meng (2014) compared two procedures for flagging consequential IPD in an operational testing program. In the first procedure, they flagged items that exhibited a substantive magnitude of drift based on a critical value that was designated to indicate barely acceptable level of IPD. In the second procedure, they used D^2 statistics to flag IPD items. The items with D^2 statistics more than two standard deviations from the mean were flagged for IPD. They implemented both methods using an iterative purification approach to detect IPD. Using simulations, they compared the effectiveness of both procedures in flagging consequential IPD. Their results revealed both procedures effectively identified consequential drift and the iterative purification approach yielded meaningful information about the consequences of keeping or removing a flagged item.

b. <u>Detection of item parameter drift in computer adaptive tests</u>

The detection of IPD and addressing its effect on examinee decisions examinees become more important with CATs. In CATs, examinees are not always exposed to the same sets of items and the integrity of an individual examinee's score is dependent upon the integrity of the calibrations of all items in the pool. Therefore, researchers have developed and used techniques to investigate IPD in CATs.

Veerkamp and Glas (2000) used a statistical quality control method based on the CUSUM charts to detect disclosed items that may result in parameter drift. The method uses cumulated deviations of item parameter estimates from the original values to test drift under 3PL and 1PL IRT models. The method was implemented on a simulated adaptive test example and a power study was conducted. In the adaptive test design, examinee proficiency parameters were

drawn from a normal distribution with mean 0.2 and variance 1. The responses were generated according to the 1PL IRT model for items assumed to be unknown and according to 3PL IRT model for items assumed to be previously known. Parameter drift was imposed in six conditions; in the first three items were known to 5%, 10%, and 20% of the respondents respectively and in the next three conditions parameters changed from the initial calibration by -.20, -.40, and -.60 logits, respectively. The results showed that procedure's detection rate was acceptable with a well-controlled type I error rate. The effect size *k* was defined as a constant reference value indicating the size of the standardized shift in item difficulty. The smaller effect sizes, *k*=0.50 and *k*=1.00, and larger model violations resulted in the highest percentages of drift detection. The best detection was observed using the combination *k*=1.00 and a shift in difficulty of -.60, with an almost perfect detection rate of 99%. The worst observed performances were with combinations of effect size *k*= .5 and *k*= 2.00, with small model violations, such as item disclosure to 5% or 10% of the respondents, or a shift in difficulty of -.20 logits. They added that parameter change significance should also be checked using some other statistical tests after IPD detection.

Bergstrom, Stahl, and Netzky (2001) analyzed factors influencing item parameter drift in a CAT environment. Their purpose was to control the trade-off between sample size and exposure rate and examine how this trade-off impacts IPD detection. Their data came from an adaptive national certification exam with 1,000 examinees and 70-140 items per test. Prior to the first exam administration, an item bank was developed using data from paper-pencil administrations and subsequent item banks were equated to the benchmark scale using common item equating (Lunz & Bergstrom, 1991). The tests were composed from one of four item banks, which were calibrated with the Rasch model. The items could be present in one or more tests.

Each item was followed through its life cycle from Bank 1 to Bank 4, and six phases in an item's use were studied: pretest to pretest, pretest to active, active to active, active to pretest, pretest to retired, and active to retired. Three methods were used to detect drift:

1. Mean-centered difference to indicate magnitude of drift: Calibrations from pairs of items that appeared in more than one bank were compared to assess whether items calibration had significantly changed.
2. Standardized difference to assess how substantial the change in item parameter estimation is given the standard error of the estimation.
3. Cumulative sum of standardized differences to assess trends in IPD that a single comparison would not have provided over the four banks.

The results revealed that smaller drift was detected in cases in which less items were pretested and in which one year passed after the first calibration. The standardized difference of item difficulties increased from pretest phase to active phase as a result of the item's exposure to a more appropriate sample. In addition, items that were exposed to too few examinees were not flagged as drifting while items that were seen by too many examinees were detected as drifting although the logit change was minor. Finally, they found that the number of examinees exposed to items impacts drift detection. Thus, item bank size, number of examinees, and item exposure impacted the IPD detection.

In another study, Han (2003) examined IPD in CAT by using a procedure called "moving averages. He defined moving average as a form of average that takes into account both systematic and random components of time series data. Item performance can be monitored over time and any changes can be used to identify potentially drifting items by comparing plots of item *p*-values for test takers within time intervals. Although he successfully identified drifting

items in this study, the "moving averages" approach is not always applicable to every case in practice as it assumes comparable examinee populations across time intervals.

In a later study, Hatfield and Nhouyvanisvong (2005) investigated IPD presence within item banks for two high-stakes licensure exams of registered and practical nurses. They employed a repeated measures hierarchical linear model (HLM) approach to analyze the degree to which IPD existed in a set of anchor items. The level-1 model composed of repeated observations nested within the items (level-2 model). Using 440 anchor items appearing in three or more examination years over a 10-year period, they analyzed whether there was a general trend for item parameters to shift over the range of occasions and whether it is related to any specific item factor, such as content area and item exposure rates. Their results revealed no systematic increase or decrease in *b*-values, point-biserial correlations, or item response times over time. They noted that HLM can be used to detect IPD in item pools, but called for future research to test its utility.

In another study, Masters, Muckle, and Bontempo (2009) compared two approaches to recalibrate IPD items in an operational exam. Their data included 152 operational items and 450 examinees in which the items exhibited significant displacement in at least two of three operational item pools that were available. They compared the performances of "adjusted" operational item difficulty (that is, adding original anchor items to the displacement statistics), with the "fresh" item calibration (that is, re-calibrating an already anchored item in accounting for the drift). Their results revealed the adjusted and the fresh calibrations were highly correlated for detecting drifted items; however, a *t*-test between the two values showed statistically significant differences for 40 of the 152 items, yet, the actual difference was small. The findings yielded mixed evidence for using displacement statistics to adjust calibrations of drifting items.

Meng, Steinkamp, and Matthews-Lopez (2011) examined IPD in a fixed-length CAT using the 3PL IRT model to develop procedures for efficiently identifying drifting items. Their data included 1,921 operational items with 1,208 items in IPD and 15,000 examinees. They used two drift identification techniques for the 1,208 items: BILOG-MG item model-fit, Chi-Square and modified Chi-Square statistics. Their findings indicated that identifying IPD items in a CAT should not merely rely on Chi-Square statistics. They suggested using the Chi-Square statistics jointly with other indicators of ICC drift, that are the difference between expected probability and observed success percentage for each θ interval, absolute ICC drift, standard ICC drift, and lower and higher asymptote ICC drift to identify IPD.

Zhang (2014) developed a sequential monitoring procedure to detect compromised items in the item pool of a CAT system. He conducted two simulation studies; one was a type I error study assessing family-wise type I errors when no items were actually compromised and the other was a type II error (power) study, assessing the procedure's power to identify a compromised item. The data was composed of 400 items from a real large-scale assessment. Items were calibrated using the 3PL IRT model and there were 10,000 examinees in the simulations. The true ability parameters were generated from the standard normal distribution, N (0, 1). The CAT was considered as a fixed-length test with 40 items. Results from the first simulation study showed the procedure could control type I errors by choosing an appropriate cutoff point at any reasonable significance level. Results from the second simulation revealed the procedure's power was high when the items compromise at the middle stage of CAT; however, the power was found to be low if an item was compromised when the CAT just started, which the researcher indicated as a rare situation. Therefore, under the simulated conditions this procedure could control type I errors with a very low rate of type II errors.

4. Impact of item parameter drift

Potential negative outcomes associated with IPD presence led researchers to investigate the impact and consequences of IPD on both fixed-item tests and CAT. They examined how IPD presence influenced scoring of exams, pass-fail decisions, diagnosis of mastery of skills, and equating and linking processes. In this section, I will explain available research regarding the impact of IPD with varying conditions and different IRT models for fixed-item tests and CATs separately.

a. Impact and consequences of item parameter drift in fixed-item tests

Sykes and Ito (1993) analyzed the impact of systematic, non-zero shifts between pairs of item difficulties on previous exam cut scores and the proportion of candidates that passed based on these cut scores. They used licensure examination data from two separate item banks with 382 items and 487 items, respectively. Their results revealed that item difficulty values changed systematically as a function of time for both exams. Therefore, item bank drift was present. The differences between the actual form cut scores and the cut scores adjusted for bank drift changed within a range from one to five score points for both exams. The proportion of candidates who passed was unstable over the eight years of test administration. Their results indicated that IPD can impact the validity of decisions made based on cut scores affected by drift.

In another study, Wells, Subkoviak, and Serlin (2002) investigated impact of IPD on examinee ability estimates. Using the 2PL IRT model, they simulated two testing occasions, one with 40 items and 300 examinees, and the other with 80 items and 1,000 examinees. Item parameter estimates were generated using the University of Wisconsin-Madison Math Placement exam and English Placement test. Examinee ability estimates were sampled from a normal

distribution, *N* (0, 1). Forty-eight drift conditions were created: test length (2) x sample size (2) x type of drift (*a*, *b*, both *a* and *b*) x percentage of drift (5, 10, 15, 20%). Drift items were randomly selected and the magnitude of drift was determined based on a large-scale standardized test. Unidirectional drift was presented to alleviate cancellation of positively and negatively drifting items. The drift items were not included in the set of items used to link two occasions. Root mean-squared error (RMSE) was used to determine how well the item parameters were recovered, and root mean-squared differences (RMSD) were used to evaluate the influence of drift on ability estimates. The differences between ability estimation values, obtained based on item parameter estimates that included drift items and those that did not include drift items were examined. Their findings revealed that under the studied conditions, IPD has a minor effect on the ability estimates between the two simulated testing occasions; however, they noted that the percentage of drift items, the magnitude of drift, and the effect of drift may be greater for testing occasions that were more than one year apart, such as CAT item banks being used for many years.

Witt, Stahl, Bergstrom, and Muckle (2003) studied the impact of IPD on Rasch ability estimates and pass-fail decisions. They used non-normal distributions of examinee abilities and item difficulties to represent true parameters of many assessment situations, particularly in the context of licensure and certification testing. Then, they simulated response data for eighteen IPD conditions based on two test scenarios, one with 100 items and 187 examinees and the other with 200 items and 260 examinees. The IPD impact on ability measures was evaluated according to correlation values between true and estimated measures and the IPD impact on pass-fail rates was evaluated by comparing the number of false positive and false negative decisions in the baseline condition and drift conditions. Under the baseline condition, their results suggested that

the correlation values between estimated and true abilities changed within a range from .81 to .92 for the 100-item test and from .95 to .97 for the 200-item test. Similarly, the correlation values for the IPD conditions for the scenario with 100 items ranged from .85 to .94 and ranged from .96 to .97 for the scenario with 200 items. Across all eighteen simulation conditions, the number of misclassifications stayed within an acceptable range that could be observed as a result of measurement error solely. For example, only four of 187 examinees were misclassified outside the 95% CI in the scenario with 100 items, and seven of 260 examinees were misclassified outside the 95% CI for the scenario with 200 items. In all drift conditions, false negative classifications were more common than false positive classifications for both scenarios. The study provided evidence that the Rasch model is robust to estimate ability under undetected drift conditions even when item and person parameters are not normally distributed.

Wollack, Sung, and Kang (2006) examined the longitudinal effects of IPD and the linking model on examinee ability using empirical data from a college-level German placement test over a 7-year period. Examinees taking the test were administered approximately 55 items each year; 30 to 53 items being scored and the remaining items pilot-tested. The researchers calibrated item parameters for the data from form 90X under the 3PL IRT model and linked the remaining forms to the 90X metric using ten different methods. To assess the impact of IPD and linking model on ability estimation, ability estimates from form 90X under the ten linking models were compared using correlations, mean differences, and RMSD between ability estimates for all pairs of models. Additionally, true score functions and passing rate functions were analyzed. Their results revealed that the ten methods displayed differences in terms of the impact on ability estimations and passing rates that are not practically important. This may indicate that IRT is sufficiently robust to identify IPD; however, they noted that it is impossible

to conclude which of the ten methods is most robust to identify IPD without being tested under different types, magnitudes, and amounts of IPD.

IPD has often been characterized as a lack of parameter invariance at the item level, and effects of lack of parameter invariance on parameter estimation were investigated by researchers. Rupp and Zumbo (2006) investigated the impact of lack of parameter invariance in unidimensional IRT models on examinee ability. Their visual, numerical, and analytical illustrations of the results revealed that effects of IPD on examinee response and true scores depend on the magnitude of drift. In addition, they noted that across various theoretical conditions, IRT model inferences about ability parameters are robust toward small-to-medium amounts of lack of invariance.

In another study, Meyers, Miller, and Way (2009) investigated the impact of item difficulty parameter change as a function of item position change on IRT-based common item equating. Using large-scale K-12 program data, they modeled the Rasch item difficulty parameter shifts from field-test stage to operational test stage as a function of differences in position of items in the test, test level, content, and item format. Then, they used a series of simulations to examine the impact of item position change on equating. They modeled the change in Rasch item difficulties using multiple regression for each grade from 3rd to 8th for each subject area (reading and math) and then across grades, using item position as a predictor. The regression analysis revealed that 56% of the variance in change of Rasch item difficulty in math and 73% of the variance in Rasch item difficulty in reading could be explained by the change in item position. In addition, they found that Rasch item difficulties were impacted by item position change from field testing to operational testing. Yet, the observed effects due to position change were mitigated as the items were ordered by difficulty, with easier items placed at the beginning

and through the end of the test and harder items placed in the middle of the test. Then, they used 281 field-test math items from grade 5 to further examine effects of item position change in equating. Their simulation results provided evidence that the effects of item position change on item difficulty were canceled out due to item ordering by the testing program. The other simulation conditions, however, where items were ordered from easier to harder indicated measurable effects of item position change on item difficulty and equating. Hard items became harder as they were seen toward the end of the test, and easier items became easier as they were seen at the beginning of the test. Therefore, the operational test seemed harder than it actually was for the examinees with higher abilities and easier than it actually was for the examinees with lower abilities.

Kingsbury and Wise (2011) examined stability of scales measuring growth of individual students across grades and change of groups across years in a K-12 setting. In addition, they investigated to what extent changes in item difficulty estimates influence student scores. They used reading and math scales developed by the Northwestern Evaluation Association for their inquiry. These scales were constructed based on the Rasch model and used to develop achievement tests for various school districts. The analyses were composed of two parts: scale drift analysis and impact analysis. In scale drift analysis, they calculated correlations between the new and original item difficulty estimates and compared them with the values obtained from other studies using the same scale. They also calculated bias and mean absolute difference values and compared these values to standard deviations of student performance to assess IPD impact. Finally, they analyzed item parameter estimate differences as a function of the original data calibrated initially in order to assess whether time between two calibrations influenced the amount of drift. In the impact analysis, two representative test forms of students were used to

create tables showing conversion from raw score to Rasch unit score, one relying on original parameter estimates and the other relying on the new item parameter estimates. Then, they compared the scale scores obtained from the two scoring tables to examine to what extent a particular student's test score would have changed due to scale drift. Their findings revealed significant correlations between original difficulty estimates and new difficulty estimates. The correlation value was .967 for math scale and .976 for reading scale. The mean shift associated with item difficulty estimates was obtained as -.11 for math scale and .17 for reading scale. No significant drift was present in the scale values and the new estimated difficulty values did not present any systematic shift as a function of time. The impact analysis showed that the largest change in student scores from original calibration to the new calibration was 1.1 point for both scales and 99% of the observed changes were smaller than 1.00 point. Thus, IPD impact on student scores was not substantial in both scales.

In another study, Han, Wells, and Sireci (2012) analyzed the impact of multidirectional IPD (e.g., some items drifted in harder direction while other items drifted in easier direction) on the linking procedure and proficiency estimates. In a series of simulations, they investigated the impact of varying linking item combinations with multidirectional IPD on the scaling coefficients, rescaled item parameter estimates, and equated proficiency estimates when linking scales between two test administrations. They mimicked an actual state-wide, large-scale assessment using 40 items and 10 different test forms. The forms varied with respect to pilot items and items that were linking the current 40-item test form to the base form. Proficiency parameter estimates for 5,000 simulees were drawn from two standard normal distributions: N (0, 1) for year 1 and N (0.1, 1) for year 2. They simulated conditions based on four crossed factors: two anchor test lengths, four patterns of multidirectional IPD, three magnitudes of IPD,

and three scaling methods. Their results revealed that choice of linking method and pattern of IPD has substantial effect on the examinee scores and classification results under some of the studied conditions. Even though the effect of IPD was mostly cancelled out by balancing multidirectional IPD for some conditions, the authors suggest not to just presume the effect of multidirectional IPD will be cancelled out by balancing IPD items. Instead, practitioners should carefully examine whether the IPD pattern is likely to deteriorate the scaling/equating processes.

Another study on equating was conducted by Jurich, DeMars, and Goodman (2012). Under various simulation conditions, they assessed recovery of equated scores and scaling constants for the Stocking-Lord IRT scaling method. The simulation mimicked a situation in which two administrations were given at different testing occasions using two forms. In the first administration with the original form, none of the items were compromised. The second form, where common items used under the non-equivalent groups with anchor test (NEAT) equating, included compromised anchor items. Thus, only anchor items on the second form were likely to be involved a possible cheating situation. Both forms had 100 items and the probability of correct response on each item was generated for 3,000 simulees using the 3PL IRT model. Four different conditions were studied to create multiple amounts of cheating: (a) the percentage of compromised anchor items (25% and 100%), (b) the percentage of examinees who access to compromised items (5%, 10% ,25%, 50%), (c) ability distributions (M:0, SD:1, M: -.5, SD:1, M:0, SD:1, M: -.5, SD:1.25), and (d) anchoring methods (external and internal scoring to the test). They created the cheating situation by adding .5 to the probability of giving a correct answer to a compromised item for cheaters. To establish equivalence between the two forms, each examinee's number-correct score on the second form was converted to an equivalent score on the original form. They found that an increase in the proportion of compromised anchor items

and examinees with access to these items resulted in positively biased equated scores. The most extreme cheating condition where 100% of the items compromised items for 50% of the examinees resulted in large positively biased equated scores. When the degree of cheating was substantial, compromised items had less impact on the equated scores' accuracy under the external anchor condition as opposed to the internal anchor condition. Even at moderate degrees of cheating, the extent of bias was relatively large, indicating that equated scores calculated based on slightly compromised test forms overestimate examinees' ability. Thus, considering compromised items as a form of IPD, they pose a threat to ability estimations and equating processes.

b. Impact and consequences of item parameter drift in computer adaptive tests

The potential impact of IPD in CAT can be more detrimental than in fixed-item tests due to the restriction of range of the ability measures of the examinees in CAT (J. Stahl, personal communication, January 16, 2015), as IPD may directly influence score estimations of examinees at the individual level. Additionally, it may lead to serious problems at the testing program level as the test scores are also used to calibrate newly-added pretest items (Han & Guo, 2011). Despite these potential threats, there are only a handful of relevant studies in the literature focusing on impact and consequences of IPD in a CAT environment. In this section, I will explain these particular studies examining impact of IPD on ability estimates, pass-fail and classification decisions.

Guo and Wang (2005) examined scale drift in CAT using GMAT® Quantitative measure data collected online. Addressing the need for research evaluating scale drift in CAT, they collected two sets of data including 31 items; Time Point One (T1) and Time Point Two (T2).

Two sets of item parameters from these two time points were calibrated and scaled. They quantified the differences between the two sets of item parameter estimates and their ten simulations by a modified root mean squared differences (RMSD) on test characteristic curves (TCC). They established an empirical baseline condition for random variations as a result of calibration and scaling over time via ten simulations. Then, they compared the baseline condition with the differences between T1 and T2 item parameters to determine whether scale drift was present. Their results showed that scale drift was not present in the GMAT® Quantitative measure and the observed differences between T1 and T2 item parameters occurred due to random variations. Yet, they called for further research with more simulations (at least 100) and simulation parameters (e.g., IRT model, sample size, calibration and scaling methods) that are identical to the real pretest calibrations for a more realistic evaluation of scale drift in CAT.

As mentioned previously, some researchers saw compromised items as one type of IPD. For example, Guo (2009) used a simulation method to quantify the impact of compromised items on the GMAT® CAT. Five thousand examinees were randomly selected from the GMAT® administration in January 2007 and their ability estimates in quantitative and verbal sections of the exam were used as true ability estimates. The same items used in the January 2007 test were used as the items with scaled items parameters in the study. He implemented a "two-path simulation" approach in which each examinee received two scores; one from the first path CAT without compromised items and another from the second path CAT involving five compromised items per section. The two paths yielded two scaled scores for each simulee, converted from their two ability estimates. The impact of compromised items was measured as the difference between two scores of the simulees who had been exposed to one or more compromised items. His results revealed that about 95% of the simulees had no gain on their scores at all; about 0.5% or 0.6% of

the simulees lost 10 or more points; about 3% of the simulees improved their scores by 10 points; about 1% of the simulees improved their scores by 20 points; and about 0.5% of the simulees improved their scores by 30 or more points, on the total score scale ranging from 200 to 800. Considering the standard error of measurement (SEM) of GMAT® is 30 in total score, about half a percent of examinees gained one SEM or more. The impact of compromised items on the GMAT® verbal and quantitative scores showed the same patterns with the total score.

McCoy (2010) conducted a study using American Society for Clinical Pathology (ASCP) CAT data to analyze if IPD exists in the item pool and to assess IPD impact on examinee scores. His data was composed of 2,555 examinees and 1,270 items in eight content areas. He employed a hierarchical generalized linear model (HGLM) that utilizes an extended Rasch model controlling for drift by incorporating time as covariate for the item difficulties. In this study the model is identified as the Rasch Linear Model (RML). Using a Bayesian analysis, he then examined the impact of using the extended model on classification decisions for each examinee and each subscale in the exam. His results showed that RLM, as a new IPD detection method, provided more accurate results for identifying IPD than traditional methods, such as the displacement method used in the Rasch framework. He identified a minor presence of IPD, ranging from 2 to 8 items, in each subscale. Additionally, RLM provided more precise ability estimates because of the adjustment on the estimates when drift existed. His findings also suggested that pass-fail decisions could be altered when ability is estimated by taking into account IPD under RLM. After the alterations, changes in pass-fail rates were minor for all the subscales, with the largest change being 0.6%, corresponding to approximately 16 examinees out of the 2,555.

Deng and Melican (2010) evaluated scale drift in CAT in order to improve quality control and calibration process for ACCUPLACER®, a large-scale adaptive assessment of academic progress. Their data was based on the operational data from years 2004 to 2007 with 223 items and more than 800,000 examinees per year. They calibrated item and person parameters based on the 3PL IRT model for each year and transformed the item parameters to the baseline scale using mean/sigma transformation. They used Raju's nonconfirmatory differential item functioning (NCDIF) index to examine IPD at the item level. They computed NCDIF for each item using the 2004 sample as a reference group and other years as focal groups. Then, for each item, they compared NCDIF values for each of the 2005/2004, 2006/2004, and 2007/2004 comparisons to the cut-score obtained from the empirical distribution of the NCDIF to flag drift items. Their results showed that only two items out of 222 indicated IPD. One of these two items became easier in years 2006 and 2007. The other item became harder and more discriminating after 2004 and was flagged as drifting in all years except 2004. They noted that the application of NCDIF to examine IPD in CAT was reassuring with regard to practicability of these statistics for further simulation studies.

Wei (2013) examined the impact of IPD on examinee ability estimates in a computer adaptive multistage test (MST). The unit of test administration was a testlet instead of an item in the MST design. She used a 1x2x2 MST design. In this design, all examinees are administered the first testlet of medium difficulty at the first stage. Then, two testlets, one of medium difficulty and one of hard difficulty are administered at the second stage. An examinee is routed to one of these two depending on their performance of the first stage. Similarly, the third stage is composed of two testlets, with the selection determined by an examinee's performance on the previous two stages. The examinee response data for 1,000 examinees was generated under the

3PL IRT model to simulate two test forms of MST, one with 24 items and another with 75 items. Ninety-six different conditions with fully-crossed five design factors were used to simulate IPD: (a) test length: 24 items and 75 items, (b) test form difficulties: medium-medium-medium and medium-hard-hard, (c) percentage of IPD items: 5%, 10%, and 20%, (d) type of drift: both *a* and *b* increase, both *a* and *b* decrease, and (e) magnitude of drift: *b* parameter by 0.4/0.8 and *a* parameter by 0.3/0.6. RMSDs of the two sets of ability estimates (one set estimated using drift items, another set estimated without drift items) were used to evaluate the impact of IPD on person ability estimates. The results showed that IPD did not have a significant effect on ability estimates under most of the 96 conditions. Although the mean difference between the two most extreme sets of ability estimates (without drift and with 20% IPD items) were relatively larger than other conditions, the average RMSDs of these conditions did not exceed 0.30. In addition, the author suggested that these conditions with 20% of total IPD items represent the extreme case and are not likely to occur in practice. While the drift direction and percentage of drifting items appear to have detectable effects on ability estimates, test length and magnitude of drift had very slight impact on ability estimates. Despite non-significant results from this particular study, she called attention to the importance of the detection of IPD and minimization of its effects on ability estimates for large-scale, multistage adaptive tests.

More recently, Risk (2015) conducted a simulation study to investigate impact of IPD on estimation precision of scores, classification accuracy, and test efficiency. Using a series of CAT simulations, she manipulated several variables including IPD magnitude, IPD amount, and item pool size. The drift magnitude varied by three levels: 0.5, 0.75, and 1.00 logits. The drift amount also changed by three levels: 100, 75, and 50 IPD items in the item pool. She used three item pools with different sizes; a small item pool with 300 items, a medium item pool with 500 items,

and a large item pool with 1000 items. IPD items drifted in both directions (75% of the items drifted easier, 25% items drifted harder) and the direction was kept constant across all the conditions. The baseline CAT with no drift items was compared to the simulation conditions to evaluate impact using root mean square error (RMSE), absolute average difference (AAD), bias, item exposure rates, test lengths, and the total percentages of misclassification. Her findings suggested that drift magnitude had a greater impact on measurement precision than the number of drift items. The overall findings revealed that IPD does not substantially impact estimation precision or classifications accuracy and supported the robustness of CAT even under large amounts of IPD.

D. Summary of Literature and Proposed Study

A change in the interaction between an item and an examinee, IPD poses a serious threat to pre-test item calibrations, parameter estimates accuracy, and classification decisions made from test scores. Despite these threats, research on IPD impact is limited and resulting in discrepant findings. A summary of the previous study findings on IPD are presented in Tables 1 and 2 below. Few studies have found IPD had little effect on ability estimates and pass-fail classification decisions (Kingsbury & Wise, 2011; Wei, 2013; Wollack, Sung, & Kang, 2006); however, some researchers have demonstrated that IPD can have substantial effect on examinee scores and decisions made based on these scores (Han, Wells, & Sireci, 2012). Even researchers who found minimal IPD effects under the studied conditions called for further research to investigate IPD impact under various conditions (Guo & Wang, 2005).

Major IPD sources, such as differences in curriculum, instruction, school resources, and practice, create differential opportunities to learn and perform on a test. The effects of these differences are not easy to detect in real world test settings as their impact on person ability can

greatly vary across examinees; however, most simulation studies have assumed that IPD occurs equally across all examinees. As a result of changing interaction between items and examinee population, IPD realistically occurs only when drifting items interact with examinees whose abilities change due to various factors such as curriculum updates, historic events, poor teaching, and lack of practice and resources. One purpose of my research was to fill an existing gap in the literature by examining how IPD resulting from changes in examinees' content knowledge and skills can impact measurement precision and classification accuracy in CAT. The main goal of my research was to examine IPD impact when only a portion of examinees are affected by IPD. I offered recommendations to testing organizations and states that use CAT, particularly in K-12 settings, about how potential factors such as infrastructural, curricular, and instructional differences can be reflected in the ability estimations and may impact classification decisions; however, the study's implications are not solely limited to the education field. Certification and licensure programs using CAT might also benefit from these results, since they are also susceptible to such differences.

TABLE I

SUMMARY OF PREVIOUS STUDIES' FINDINGS ON IPD IN FIT

Sources of IPD in FIT	IPD detection methods in FIT	Impact of IPD in FIT
• Changes in curriculum or content (Bock, Muraki, & Pfeiffenberger, 1988; Goldstein, 1983; Mislevy, 1982). • Item characteristics such as content of items (Bock, Muraki, and Pfeiffenberger, 1988) • Item overexposure (Bock, Muraki, & Pfeiffenberger, 1988) • Construct irrelevant changes made to items such as position change (Meyers, Murphy, Goodman, & Turhan, 2012; Whitely & Dawis, 1976; Yen, 1980).	• By examining item parameter invariance in IRT models (Bock, Muraki,& Pfeiffenberger, 1988; Cook, Eignor, & Taft, 1988, Stone & Lane, 1991; Sykes & Ito, 1993). • DIF detection methods such as Mantel-Haenszel, Chi-square based methods (Donoghue & Isham, 1998; Chan, Drasgow, & Sawin, 1999). • Using statistics provided by software (DeMars, 2004a, 2004b; Stahl & Muckle, 2007).	• Classification decisions (Sykes & Ito, 1993; Witt et. all., 2003; Wollack, Sung, & Kang, 2006) • Ability estimation precision (Kingsbury & Wise, 2011; Rupp & Zumbo, 2006; Wells, Subkoviak, & Serlin, 2002; Witt et. all., 2003; Wollack, Sung, & Kang, 2006) • Equating (Jurich, DeMars, & Goodman, 2012; Meyers, Miller, & Way, 2009) • Linking (Han, Wells, & Sireci, 2012)

TABLE II

SUMMARY OF PREVIOUS STUDIES' FINDINGS ON IPD IN CAT

Factors cause IPD in CAT	IPD detection methods in CAT	Impact of IPD in CAT
• Changes in curriculum or content (Han & Guo, 2011). • Item characteristics such as content of items (Chan, Drasgow, & Sawin, 1999). • Test security, developing test-wise strategies, item overexposure (Bergstrom, Stahl, & Netzky, 2001). • Construct irrelevant changes made to items such as position change	• CUSUM: cumulative deviation of item parameter estimates (Bergstrom, Stahl, & Netzky, 2001; Veerkamp & Glas, 2006) • Hierarchical linear models where repeated observations nested within items (Hatfield & Nhouyvanisvong, 2005) • Using statistics provided by calibration software (Masters, Muckle,& Bontempo, 2009; Meng, Steinkamp,& Matthews-Lopez, 2011) • Graphical methods (Han, 2003) • Sequential type I and type II error testing (Zhang, 2014)	• Estimation of pre-test item parameters/ item bank maintenance (Deng & Melican, 2010; Guo & Wang, 2005) • Ability estimation/ score estimation precision (Guo, 2009; Risk, 2015; McCoy, 2010; Wei, 2013) • Classification decisions (Risk, 2015) • Test efficiency (Risk, 2015)

III. METHODS

The overarching research question is, "To what extent are person measure estimates and classification accuracy impacted when IPD exists in CAT?" The following particular questions were addressed:

1. What is the impact of IPD on person measure estimates when only a sub-group of people are affected by the drift?
2. What is the impact of IPD on classification of examinees to achievement levels when only a sub-group of people are affected by the drift?
3. Are the effects of IPD on person measure estimates and classification of examinees consistent across three factors of drift: proportion of IPD items in the pool, proportion of PPD examinees in the sample, and item pool targeting?
4. Does item pool targeting change the effect of IPD on low, medium and high ability PPD examinees' classification to achievement levels differently?
5. Holding all else constant, does the change in the impact of IPD over the levels of one factor depend on the level of another factor?

A. Chapter Overview

For this study, I performed a series of simulations to investigate the potential impact of IPD on examinee ability estimation and classification accuracy in a CAT exam when only a portion of examinees were affected by IPD. I employed a fully crossed design with a total of three independent variables for each condition. The independent variables were: (a) targeting of the item pool, (b) percentage of people affected by IPD in the population, and (c) percentage of a content area with IPD items in the item pool. Using several criteria, I evaluated various drift conditions and determined to what extent IPD influenced ability estimation precision and

examinee classification into achievement levels. I also examined if IPD impact on examinees with different ability levels varies, depending on the item pool targeting. As a supplemental analysis, I looked to determine if WINSTEPS person fit statistics are useful indicators for detecting misfit due to IPD.

My test design mimicked a statewide CAT assessment in terms of mean of the item difficulties, mean of the person ability, and cut scores for the achievement levels. Since I was unable to access test specifics of educational CAT programs at this stage of research, I used more general CAT specifications, which would likely be used on a statewide educational CAT assessment.

The modeled assessment was a norm-referenced measure of performance and growth across grades in K-12. This type of interim assessment helps educators determine where each student is performing in relation to local or state standards and national norms. One challenge while developing adaptive tests and maintaining item pools in the K-12 context are the differences in practice, curriculum, infrastructure, and instruction among schools (Kingsbury & Wise, 2011). Yet, the effects and consequences of this challenge on CAT have not been researched adequately. With this study, I will contribute to the literature by addressing the potential consequences of this challenge. In the following sections, I described specifics of the examinee population, the item pool properties, the simulation parameters, and the evaluation criteria.

B. Independent Variables

In this section, I describe the independent variables manipulated in the study. For ease of presentation, the independent variables and the levels are shown on Table 3 below. Specifically,

I address item pool targeting, content areas in the item pool, and examinees affected by drift in following section.

TABLE III

INDEPENDENT VARIABLES

Variable name	Item pool targeting	Percentage of content area in the item pool	Percentage of examinees affected by IPD
Levels	• Well-targeted item pool • Poorly targeted item pool	• 20% • 40% • 60%	• 20% • 30% • 40% • 50%

1. <u>Item pool targeting</u>

I simulated two item pools with different targeting to assess the impact of IPD on low, medium and high ability examinees. The first item pool was well-targeted to the examinee population with the mean item difficulty very close to the mean examinee ability. I specified the mean values and standard deviations based on the mirrored assessment's blueprint in order to design my study to be as realistic as possible. The second item pool was an off-targeted item pool in the easy direction. I set the mean item difficulty for this item pool one logit below the mean item difficulty of the well-targeted pool. The rationale behind setting one logit below was that the standard deviation of the examinee ability distribution was one. Therefore, generating an item pool with a mean one standard deviation below allowed me to simulate a poor targeting situation for the given examinee population. By simulating these two item banks with different targeting, I was able to consider the difficulty of IPD items that were exposed to PPD examinees and to evaluate the change in the impact of IPD when difficulty level of the IPD items changed. Specifically, when the item pool is well-targeted to the examinee population, the chance of an

examinee encountering an IPD item is determined by the distribution of the IPD items in the item pool. When the pool is off-targeted in easy direction, the likelihood of a high ability PPD examinee seeing an IPD item is expected to be reduced. On the other hand, low ability PPD examinees would likely encounter more IPD items since their ability values are closer to the difficulty values of the IPD items in an off-targeted pool. Therefore, the number of IPD items that is seen by PPD examinees with different abilities is expected to change by item pool targeting. For example, low ability examinees are expected to miss many more items in the poorly targeted pool. IPD in the harder direction can significantly affect the probability of these low ability examinees responding to IPD items correctly and may result in disadvantage particularly for them. Taking into account that CAT use could disadvantage examinees with less opportunity and motivation to learn, such as examinees from lower-performing schools, focusing on low ability examinees carried importance. Hence, with these two item pools, I aimed to take into account IPD item difficulty as a potential factor changing IPD impact on PPD examinees from different ability levels.

2. Percentage of the content area in the item pool

The percentage of content area with IPD items varied by three levels: 20%, 40%, and 60% of the item bank. Based on my simulation scenario, I assumed that due to poorer learning opportunities and lack of knowledge on this content area, disadvantaged PPD examinees had lower ability on the IPD items and the items drifted harder for these examinees. The difficulty parameter values for items in this content area differed in a range from -0.50 to -1.00 logits from their original values only for those PPD examinees. As stated by Han and Guo (2011) "the standard error of estimation for *b*-parameter usually ranges between 0.30 and 0.50, even without IPD" (p. 3). Therefore, it is important to study an acceptable and realistic IPD range.

Setting a range of difficulty shift between 0.5 and 1.00 logits produced varying effects of the factors that may result change in knowledge and skills (i.e. infrastructural, curricular, instructional, or practice differences) across the sample. This modification provided flexibility in simulating lack of knowledge that occurs due to differential opportunity to learn content.

3. Percentage of examinees affected by drift

The existing simulation studies exposed drift items to all examinees equally; however, the IPD effects that occur due to the factors such as curricular, instructional, infrastructural, and practice differences among schools and classrooms, are likely to vary across examinees. Thus, in this study, IPD was exposed to a partial group of examinees who did not have equal opportunity to learn the content with the other examinees. The percentage of PPD examinees in the sample varied by four levels. Here, 20%, 30%, 40%, and 50% of the sample contained PPD examinees who had less opportunity to learn and practice over certain topics in the assessment. There was no IPD for the remaining examinees.

C. Test Properties

In this section, I describe the simulated test properties including item pool, examinee distribution and achievement levels.

1. Item pool

The item difficulty distribution in the well-targeted item pool mirrored a state-wide CAT exam (M= 0.197 and SD=1.65) for the test. I generated the poorly targeted item pool by setting the mean difficulty one logit below the mean difficulty of the well-targeted item pool for the test (M= - 0.803 and SD=1.65). Note that one of the variables that was manipulated in this study is the proportion of content area that includes IPD items. There were two content areas in the exam; one with all IPD items and one with non-IPD items. The percentages of these

content areas in the pool varied in each IPD condition. The percentages of content areas in the exam mimicked the percentages in the item pool. Table 4 outlines the test specifications. As aligned with the mimicked assessment, I used a fixed-length CAT approach with 40 items drawn from the 1200-item pool.

TABLE IV

TEST PROPERTIES AND ITEM CHARACTERISTICS FOR THE ITEM POOLS

Item Pool	Min	Max	Mean	SD	Number of items in the CAT
Well-targeted	-3.2	3.4	0.197	1.65	40
Poor-targeted	-4.2	2.4	-0.803	1.65	40

2. Examinee distribution

The ability parameters (θ) for 500 examinees were drawn from N (0.3, 1) to obtain an approximate match with the well-targeted item pool difficulty distribution. The mean ability of the grade level was obtained from the modeled assessment. As size of districts and schools greatly vary by state, I could not define a "typical" sample size for the study in the context of the mimicked assessment. Hence, I used a suggested sample size for robust measures under a Rasch model framework. Linacre (1994) recommended that a sample size of 500 would provide robust item and person measure calibrations even under adverse circumstances such as poor targeting.

3. Achievement levels

I identified the achievement levels and cut scores using national grade level scores provided by the modeled assessment initially. The CAT assessment I used was a fixed-length 40-item test. This may have resulted in large standard error associated with the achievement level

classifications. Large standard error could even exceed width of the achievement levels specified in the modeled assessment. In order to eliminate this problem, I decided to collapse top and bottom categories and reduced the achievement levels to three levels; at the grade level, below grade level, and 1-grade below. Table 5 specifies the achievement levels and cut scores in the original assessment. Table 6 specifies the achievement level and cut scores in the simulated CAT.

TABLE V

ACHIEVEMENT LEVELS AND CUT SCORES IN THE ORIGINAL CAT

Achievement level	Cut score
Above grade level	≥ 0.4 logits
At grade level	≥ 0.2 logits to < 0.4 logits
Below grade level	≥ -0.7 logits to < 0.2 logits
1-grade below	≥ -2.1 logits to < -0.7 logits
2-grades below	< -2.1 logits

TABLE VI

ACHIEVEMENT LEVELS AND CUT SCORES IN THE STUDY

Achievement level	Cut score
At grade level	≥ 0.2 logits
Below grade level	≥ -0.7 logits to < 0.2 logits
1-grade below	< -0.7 logits

D. Computer Adaptive Test Simulation

Promissor® (Becker, 2013), a software simulating CAT administrations, was used for this study. The adaptive test specifications are provided below.

1. Ability estimate

The national grade level mean score was assigned as each examinee's ability estimate at the beginning of the test. After an examinee responded to an item, the simulator updated the estimated ability for a particular examinee based on unconditional maximum likelihood estimation (UCON) method. In this method, item and person parameters are estimated when the observed score for the parameters matches the expected score within a specified precision level (Wright & Panchapakesan, 1969). UCON is derived from the joint maximum likelihood estimation method (JMLE) where the likelihood function is updated based on the raw number-right score (Wright & Panchapakesan, 1969). As the name infers, it is an iterative, "joint" process where estimates of persons and items are obtained simultaneously.

2. Item selection algorithm

The system of the mirrored assessment selected the first item to be given to a student based on the national grade level in which the student was enrolled. Specifically, the first item was selected so its difficulty parameter equaled the mean performance for students in the same grade from the norming sample, which was 0.3 logits for the mimicked assessment. Each subsequent item was selected based on content balancing and item exposure constraints specified below.

3. Content balancing and item exposure

The simulator adaptively selected subsequent items that match the content balancing test specifications entered to the simulator. I used a *randomesque procedure* to control

item exposure. In this procedure, the simulator selected the ten best items that provided the most statistical information concerning the examinee. These items were ones with difficulty measures closest to the current ability level estimate for the student. After identifying the ten items, one was selected at random and administered to the examinee. With this approach, the adaptive algorithm maximized the information obtained from each item. This procedure also maximized the assessment efficiency while balancing item usage with similar psychometric characteristics.

4. Stopping rule

I used a fixed-length stopping rule for this study. The simulator administered 40 items to each examinee, which mirrored the number of items administered in the modeled assessment.

E. Item Parameter Drift Evaluation Criteria

I evaluated the impact of IPD on person measure estimates and classification of examinees into achievement levels using several measures. I obtained those measures for each replication and averaged them over the 100 replications. I also evaluated how the impact of IPD on low, medium, and high ability PPD examinees varies by the change in targeting of the items between two item pools.

1. Impact on person measure estimates

In order to assess the impact of IPD on person measure estimates, I compared true ability measures,θ_i, for i=1….*N* of the simulated examinees (where θ_i is the ability measure of the *i*th examinee and *N*= 500) to the estimated ability measures computed by the simulator using three indices: (a) bias values, (b) root mean squared error values (RMSEs), and (c) mean absolute differences (MADs). These statistics are frequently used by studies to evaluate precision of ability estimation (Chen, 2013; Han & Guo, 2011; Kingsbury & Wise, 2011; Risk, 2015;

Store, 2013; Ye & Xin, 2014). Bias is used for measuring systematic deviation of the estimated ability from the true ability, indicating whether model might be over- or underestimating the true values. The lower bias value indicates that the estimated ability is closer to the true ability. RMSE (also called root mean squared deviation, RMSD) provides an accurate indication of the total distance between the estimated and true ability. The lower the RMSE value, the more accurate estimates of true ability are obtained. Lastly, MAD is a measure of mean absolute bias among true and estimated ability parameters. Similarly, the smaller MAD indicate higher precision in ability estimates. Formulas associated with each index are provided in Table 7 below. I also calculated the Pearson correlation coefficient between true and estimated ability parameters for each condition as measures of degree of successful ability parameter recovery.

TABLE VII

EVALUATION CRITERIA FOR ESTIMATION PRECISION

Purpose	Measure Index	Description	Calculation
Evaluation of estimation precision	Bias	Systematic deviation of estimated ability from true ability	$\frac{\sum_{j=1}^{n}(\hat{\theta}_i - \theta_i)}{n}$
	Root Mean Square Error (RMSE)	A measure of total distance between the estimated and true ability	$\sqrt{\frac{\sum_{j=1}^{n}(\hat{\theta}_i - \theta_i)^2}{n}}$
	Mean Absolute Difference (MAD)	A measure of average absolute bias among true and estimated ability	$\frac{\sum_{i=1}^{n}\lvert\hat{\theta}_i - \theta_i\rvert}{n}$
	Pearson product-moment correlation coefficient *(r)*	A measure of successful recovery of the generating ability parameters	$\frac{\sum_{i=1}^{n}(Q_i - \overline{Q_\iota})(\widehat{Q_\iota} - \overline{Q_\iota^\wedge})}{\left(\sqrt{\sum_{i=1}^{n}(Q_i - \overline{Q_\iota})^2}\right)\left(\sum_{i=1}^{n}(\widehat{Q_\iota} - \overline{Q_\iota^\wedge})^2\right)}$

Note. $\widehat{\theta_\iota}$, θ_i *represents the estimated and true thetas for examinee i, n is the total number of examinees in each condition*

2. Classification accuracy

I examined classification accuracy based on the cut scores for the three achievement levels. These cut scores were specified based on the mimicked assessment. Specifically, I identified the number and percentages of false decisions focusing on the false negative decisions that were outside 95% CI for each condition. I also calculated misclassification rate of PPD examinees in order to examine individual impact of IPD items on classification accuracy regardless of the varying number of PPD examinees in each condition. Lastly, I analyzed rank ordering change of examinees due to IPD using Spearman's rho rank order correlation.

As a supplementary analysis, I analyzed WINSTEPS person Infit and Outfit mean-square fit statistics to understand if any PPD examinees affected by IPD could be detected using traditional fit statistics in real life tests. The Infit and Outfit mean-square statistics are measures of unexpectedness in responses, are calculated based on conventional chi-square statistics. Their expected value is 1.0. Values larger than 1.0 indicate construct irrelevant variance in the responses while values smaller than 1.0 indicate that the measurement model predicts the data too well (Linacre, 2005).

TABLE VIII

EVALUATION CRITERIA FOR CLASSIFICATION ACCURACY

Purpose	Measure Index	Description	Calculation
Evaluation of classification accuracy	False negatives (FN)	The number of examinees being classified into a lower achievement level	NA
	Percentage of misclassifications	The total percentage of false negatives resulting from each condition in the study	$\frac{FN}{n} * 100$
	Person mean-square fit statistics of those misclassified (MNSQ)	To analyze if examinees get hard items unexpectedly right or easy items unexpectedly wrong	$Outfit = \frac{1}{L}\sum_{i=1}^{L}\frac{(X_{ni}-E_{ni})^2}{V_{ni}}$ $Infit = \frac{\sum_{i=1}^{L}(X_{ni}-E_{ni})^2}{\sum_{i=1}^{L}V_{ni}}$
	Spearman's rho rank order correlation	A measure of association between rank ordering of examinees based on true and estimated abilities	$1-\frac{6\sum_{i=1}^{n}di^2}{n(n^2-1)}$
	Misclassification rate	Proportion of misclassified PPD examinees to all PPD examinees in each condition	$\frac{\#\ of\ misclassified\ PPD\ examinees}{\#\ of\ PPD\ examinees}$
Evaluation of IPD impact on classification by ability	False decisions by ability group	Number and proportion of misclassifications by ability group in two item pools	NA

Note. $\widehat{\theta_i}$, θ_i *represents the estimated and true thetas for examinee i, n is the total number of examinees in each condition.* X_{ni} *is the observed response for person n on item i,* E_{ni} *is the expected response for person n on item i,* V_{ni} *is the variance, L is the number of items on the test.*

3. Impact of item parameter drift on high, medium, and low ability examinees

I evaluated how impact of IPD on classification accuracy varies by item pool targeting based on the differences in the number and proportion of misclassifications for low, medium and high ability PPD examinees. The high ability PPD examinees were those with ability estimates one standard deviation above the mean ability of the sample. The low ability

PPD examinees had ability estimates one standard deviation below the mean ability of the sample. I set this criterion based on the examinee sample distribution and standard deviation. In a normal distribution, the upper tail, which is one standard deviation above the mean, represents higher scores and the lower tail, which is one standard deviation below the mean, represents lower scores.

F. Analysis Methodology

1. Response model

I employed the Rasch dichotomous model to calibrate difficulty and ability parameters for the simulated CAT (Rasch, 1960). The dichotomous model can be used when a response to an item is scored dichotomously; either correct or incorrect (Linacre, 2005). The probability of giving a specified answer (e.g., correct/incorrect) was quantified as a function of person and item parameters. The mathematical model is given as:

$$\Pr\{X_{ni} = 1\} = \frac{e^{\beta n - \delta_i}}{1 + e^{\beta n - \delta_i}}, \quad (1)$$

where *Pr* is the probability of examinee *n* scoring 1 on item *i*, δ is the difficulty parameter of item *i*, and β is the ability parameter of examinee *n*.

2. Baseline (zero item parameter drift) condition

A condition that includes zero IPD items served as a baseline for comparisons with other conditions including IPD items. This baseline condition allowed me to obtain initial rates for estimation precision and classification accuracy without the influence of IPD. I then compared the values that I got under each IPD condition to the values in the zero IPD condition and evaluated the extent of change in precision of estimates and classification accuracy based on the evaluation criteria.

3. Research questions

In order to address my research questions, I employed a crossed 2 (item pool) × 4 (percentage of content area) × 3 (percentage of PPD examinee) factorial design with a total of 24 conditions. I examined both main and interactions effects using several statistical criteria. In order to examine impact of IPD on ability estimates and classification accuracy, I compared values obtained experimentally for each condition to the values obtained in the zero IPD condition. I calculated bias, RMSE, and MAD statistics to assess impact on estimation precision. I compared the number of false negatives and total percentage of false negatives across IPD conditions in order to examine the impact on classification accuracy. Finally, in order to evaluate the effect of IPD on high, medium and low ability PPD examinees, I examined the total number and percentage of false decisions at each ability group separately across two item pools.

IV. RESULTS

A. Ability Estimation

The ability parameters (θ) of 500 examinees were estimated in 100 replications for each of the 24 conditions. The ability estimates were then compared to the values in baseline conditions and true ability parameters via the RMSE, bias, and MAD statistics. The correlation between true and estimated ability parameters were also reported for each condition. The changes in the RMSE, bias, and MAD values compared to the baseline condition for well targeted and poorly targeted item pools are shown in Table 9. The results indicated that the lowest amount of bias, RMSE, and MAD and the highest correlation between true and estimated ability occurred for the baseline conditions in both item pools (see Table 9), as expected. As the percentage of IPD items and examinees with the IPD effect increased, the mean bias, RMSE, and MAD values increased indicating that measurement precision decreased. Similarly, as the number of IPD items and number of affected people by IPD increased, the correlation between true and estimated ability measures decreased, supporting that measurement precision decreased. In the following sections, the relation between the measures of precision and IPD is discussed.

1. Bias

The impact of bias can be cancelled out if items exhibit IPD in different directions (Wei, 2013). The sign of bias is about the direction of IPD (C. Han, personal communication, December 7, 2015). Since drift was presented in one direction in this study (all items becoming harder), bias values were found to be negative in all IPD conditions. The smallest bias in ability estimates was obtained for the baseline conditions for both item pools. These were expected findings since the baseline conditions did not reflect drift in any item parameters and thus they were

TABLE IX

PRECISION OF ABILITY ESTIMATES

Item pool	% of IPD items	% of PPD examinees	Bias	RMSE	MAD	Correlation
Well-targeted item pool	Baseline	N/A	-0.00336	0.035187	0.028115	0.99932
	20%	20%	-0.028998	0.073404	0.049492	0.997464
		30%	-0.042288	0.085336	0.061000	0.996929
		40%	-0.057922	0.097987	0.074063	0.996506
		50%	-0.076512	0.111446	0.088163	0.996336
	40%	20%	-0.061339	0.138257	0.081413	0.991629
		30%	-0.088650	0.167246	0.107494	0.988887
		40%	-0.117125	0.190369	0.134280	0.987443
		50%	-0.150767	0.215506	0.163212	0.986958
	60%	20%	-0.094226	0.205698	0.112666	0.982027
		30%	-0.132632	0.245921	0.152485	0.976489
		40%	-0.180297	0.284752	0.194197	0.973383
		50%	-0.223350	0.319164	0.236687	0.971826
Poorly targeted item pool	Baseline	N/A	-0.00666	0.033015	0.034968	0.99939
	20%	20%	-0.031034	0.074366	0.051404	0.997457
		30%	-0.045922	0.087732	0.063045	0.996887
		40%	-0.059188	0.100508	0.076029	0.996315
		50%	-0.072755	0.106334	0.084482	0.996642
	40%	20%	-0.058934	0.135935	0.079991	0.991797
		30%	-0.087231	0.166596	0.107624	0.988852
		40%	-0.120802	0.191568	0.134104	0.987688
		50%	-0.148671	0.214975	0.162064	0.986680
	60%	20%	-0.090394	0.205579	0.113461	0.981765
		30%	-0.134505	0.245498	0.152072	0.976873
		40%	-0.179948	0.285232	0.194847	0.973104
		50%	-0.222363	0.319585	0.236357	0.971677

created to provide a recovery rate without IPD impact. The mean bias in ability increased linearly in a negative direction as the percentage of IPD items and percentage of PPD examinees increase for both item pools. The largest amount of bias was observed in the most extreme conditions where 50% of items in the pool drifted for 60% of examinees (see Figures 1 and 2).

There was not any systematic pattern between two item pools in terms of the values of bias statistics. In some conditions, the bias values were higher in magnitude in the well (poorly) targeted pool, in some others, they were equal (see Figures 1 and 2.)

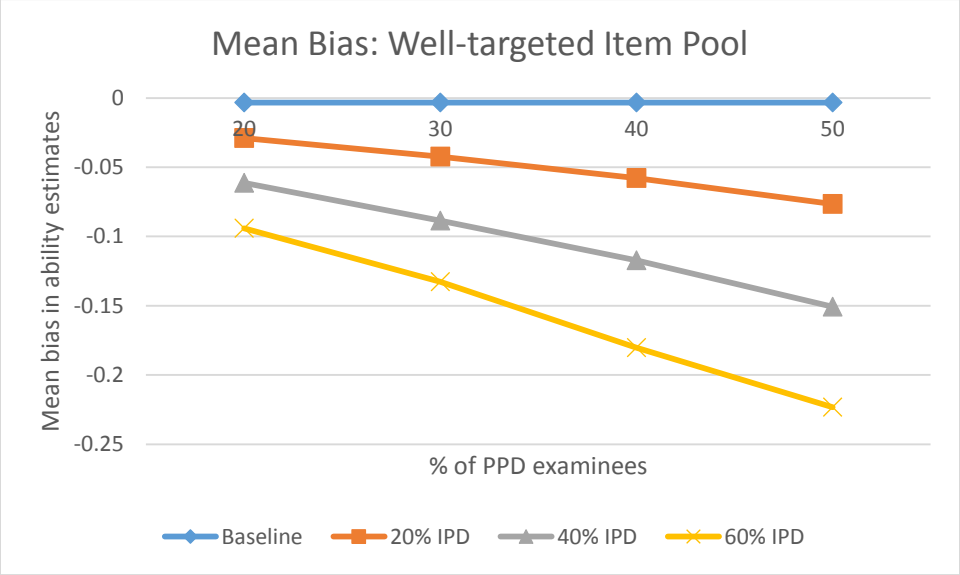


Figure 1. Mean bias values for the well-targeted item pool.

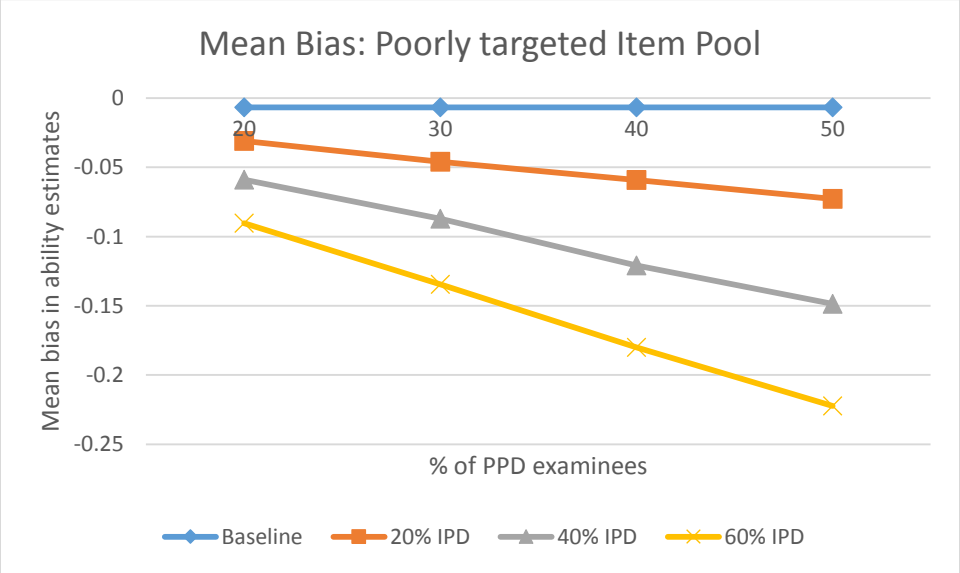


Figure 2. Mean bias values for the poorly targeted item pool.

2. Root mean-squared error

As expected, the smallest RMSE values were observed in the baseline conditions for both item pools. The RMSE value increased linearly when the percentage of IPD items increase in the item pool and when the percentage of PPD examinees increased in the examinee sample. This systematic pattern occurred in both item pools. Within each percentage of IPD items (20%, 40%, and 60%), conditions with 50% PPD examinees yielded the highest RMSE values followed by 40%, 30%, and 20% of PPD examinees, respectively. Unlike the bias values, RMSE values were almost equal in both item pools across conditions except the condition with 20% IPD items and 50% PPD examinees (see Figures 3 and 4).

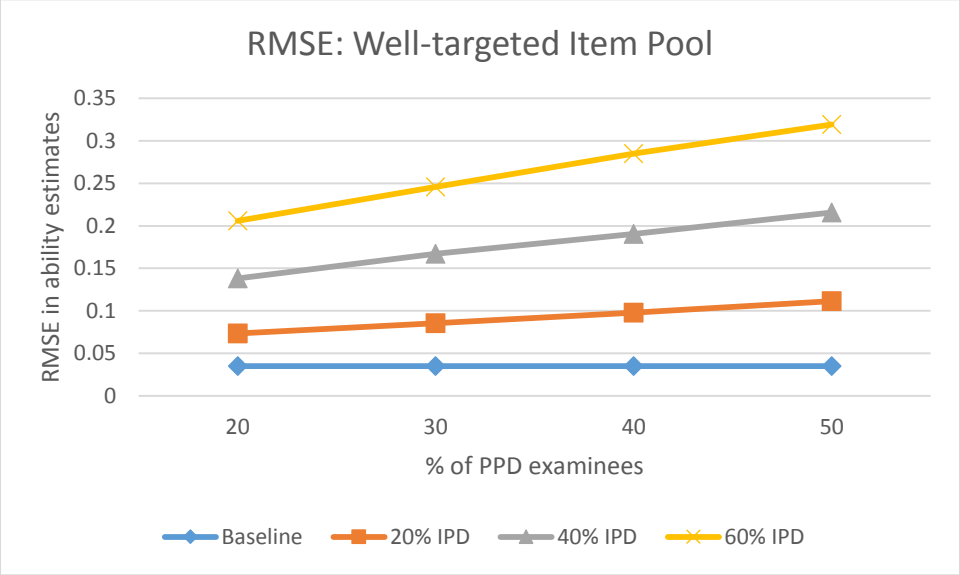


Figure 3. RMSE values for the well-targeted item pool.

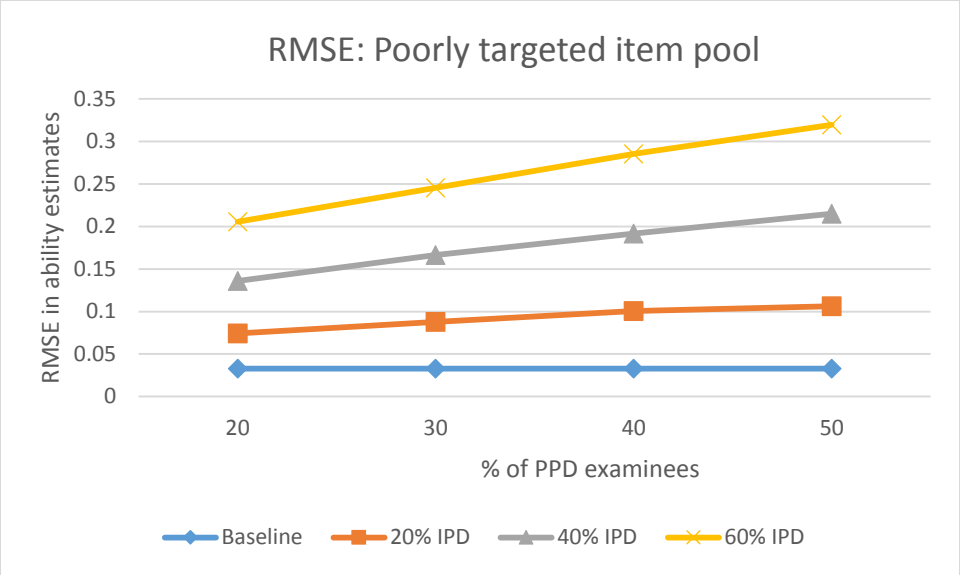


Figure 4. RMSE values for the poorly targeted item pool.

3. Mean absolute difference

Similar to the RMSE and bias values, the smallest MAD was observed in the baseline conditions as expected. The MAD between true and estimated ability increased as the percentage of IPD items increased in the item pools. Within the same percentage of IPD items, the more PPD examinees, the higher MAD values. The MAD values of two item pools were very close for the same percentage of IPD items and PPD persons (see Figures 5 and 6). In the worst case scenario where 60% of items drift for 50% of examinees, the MAD value increased to 0.23. This result is promising when considering the standard error of ability estimates is usually about 0.3 logits in practice (Han & Guo, 2011).

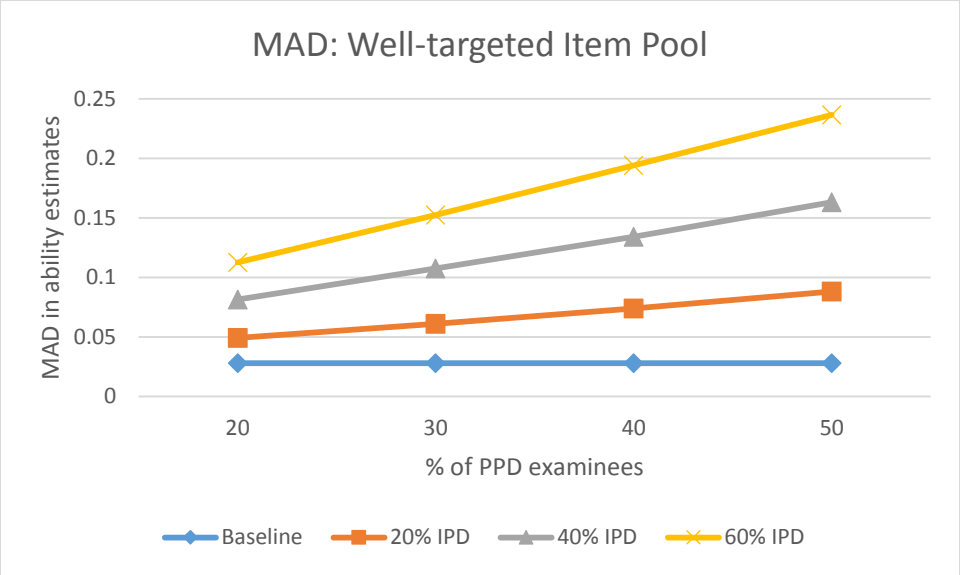


Figure 5. MAD values for the well-targeted item pool.

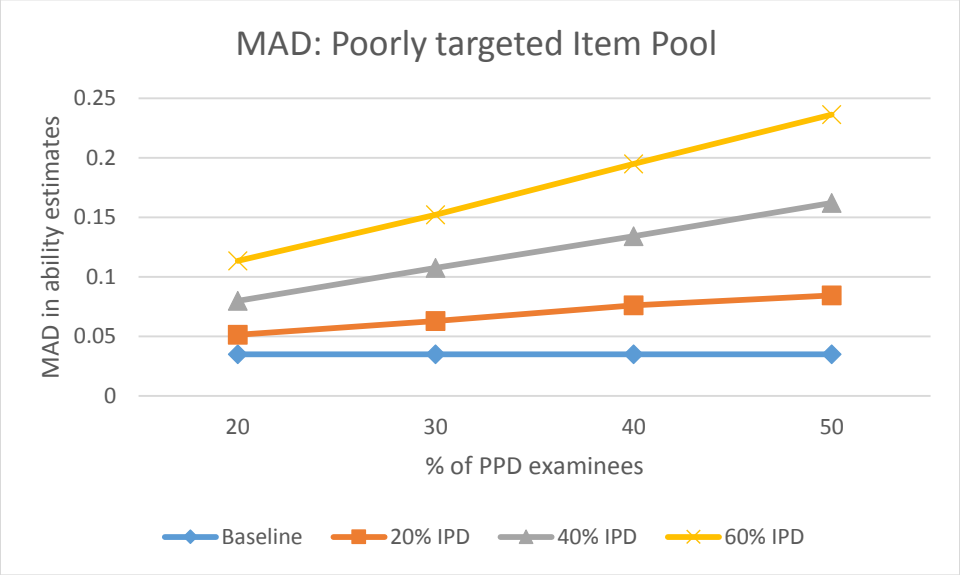


Figure 6. MAD values for the poorly targeted item pool.

4. Correlation

The Pearson correlation coefficient between true and estimated ability parameters were very close to 1.00 for the baseline conditions indicating successful recovery of the generating parameters. As IPD was introduced to the item pools, the correlation between true and estimated ability parameters decreased (see Table 9), but the change was negligible. The lowest

correlation was observed in the conditions where 60% of the items drift in both pools. The increasing percentage of PPD examinees in the sample did not influence the correlation as much as the increasing percentage of IPD items. For example, within the same percentage of IPD items, the correlation values slightly decreased as the percentage of PPD examinees increased. On the other hand, within the same percentage of PPD examinees, the correlation decreased from 0.99 in the 20% IPD condition to 0.97 in the 60% IPD condition (see Table 9).

B. Classification Accuracy

Similar to the precision measures for the ability estimates, classification accuracy measures yielded consistent findings across conditions. As expected, the best results with the least percentage of misclassified examinees ((number of false decisions/sample size) * 100) occurred in the baseline conditions since the person parameter estimates in the baseline conditions were not affected by IPD. As IPD was introduced to the item pools, classification accuracy declined (see Table 10). The most misclassifications were observed at the most extreme IPD conditions in both item pools. Changes in the measures of classification accuracy are discussed below.

TABLE X

MEASURES OF CLASSIFICATION ACCURACY

Item pool	Percentage of IPD items	Percentage of PPD examinees	Number of False Negatives	Number of False Positives	Total percentage of mis-classification	False Negatives outside 95%CI	False Positives outside 95%CI
Well-targeted item pool	**Baseline**	**N/A**	**1**	**5**	**1.2%**	**0**	**0**
	20%	20%	10	1***	2.2%	2 (0.4%)	0
		30%	19	0	3.8%	7 (1.4%)	0
		40%	21*	1***	4.4%	5 (1.0%)	0
		50%	26	0	5.2%	9 (1.8%)	0
	40%	20%	18	0	3.6%	13 (2.6%)	0
		30%	35*	0	7%	26 (5.2%)	0
		40%	38	0	7.6%	22 (4.4%)	0
		50%	55	0	11%	42 (8.4%)	0
	60%	20%	31	0	6.2%	20 (4.0%)	0
		30%	57*	0	11.4%	41 (8.2%)	0
		40%	58	1	11.8%	44 (8.8%)	0
		50%	87*	1	17.6	66 (13.2%)	0
Poorly targeted item pool	**Baseline**	**n/a**	**4**	**2**	**1.2%**	**0**	**0**
	20%	20%	11	1***	2.4%	3 (0.6%)	0
		30%	22	0	4.4%	11(2.2%)	0
		40%	25**	0	5%	7 (1.4%)	0
		50%	32	2***	6.8%	11 (2.2%)	0
	40%	20%	19*	0	3.8%	12 (2.4%)	0
		30%	36*	0	7.2%	30 (6%)	0
		40%	35	0	7%	25 (5%)	0
		50%	54	1***	11%	39 (7.8%)	0
	60%	20%	34*	1***	7%	25 (5%)	0
		30%	55	0	11%	38 (7.6%)	0
		40%	57	1***	11.6%	40 (8%)	0
		50%	83	0	16.6%	61 (12.2%)	0

Note: All PPD unless indicated; * 1 non-PPD; ** 2 non-PPD; *** non-PPD

1. Number and percentage of misclassifications

The percentage of misclassification patterns in the two item pools were consistent with each other. The lowest misclassification percentage occurred in baseline conditions where no IPD items were present (see Table 10). The percentage of misclassified examinees increased as the percentage of IPD items in the pool increased. Similarly, more examinees were misclassified as the percentage of PPD examinees in the sample increased. The highest misclassification percentage occurred in conditions where there were 60% IPD items for 50%

examinees in each item pool. Since IPD occurred in one direction in this study, misclassification due to IPD was expected to be observed as false negative decisions; however, in some conditions a few PPD examinees were misclassified into a higher achievement level. In addition, some of the examinees misclassified into a lower achievement level were among the non-PPD examinees. These results were likely to occur due to measurement error, given the margin of error around the cut scores. Therefore, 95% confidence interval (CI) around the cut scores were calculated to elaborate on the significance of misclassifications. The results showed that all of the non-PPD misclassified examinees were within measurement error. In other words, only PPD examinees were effected by false negative decisions by being significantly misclassified into a lower achievement level. The percentage of significantly misclassified examinees increased as the percentage of IPD items increased in the item pool. There was not a consistent pattern in terms of percentage of significantly misclassified examinees within the same percentage of IPD items.

2. Misclassification rate of person parameter drift examinees

IPD items only impacted the PPD examinees in this study. Thus, I expected to observe an increased percentage of misclassifications as the percentage of PPD examinees increased. In order to understand the individual impact of IPD items on classification accuracy regardless of the number of PPD examinees in each condition, misclassification rate (number of misclassified PPD examinees/number of PPD examinees) were calculated. The results showed that the misclassification rate increased gradually when the percentage of IPD items increased in the item pool (see Table 11). Within the same percentage of IPD items, the misclassification rate changed within a small range that can be attributable to random fluctuation. The highest misclassification rate was observed in the 60% IPD conditions in both item pools.

TABLE XI

MISCLASSIFICATION RATE OF PPD EXAMINEES

Item pool	Percentage of IPD items	Percentage of PPD examinees	Misclassification Rate of PPD examinees
Well-targeted item pool	**Baseline**	**N/A**	**N/A**
	20%	20%	0.11
		30%	0.12
		40%	0.10
		50%	0.10
	40%	20%	0.18
		30%	0.22
		40%	0.19
		50%	0.22
	60%	20%	0.31
		30%	0.36
		40%	0.30
		50%	0.34
Poorly targeted item pool	**Baseline**	**N/A**	**N/A**
	20%	20%	0.11
		30%	0.14
		40%	0.11
		50%	0.12
	40%	20%	0.19
		30%	0.23
		40%	0.17
		50%	0.21
	60%	20%	0.33
		30%	0.36
		40%	0.28
		50%	0.33

3. Rank ordering change

Despite the fact that most K-12 assessments focus on examinees' classification into achievement levels as their outcomes, some of them also report the percentile rank of examinees (i.e., NWEA MAP). In order to assess the extent of rank ordering change of examinees based on their estimated ability parameters, Spearman's rank-order correlation was

used. The Spearman's correlation coefficient measured the strength of association between examinees' true percentile rankings based on their true ability parameter and percentile rankings based on their estimated abilities. The results were consistent across conditions (see Table 12). All conditions yielded high and significant rank-order correlations. This indicated a non-significant change in examinees' rank ordering in all conditions, even the most extreme ones. The highest observed rank-order correlation values occurred in baseline conditions, indicating the closest rank-ordering to the rank-ordering with the 'true" ability parameters. Overall, positive and significant rank-order correlation values associated with IPD conditions indicated stable rank ordering of examinees even when IPD exists in an item pool.

4. Evaluation of person fit

While the overall impact of IPD on PPD examinees is evaluated by bias, RMSE, MAD, and misclassifications, one important question is how to detect the affected people (i.e., PPD examinees) based on their unexpected responses in real-life test situations. Person fit statistics are used to detect examinees that are not performing in accordance with the measurement model expectations. Once such examinees are detected, potential reasons of unexpected responses (i.e., cheating, curricular, instructional and practice differences, security breaches) can be diagnosed based on unexpected response patterns. In order to assess the effectiveness of person fit statistics in detecting PPD examinees in this study, I examined WINSTEPS person Infit and Outfit mean-square (MNSQ) fit statistics. At this part of the analyses, all IPD items were anchored at their non-IPD difficulties in order to model a situation where item difficulties have changed due to IPD, but this change has yet to be recognized by the testing organization.

TABLE XII

SPEARMAN'S RANK-ORDER CORRELATION OF RANKINGS BASED ON TRUE AND ESTIMATED ABILITIES

Item pool	% of IPD items	% of PPD examinees	Spearman's rho rank-order correlation
Well-targeted item pool	**Baseline**	**N/A**	**0.999**
	20%	20%	0.997
		30%	0.997
		40%	0.996
		50%	0.996
	40%	20%	0.991
		30%	0.988
		40%	0.986
		50%	0.986
	60%	20%	0.981
		30%	0.974
		40%	0.972
		50%	0.970
Poorly targeted item pool	**Baseline**	**N/A**	**.999**
	20%	20%	.997
		30%	.996
		40%	.996
		50%	.996
	40%	20%	.991
		30%	.988
		40%	.987
		50%	.986
	60%	20%	.981
		30%	.975
		40%	.971
		50%	.970

Then, CAT response strings were calibrated and person MNSQ fit statistics were averaged over 100 replications in each condition to examine if any misfit patterns associated with PPD examinees could be observed.

The results showed that WINSTEPS person MNSQ fit statistics didn't perform efficiently to detect unexpectedness in the PPD examinee responses. Across 100 replications in each of the 24 IPD conditions, the average person Outfit MNSQ statistic values fell within the range of 0.94 to 0.96. These values are not usable to detect aberrant responses given the distribution of Outfit statistics with an expected mean of 1.00. The baseline condition person Outfit MNSQ statistic values were almost equal to the expected value of 1.00, indicating that IPD resulted in a very slight deviance that did not yield informative results to detect PPD examinees in this study. Figure 7 shows the distribution of Infit and Outfit statistics values for PPD and non-PPD examinees in the most extreme IPD condition in which 60% IPD items drift for 50% of the examinees in the sample. Examinees flagged with zero represent non-PPD examinees and examinees flagged with 1 are PPD examinees. As seen in the Figure 7, the fit statistics values are not substantially different for the two groups even in the most extreme IPD condition.

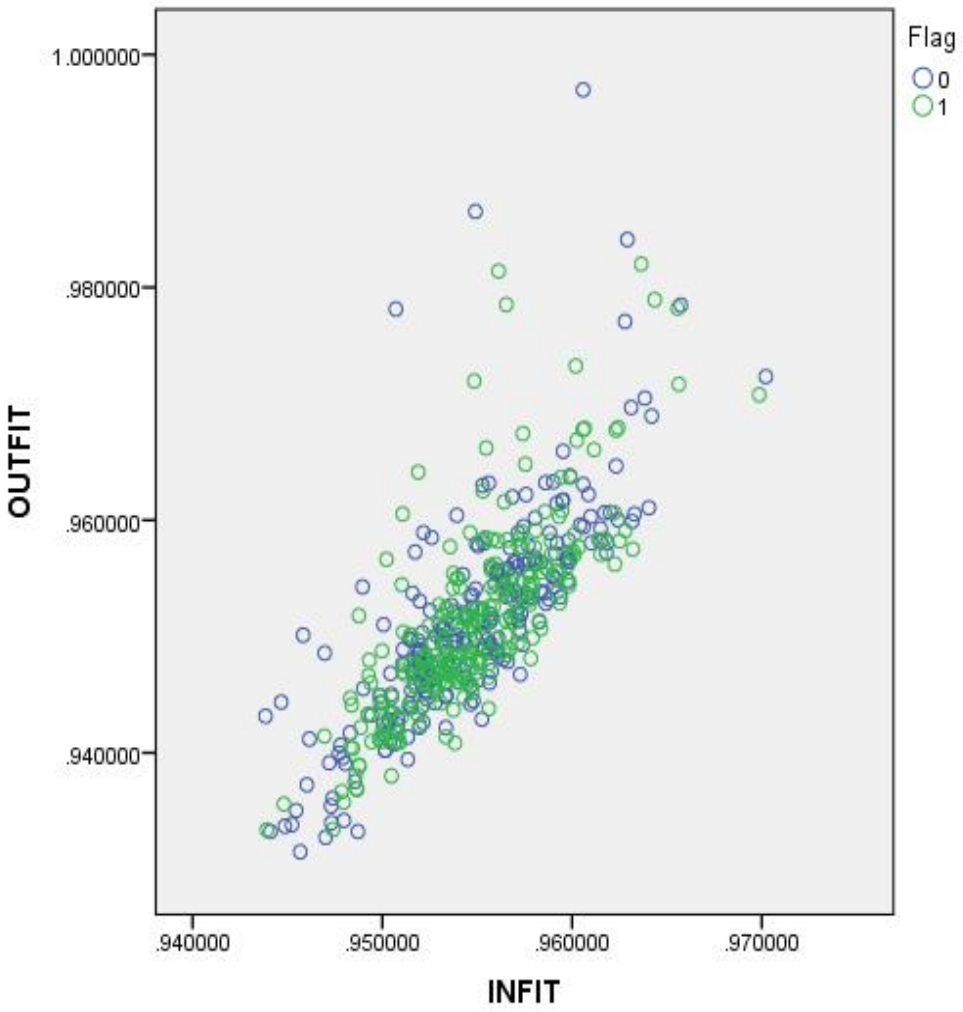


Figure 7. Infit and Outfit MNSQ plot of PPD vs non-PPD examinees in the most extreme IPD condition.

C. Ability Level Composition of Misclassified Examinees

In this part of the analyses, I dissected misclassification results by examinees' ability levels to examine if the proportion of low, medium, and high ability misclassified examinees vary across conditions. Most misclassifications occurred around the higher cut score (0.2 logits) in both item pools while most misclassified examinees were among the medium-ability examinees, followed by low-ability examinees in all conditions. In both item pools, there were

no high-ability examinee who was misclassified (see Table 13). The following describes the misclassified examinees' ability composition in each item pool.

TABLE XIII

MISCLASSIFIED EXAMINEES BY ABILITY LEVEL

Item pool	% of IPD items	% of PPD examinees	Number and proportion of misclassified low ability examinees	Number and proportion of misclassified medium ability examinees	Number and proportion of misclassified high ability examinees	Total Misclassification
Well-targeted item pool	**Baseline**	**N/A**	**2 (0.33)**	**4 (0.67)**	**0**	**6 (1.2%)**
	20%	20%	3 (0.27)	8 (0.73)	0	11 (2.2%)
		30%	3 (0.16)	16 (0.84)	0	19 (3.8%)
		40%	6 (0.27)	16 (0.73)	0	22 (4.4%)
		50%	7 (0.27)	19 (0.73)	0	26 (5.2%)
	40%	20%	4 (0.14)	14 (0.86)	0	18 (3.6%)
		30%	5 (0.14)	30 (0.86)	0	35 (7%)
		40%	5 (0.13)	33 (0.87)	0	38 (7.6%)
		50%	9 (0.16)	46 (0.84)	0	55 (11%)
	60%	20%	5 (0.16)	26 (0.84)	0	31 (6.2%)
		30%	5 (0.08)	52 (0.92)	0	57 (11.4%)
		40%	5 (0.08)	54 (0.92)	0	59 (11.8%)
		50%	10 (0.11)	78 (0.89)	0	88 (17.6%)
Poorly targeted item pool	**Baseline**	**N/A**	**2 (0.33)**	**4 (0.67)**	**0**	**6 (1.2%)**
	20%	20%	4 (0.33)	8 (0.67)	0	12 (2.4%)
		30%	5 (0.22)	17 (0.78)	0	22 (4.4%)
		40%	7 (0.28)	18 (0.72)	0	25 (5%)
		50%	9 (0.26)	25 (0.74)	0	34 (6.8%)
	40%	20%	4 (0.21)	15 (0.79)	0	19 (3.8%)
		30%	3 (0.08)	33 (0.92)	0	36 (7.2%)
		40%	5 (0.14)	30 (0.86)	0	35 (7%)
		50%	9 (0.16)	46 (0.84)	0	55 (11%)
	60%	20%	4 (0.11)	31 (0.89)	0	35 (7%)
		30%	6 (0.1)	49 (0.9)	0	55 (11%)
		40%	7 (0.12)	50 (0.88)	0	58 (11.6%)
		50%	10 (0.12)	73 (0.88)	0	83 (16.6%)

1. Well-targeted item pool

The proportion of low, medium and high ability misclassified examinees were calculated for each condition. The results showed that the proportion of misclassified low-ability examinees decreased as the percentage of IPD items increased in the well-targeted item pool (see Table 13). The baseline condition yielded the highest proportion of misclassified low-ability examinees at 0.33. In the 20% IPD conditions, the proportion of low-ability misclassified examinees ranged between 0.16 and 0.27. The proportion of misclassified low-ability examinees was .10 on average in the 60% IPD conditions. The more IPD items in the item pool, the higher proportion of medium ability examinees in the misclassified examinees observed. These findings indicated that IPD items tended to influence classification accuracy of medium-ability examinees in the well-targeted item pool.

2. Poorly targeted item pool

I generated a poorly targeted item pool where items were easier for the average ability examinees, allowing low-ability examinees to see as many IPD items as possible. Although the low-ability examinees were exposed to more IPD items in the poorly targeted item pool, the impact IPD had on their classification accuracy did not practically differ from the well-targeted item pool. Similar to the well-targeted item pool, the more IPD introduced in the pool, the higher proportion of medium ability examinees were detected as misclassified. On average, the proportion of low-ability examinees misclassified was slightly higher than the well-targeted item pool, but this difference was negligible. None of the high ability examinees were misclassified, indicating that IPD was not strong enough to impact them since their ability level is well-above the cut scores and the majority of IPD items in the pool.

V. DISCUSSION

A. Chapter Overview

The purpose of this research was to investigate the potential impact of IPD that occurs as a result of changes in examinee knowledge and skills over test administrations that can attributable to the factors such as infrastructural, , curricular, instructional, and practice differences on person measure estimation and classification accuracy. Simulated data was used to answer five research questions. Analysis of the data showed that IPD that was exposed to only a sub-group of examinees can affect classification accuracy of those examinees substantially but its impact on average ability estimation was small. This chapter summarizes findings in reference to the research questions, outlines strengths and limitations of the study, presents new research avenues for future studies, and proposes some implications for testing companies and states.

B. Summary of Findings by Research Question

1. Research question 1: What is the impact of IPD on person measure estimates when only a sub-group of examinees are affected by the drift?

The simulation results showed that on average, the IPD impact on ability estimates was small, but substantive when only a subgroup of examinees in the sample was affected. The magnitude of negative bias, RMSE, and MAD values started to diverge from the baseline recovery rate as IPD was exposed to the test. The highest RMSE values reached up to 0.32, which is slightly higher than the standard error of ability estimation in practice, in the 60% IPD conditions. Despite 60% of the items drifting in a CAT exam being a very rare situation, one should keep in mind that differential exposure of some examinees to one or more content areas in an exam can matter significantly in terms of estimation precision since examinees are disadvantaged for all items under the poorly learnt content area to varying degrees.

Consistent with the previous studies on IPD impact, (Han & Guo, 2011; McCoy, 2009; Risk, 2015; Wei, 2013; Wells, Subkoviak, & Serlin, 2002), larger amounts of drift resulted in less precise estimates. On the other hand, the values of bias, RMSE, and MAD statistics under the IPD conditions were larger than these previous studies. For example, studies that examined multidirectional IPD impact on estimation precision found minimal impact on ability estimates with RMSE values fluctuating within a small range around the baseline recovery rate, even under the most extreme conditions (Han, Wells, & Sireci, 2012; Risk, 2015). This study's measures of precision values were also larger than the previous studies' measures that analyzed unidirectional IPD impact on ability estimates (Babcock & Albano; 2012; Guo, 2009; Han & Guo; 2011; Wei, 2013). One factor that can explain the difference between the results is exposing a known percentage of IPD items to the PPD examinees in this study. Previous simulation studies assessing IPD impact on ability estimation exposed a probabilistic draw of IPD items to the entire examinee sample (Guo, 2009; Risk, 2015; Wei, 2013; Wells, Subkoviak, & Serlin, 2002). Their study conditions were well-suited to the IPD situations such as item overexposure, cheating, security breaches, or item position effect. Yet, they cannot model drift situations where examinee ability changes due to the factors such as learning, differential exposure to the test content or lack of practice on one or more content in a CAT exam. This study adds to the IPD literature that if examinees' ability levels change due to factors such as differential school resources, curriculum, instruction, technology or practice, which was named as "person parameter drift", the impact of IPD on precision of ability estimates can be substantial.

2. Research question 2: What is the impact of IPD on classification of examinees to achievement levels when only a sub-group of people are affected by the drift?

While IPD impact on ability estimates were significant but small, the consequential impact of IPD was indicated by misclassification results of the PPD examinees across conditions. Even in the smallest amount of IPD conditions (i.e., 20% of the test), there were a few misclassified PPD examinees outside of 95% CI. As aligned with the IPD modeling in the study, the misclassifications were observed as false negative decisions; examinees being classified into a lower achievement level. The significance of a false decision should be assessed based on the purpose of a CAT exam. Previous IPD studies elaborated on misclassifications that occurred due to type II errors, mostly in pass/fail decisions in certification and licensure exams (McCoy, 2010; Risk, 2015, Witt, Stahl, Bergstrom, & Muckle, 2003; Sukin, 2010). Their findings showed minimal misclassifications using one cut score. Unlike certification and licensure examinations, CAT is used not only for pass/fail decisions, but also for predicting examinees' true achievement levels for formative and summative evaluations in educational testing (Way et. al., 2010). This requires use of multiple cut scores. Using multiple cut scores increases likelihood of a misclassification as a result of random error.

In order to control misclassifications due to use of multiple cut scores placed within a small range of ability on the scale in the modeled assessment, I collapsed some of them and reduced the likelihood of false decisions by random error. Despite this adjustment to the number of cut scores, the percentage of misclassifications in this study was noticeable. As expected, the conditions with lowest amount of IPD yielded the smallest percentage of significant misclassifications; however, the practical importance of IPD impact on classification should not

be assessed by the small number of misclassified people, but by the exam stakes. In educational testing, an IPD situation due to lack of equal opportunity to learn and perform on a test across examinee groups, districts, or states may threat on fairness of important decisions linked to the test results. As stated in *The Code of Fair Testing Practices in Education* (1988), fairness indicates that every test taker is provided equal opportunity to prepare for the test. Unless each examinee has been given equal access to quality instruction or resources, they cannot have same opportunity to demonstrate their skills and ability with other examinees. Consequently, their test performance and decisions linked to the test performance (i.e., placement to achievement levels, teacher performance evaluations) might be derailed. This warrants further investigation on the potential impact of IPD due to lack of content knowledge that can occur as a result of fewer opportunities to learn on classification accuracy in educational CAT exams.

3. Research question 3: Are the effects of IPD on person measure estimates and classification of examinees consistent across three factors of drift: proportion of IPD items in the pool, proportion of PPD examinees in the sample, and item pool targeting?

a. Proportion of IPD items in the pool

1) Person measure estimation

The IPD impact on ability parameter estimates showed a consistent and linearly increasing pattern with increasing percentage of IPD items in the pool. The smallest bias, RMSE, and MAD values were observed with the smallest amount of IPD items in the bank (20% IPD) across the drift conditions. All values increased linearly in magnitude with increasing percentage of drift items. Unlike some of the existing studies' findings, the bias values increased in negative direction and ranged within a relatively larger interval (Babcock & Albano, 2012;

Han, Wells, & Sireci, 2012; Risk, 2015). The large and linearly increasing negative bias associated with ability estimates occurred as a result of IPD direction (all harder) and IPD item exposure to the PPD examinees in this study. When items drift in both directions, the bias statistics are canceled out. In addition, when examinees are exposed to a known number of IPD items, bias, RMSE, and MAD statistics show the substantial impact of IPD on estimation precision.

2) Classification

The PPD examinees classification results showed the practical importance of the IPD item amounts on decisions linked to the CAT exam. Increasing the percentage of IPD items resulted in more PPD examinees misclassified to a lower achievement level. The percentage of false positive decisions did not show a consistent pattern with respect to the percentage of IPD items. False negative decisions (i.e. failing an examinee if she would actually pass) tend to carry more serious threat to test validity and reduce the fairness of a test than false positive decisions in educational testing. Thus, testing companies should decide on the minimum amount of IPD that can derail the decisions linked to CAT exams, particularly in education.

b. Proportion of PPD examinees in the sample

1) Person measure estimation

In this study, a wide range of PPD examinee percentages were simulated in order to see the average IPD impact on smaller and larger examinee subgroups. On average, conditions with the smallest PPD examinee percentage revealed precision results closer to the baseline recovery rate. When more PPD examinees were added to the sample, the less precise ability estimates were observed, as expected. Even when 60% of the sample was affected by IPD, the average error associated with ability estimates did not exceed .30 logits; however, these

results might be misleading if they are not evaluated with the classifications results as they outline the practical importance of drift.

2) Classification

Within the same percentage of IPD condition, the percentage of false negative decisions increased as the percentage of PPD examinees increased, as expected. The percentage of false positive decisions were minimal and close to each other in all conditions, regardless the percentage of PPD examinees. These findings were aligned with the IPD modeling in the study. A noticeable pattern occurred when the percentages of significant false decisions were calculated. Specifically, within the same percentage of IPD items, the number of misclassifications for 30% and 40% PPD examinees did not differ substantially while the percentage of significant misclassifications increased in the 50% PPD conditions. This pattern implied that the proportion of PPD examinees in the sample may impact the problem of misclassification, particularly when it exceeds a certain amount.

c. Item pool targeting

1) Person measure estimation

The ability estimate precision indicators were very similar across conditions in the two item pools. This is an expected result given the fact that the CAT algorithm targets examinee abilities. The poorly targeted item pool had fewer items that properly target the examinee sample compared to the well targeted item pool. This may have resulted in some item overexposure, but kept the precision at levels similar to the well-targeted item pool conditions.

2) Classification

On average, CAT exams using the poorly targeted item pool resulted in slightly more misclassifications in the 20% and 40% IPD conditions and slightly fewer

misclassifications in 60% IPD conditions; however, the differences between the two pools were not large enough to indicate a meaningful conclusion. Item pool targeting did not result in a substantial difference in classification accuracy of PPD examinees in the studied conditions. Again, this result was expected in a CAT since the exam algorithm tailors the test for the specific examinee ability.

4. **Research question 4: Does item pool targeting change the effect of IPD in low, medium, and high ability PPD examinees' classification into achievement levels differently?**

While the overall impact of IPD on classification accuracy did not vary by item pool targeting, one important question in this study was whether or not targeting changed the IPD impact on classification accuracy for different ability groups. I focused on low-ability examinees, since it can be assumed that under-resourced schools, less motivated teachers, and poorer instruction in real-life testing conditions would potentially lead to lower ability and less motivated examinees. Therefore, investigating the greatest potential IPD impact on lower-ability examinees was an important inquiry. For that purpose, I simulated a poorly targeted item pool, which had a mean item difficulty below the average ability of the examinee sample. This provided a CAT situation where lower ability examinees answered as many IPD items as possible. Despite the fact that they were targeted by more IPD items in the poorly targeted item pool than the well-targeted item pool, IPD did not practically change their classification results.

The results of the IPD impact on classification accuracy was governed by multiple sources. First, most low ability examinees' ability parameters were below the cut scores. An IPD impact in the harder direction did not result in a false decision, since most were already below the lower cut score of -0.7. An IPD scenario with items drifting in the easier direction (such as

item overexposure, specific content knowledge, or cheating), may give a clearer idea about potential impact of IPD on lower-ability examinees. Second, the distribution of the examinee sample partially contributed to many more medium-ability PPD examinees were misclassified than other ability groups. Thus, cut score values and the examinee sample distribution impacted the classification accuracy results of different ability groups. Further studies should examine IPD impact due to differences in opportunity to learn test content using different ability distributions (i.e., positively skewed to simulate a poor-performing district) and different cut scores. Designing simulations with various ability distributions and cut scores will be helpful to understand the consequential impact of such an IPD on potentially vulnerable groups such as students in schools with poorer resources, ill-equipped classrooms, and/or low-SES groups.

5. Research question 5: Holding all else constant, does the change in the impact of IPD over the levels of one factor depend on the level of another factor?

In this study, I included three factors to analyze their impact on estimation precision and classification accuracy. Those factors were item pool targeting, percentage of PPD examinees in the sample, and IPD items in the pool. To examine if one factor's impact depends on the level of another factor, all possible two-way interactions of the three factors (% of PPD examinees x % of PPD examinees, % of PPD examinees x item pool targeting, and % of IPD examinees x item pool targeting) were plotted. In Appendix A, RMSE plots within each IPD and PPD condition are displayed. Bias and MAD plots were not reported, since the pattern of their values were parallel with RMSE values. Both precision and classification accuracy improved when fewer IPD items and fewer PPD examinees were involved in both item pools (see Figures 1 to 6 and Table 9). The pattern of the impact of two factors, the percentage of IPD and the percentage of PPD examinees, on precision and classification were very similar to when one

factor was held constant in the two item pools. Overall, the plots of any two factors within a level of the third factor had the same slope, indicating the IPD impact over the levels of one factor was independent from the level of another factor in this study (see Appendix A and Appendix B).

C. Evaluation of Person Fit

One advantage of a simulation study is that factors causing misfit are known. In real-life testing conditions, the reasons for any unexpectedness in the responses should be sufficiently diagnosed to maintain validity. To address the issue of person misfit, numerous person fit indices and statistical approaches were used to highlight unexpected response behaviors. The unexpectedness mostly occurred due to examinees answering easier items incorrectly or harder items correctly, given their ability level. Examinees may become careless or feel fatigue and as a result, may miss an easier item for their ability level. Likewise, low ability examinees may have special knowledge about some harder items' content or they may simply guess. Traditional fit statistics are derived to diagnose such behaviors and work well in fixed-item tests where every examinee responds to same sets of items (J. Stahl, personal communication, November 17, 2016).

On the other hand, one of the most advantageous feature of CAT is targeting examinees' ability levels to maximize information and minimize error. A consequence of CAT's targeting and item selection features is that commonly used misfit statistics may not function effectively in CAT since it is less likely that a person answers easier or harder item for his/her ability level (J. Stahl, personal communication, November 17, , 2016). Aligned with the measurement model and estimation approach used in data generation and analyses in this study, I utilized WINSTEPS mean-squared (MNSQ) person fit statistics to detect misfitting examinees. My goal was to investigate if the misfitting examinees were among the PPD examinees. As reported in Chapter

four, the misfit statistics results were inconclusive, varying within a small range around the statistic's expected value (1.00). Therefore, the person fit findings of this study were similar to earlier claims that traditional fit statistics are not informative to detect person misfit in CAT (Meijer & van Krimpen-Stoop, 2010; Reise & Due, 1991).

To date, researchers have sought alternative methods to detect person misfit in CAT (J. Stahl, personal communication, November 17, 2016). These methods included predicted percent correct analysis, the Wald-Wolfowitz run test, and regression analysis (Meijer & van Krimpen-Stoop, 2010; J. Stahl, personal communication, November 17, 2016). In the percent correct analysis approach, expected probability of correct response criteria is used to estimate the percent of correct responses that an examinee would give. Then, each examinee's observed percent correct is compared to the average percent correct of the examinee sample. The second alternative approach, the Wald-Wolfowitz run test, was specifically developed for variable-length CAT exams. An expected mean and standard deviation of run changes in an exam are calculated based on a 50% correct assumption. The deviation from the expected number of run changes (i.e. incorrect to correct, correct to incorrect) is calculated through a *z* score. If the number of run changes is significantly lower or higher than expected, further investigation of misfit should be conducted (J. Stahl, personal communication, November 17, 2016). The last approach, regression analysis, was developed to be used as graphical evidence in conjunction with other statistical evidence of misfit. Ability estimates after each response are updated and used as a dependent variable while the administration order of the items are used as an independent variable to plot a regression line. The slope of the regression line is expected to approach zero as an examinee moves to the end of the CAT exam since the standard error of measurement gets smaller and ability estimates tends to stabilize. Any significant deviance from

a zero slope is considered as evidence of misfit (J. Stahl, personal communication, November 17, 2016). These alternative approaches that specifically focus on misfit caused by construct irrelevant sources, such as lack of motivation, item pre-knowledge, and warm-up anxiety, were evaluated in CAT. Among these sources of misfit, item pre-knowledge was found to be the most challenging to detect, since it can occur at any part of the test and be easily blended with other sources of misfit such as fatigue towards the end of the test or start up effect at the beginning (Meijer & van Krimpen-Stoop, 2010; J. Stahl, personal communication, November 17, 2016). This study elaborates on a specific source of a misfit (lack of knowledge on a certain content by a certain group of examinees) in a fixed-length CAT. My findings and previous studies' findings supported that it is even more challenging to detect misfit in a subgroup of examinees and the problem of detecting person misfit in CAT still exists. Therefore, further statistical approaches should be developed to detect misfit that occurs due to changes in examinee knowledge that can be attributable to various factors such as curricular and instructional differences and impact only unknown subgroups of examinees in CAT.

D. Strengths and Limitations

1. Strengths

The main strength of this study is that it elaborates on a drift problem, which has not been researched in CAT; particularly in a K-12 context. Some studies examined curricular, practice, and instructional differences as sources of changes in ability and knowledge and resulted in IPD in fixed-item tests (Bock, Muraki, & Pfeiffenberger, 1988; DeMars, 2004b; Fitzpatrick, 1992; Mislevy, 1982); however, these studies mostly analyzed multi-directional drift, where some items become easier while other items become harder. In addition, IPD in those studies affected all examinees who took the test. Studies analyzing impact of IPD due to changes

in examinee ability and knowledge that can be explained by similar factors in CAT are relatively limited than the studies using fixed-item test data. They looked at the issue as an item-overexposure problem, where IPD provided an unfair advantage to a group of examinees (Bergstrom, Stahl, & Netzky, 2001; Guo, 2009; Han & Guo, 2011). On the other hand, curricular, instructional, infrastructural and practice changes across time and test occasions may have resulted in disadvantages for some examinees (Kingsbury & Wise, 2011). This current study elaborated on the potential IPD impact in a newly-implemented CAT context where curricular and instructional differences, de-emphasized content, poor test practice, and lack of sufficient infrastructural resources across groups, settings, and time are of concern (Ash, 2008; Babcock & Albano, 2012; Kingsbury & Wise, 2011).

Another major strength of the current study is how IPD is exposed to the examinees. Previous simulation studies in CAT assumed that IPD impacted all examinees in the same way (Guo, 2009; Risk, 2015; Wei, 2013); however, their assumption may not hold in situations where group of examinees' abilities change over time due to person-centered reasons such as cheating, differential exposure to content, lack of access to learning materials, differential practice, or curricular differences. Those person-centered factors lead to drift in person ability rather than drift in item difficulty. Consequently, assuming items' difficulties change at the same degree for all examinees is not realistic in such cases. This current research explored drift as a matter of change in person ability and called this "person parameter drift." This was a new IPD investigation that can be applied to real life educational CAT exams.

Another strength of this study is the simulated item and person parameters were grounded in available literature and an operational K-12 CAT exam. Although I was not able to access the actual educational item bank parameters while designing the study, I obtained the mean item

difficulty value as well as content and test specifications including the number of items, the stopping rule, and selection of the first item of a newly-implemented K-12 CAT exam, which were adapted to my study. This helped my study approximate a real K-12 CAT exam as much as possible.

Lastly, I based my evaluation criteria and the manipulated factors from previous IPD studies, particularly in CAT. This provided me a guideline based on the literature while evaluating practical IPD consequences that can happen in real world CAT exams.

2. Limitations

Along with this study's strengths, there are number of limitations. First of all, only one exam administration was simulated and IPD impact was examined on that data; however, it is very likely that IPD may not be able to be captured in one administration and may become more pronounced by affecting more examinees over time (Han & Guo, 2011). Further studies should examine the longitudinal IPD impact in CAT, particularly in the form of PPD due to potential factors such as infrastructural, curricular, instructional, and practice differences across groups, time and administrations.

Another limitation is that I simulated a specific IPD condition, where all IPD items are in one content area and are all biased in one direction. As stated by Han and Sireci (2007), learning effects may make items easier or harder, depending on the emphasis over time. In real life testing, there may be IPD situations where some items become easier for some examinees due to practice or content overexposure while the others become harder for another examinee subgroup due to poor teaching or lack of access to learning opportunities (i.e. skipping school on a certain day when content is taught) in the same item pool. Thus, follow-up studies should explore

multidirectional drift impact that can happen due to changes within examinee ability in the same test.

The drift magnitude was selected within the range of 0.5-1.0 logits, since change in item parameters less than 0.5 logits is usually accepted as random fluctuation (Han & Guo, 2011). Setting a range from 0.5 to 1.0 logits, as in this study, may somewhat limit consequential IPD impact detection, particularly when items with IPD less than 0.5 logits are operationally used across time points and their impact are accumulated over time in an item bank. Therefore, follow-up studies should investigate any longitudinal impacts of similar IPD in smaller magnitudes, since such impact can become consequential in the long run.

Finally, as with many other simulations studies, this study's findings should not be generalized outside the studied testing conditions. The study simulated an explicit IPD situation that can happen in a K-12 CAT exam used for placing examinees into performance levels. Thus, the generalizability of the findings is limited to the specific exam and IPD conditions studied.

E. Implications

In this chapter, I highlighted the current study's key findings and provided some practical implications for testing organizations, schools, and districts. As indicated before, CAT use in K-12 is still developing and there are still considerations and critics regarding the extent that test results are used (Way et al., 2010). In order to ensure the most valid test results use for each examinee, testing organizations should carefully handle IPD by employing one of many IPD detection methods available in the literature. On the other hand, it is very important to address the causes of IPD before it becomes problematic in a CAT. IPD due to changes in ability that can be attributable to infrastructural, curricular, instructional, and practice differences are hard to detect and handle in practice, since it affects each examinee differently (Han & Guo, 2011).

Existing person fit indices were found ineffective to detect aberrant behaviors due to IPD in CAT (Meijer & van Krimpen-Stoop, 2010). Therefore, states, districts, and schools should take great effort to provide equal opportunity to learn for every examinee taking the same test. Additionally, they should also make every effort to address construct irrelevant sources of IPD, such as poor access to resources, computer illiteracy, and English as a second language (ESL) issues, which can disadvantage certain group of examinees in CAT.

After providing every examinee an equal opportunity to learn and perform on the exam, testing organizations should carefully address psychometric and technical issues while developing and maintaining item pools. As indicated by the classification accuracy results of the study, most misclassifications occurred around the cut scores by the medium ability group of examinees. Thus, testing companies should focus on item maintenance, particularly those items with difficulties around the cut scores in achievement tests. Testing companies should also make sure there are sufficient numbers of items targeting each performance level in the item pool to reduce misclassifications as a result of measurement error.

F. Suggestions for Future Research

Although this study shed light on some problems that may be associated with one possible IPD situation in CAT, there are still many avenues of research on IPD impact that needs to be investigated. Future studies might analyze IPD impact on measurement precision and classification accuracy using different amounts and magnitudes of drift. For this current study, I used a hypothetical sample size of 500 to simulate an IPD situation in a small district; other studies may use larger sample sizes to simulate a similar IPD scenario with nested models (i.e., districts nested in states). My examinee sample followed a normal distribution similar to most of the simulation studies investigating IPD (Han, Wells, & Sireci, 2012; Veerkamp &

Glas, 2000; Wells, Subkoviak, & Serlin, 2002; Zhang, 2014). On the other hand, it was suggested that in some assessment types, such as certification exams, ability distributions are often skewed (Witt et. al, 2003). This can apply to some of the educational testing situations where the majority of examinees in the sample are performing poorer or higher than the norm. Therefore, further simulation studies investigating IPD impact due to changes in examinee knowledge usually as a result of curricular, instructional, infrastructural, and practice differences may use skewed distributions to generate examinee samples to understand IPD impact on different ability compositions.

Another research possibility is to investigate IPD impact on classification accuracy with various cut scores. For this study, I collapsed the initial cut scores to reduce misclassification probability due to random error because the CAT exam I simulated was a 40-item test with 500 hypothetical examinees. Studies using larger sample sizes and variable-length CAT exams might investigate changing IPD impact due to differential access to content knowledge with more cut scores and/or with the same number of cut scores distributed differently across the ability continuum.

In this study, I simulated two item pools; one well-targeted to the mean examinee ability and another off-targeted pool in the easier direction. I obtained very similar results from the two item pools since the range of item difficulties were large enough to compensate for off-targeting. Thus, the question, “What is the impact of IPD on precision and classification accuracy when item pool targeting varies” still exists. Further studies should investigate the changing IPD impact with various off-targeting situations.

Mixed Rasch Models (MRM, Rost, 1990), an extension of traditional Rasch models, may be another future research direction to explore strategies for identification of PPD examinees. In

MRM, it is accepted that if unidimensionality does not hold for the entire data, there may be subgroups/classes (a latent class) in which a different unidimensional latent dimension is being measured. In a PPD situation, which this study was grounded on, responses of examinees from the same latent class may be related by factors such as instructional quality, school resources, etc. Future researchers may employ MRM to explain why different latent classes/PPD groups exist in the data.

In this study, I used a unidimensional Rasch model assuming that the simulated test included two content areas measuring the same domain. Recent research suggests that multidimensional computer adaptive testing (MCAT) allows more comprehensive detection of examinee performance than the unidimensional CAT approach (Chang, 2015). MCAT may better identify a drift condition that is caused by changes in examinee knowledge and skills on a single content area. Future studies may examine a similar drift problem on a content area from the MCAT perspective using an appropriate multidimensional IRT/Rasch model.

Another research idea for further studies is looking at the longitudinal IPD impact due to differential access to quality instruction, motivated teachers, resources, and lack of practice in CAT. All these factors may result in changes to examinee ability, which may not show their consequential impact completely in a single CAT administration; however, their effect may accumulate as more examinees are affected by IPD over time. In addition, IPD may impact linking and equating CAT exams, which aim to measure student growth across administrations. Thus, further studies may examine longitudinal IPD impact due to factors that result in changes to examinee ability over time.

Lastly, follow-up studies should develop person fit indices that work with CAT response data to detect aberrant response behaviors of individual examinees in various PPD and IPD

situations. A misfit index that works effectively with CAT would be helpful to identify IPD before it becomes consequential in operational tests or PPD when interpreting person ability estimates. While other possible person fit indices may pinpoint individual examinees, use of hierarchical linear models (HLM, Bryk & Raudenbush, 2002) may help with the identification of groups (i.e. gender, ethnicity, classrooms) whose scores are affected by IPD. Group level covariates may be added as level-2 predictors in an HLM model to test their significance. Once groups whose test performance is impacted due to group-related factors are identified, sources of IPD can be better addressed. Future studies may utilize HLM to investigate if a change in the interaction between items and a subgroup of examinee population can be attributed to group-related factors.

G. Conclusion

The most significant difference between this research and previous IPD studies in CAT is that a subset of flagged examinees received a known percentage of IPD items in the test. This allowed me to simulate a situation which can happen in a real life CAT exam; all items in one or more content areas drifting for a subgroup of examinees taking the test. Exposing flagged examinees to a known percentage of IPD items yielded results that have practical implications. The most relevant finding was that a small amount of IPD (i.e. 20% of the test) can matter substantially when examinees are affected by factors such as differential instruction or the lack of practice on a particular content area. Depending on the stakes linked to a CAT exam, even one person misclassified into a lower achievement level due to unequal opportunity to learn and perform may result in fairness issues. For instance, in a K-12 CAT exam in which examinee rank ordering is reported, a substantive number of misclassifications might result in significant changes in examinee rank order. As a result of a change in rank ordering, not only are individual

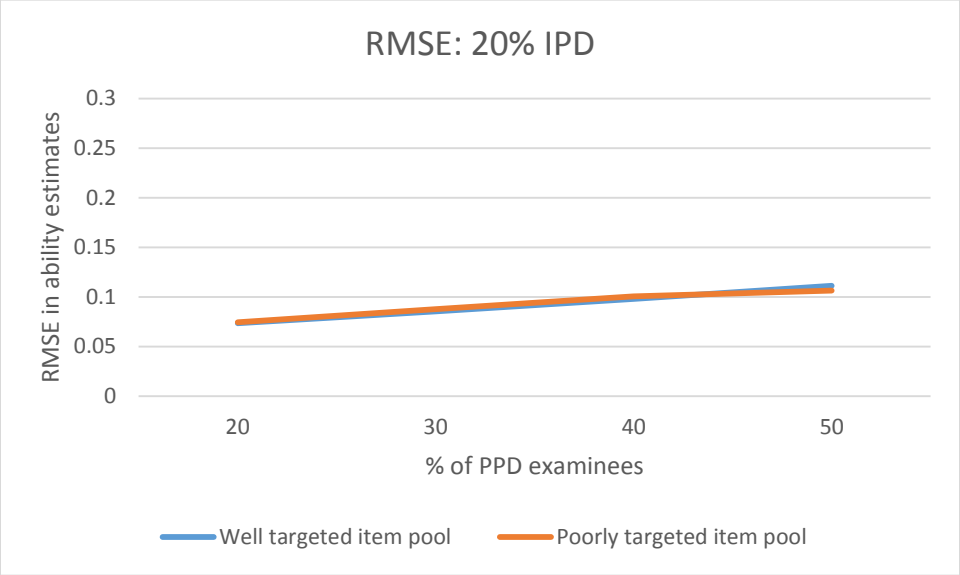
examinees impacted, but teachers and schools may also be effected by the test results, since CAT exams can be used for performance evaluations of teachers and schools in K-12 settings (Way et al. 2010). Thus, infrastructural, curricular, instructional, and practice differences across examinees should be carefully identified in CAT to make sure if every examinee is exposed to the content in a similar manner and given equal opportunity to demonstrate their true ability on the examination.

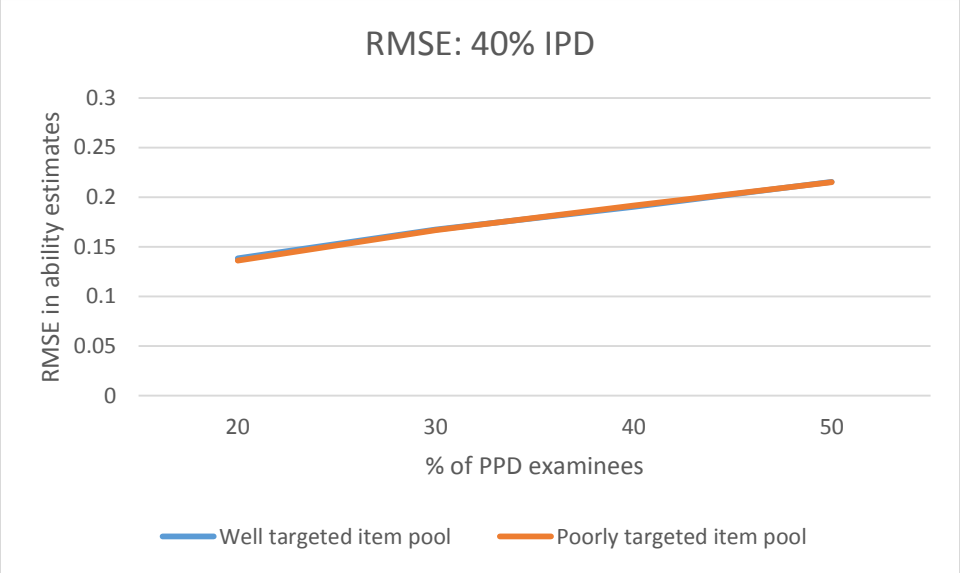
Another important finding of this research was that examinees whose abilities were around the cut scores were more likely to be misclassified due to IPD. A small deviation of an item's difficulty for an examinee may become consequential if the examinee's true ability is near a cut score. Therefore, testing companies should focus on maintenance of the items around the cut scores in order to better target the examinees whose abilities are around the cut scores in CAT.

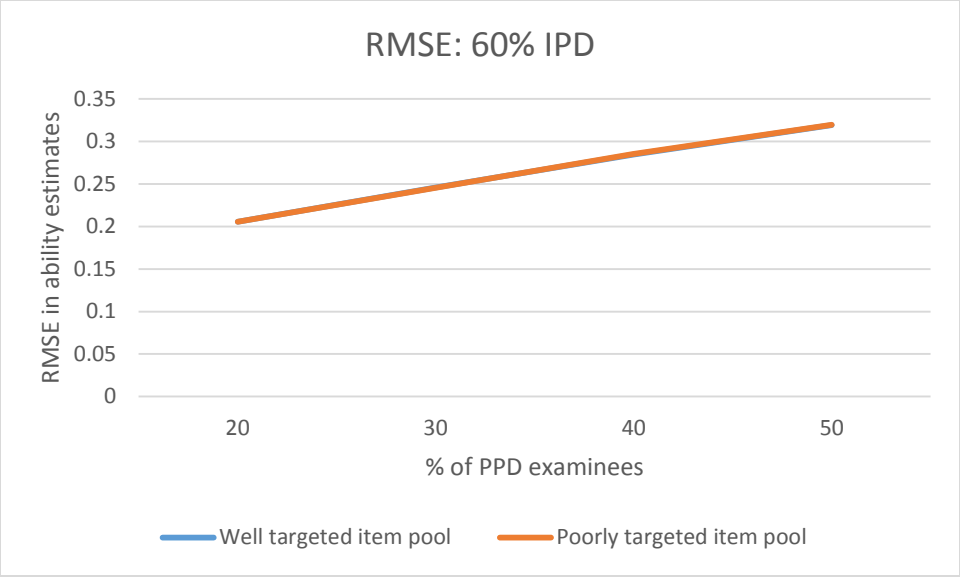
One promising finding of this study was CAT is robust to off-targeting as long as the variability of item difficulties in the bank are large enough for the ability of examinees taking the test. Although poor-targeting may result in overuse of some items, particularly the ones around the mean examinee ability, when an item pool has items with adequate variability in difficulty, testing organizations can be less concerned with targeting as a factor of IPD. While this research answers some of the existing questions about IPD impact in CAT, there are still many questions that need to be addressed by future studies, particularly in the contexts where CAT has newly been implemented (i.e., K-12).

APPENDIX A

RMSE PLOTS FOR THREE IPD % CONDITIONS

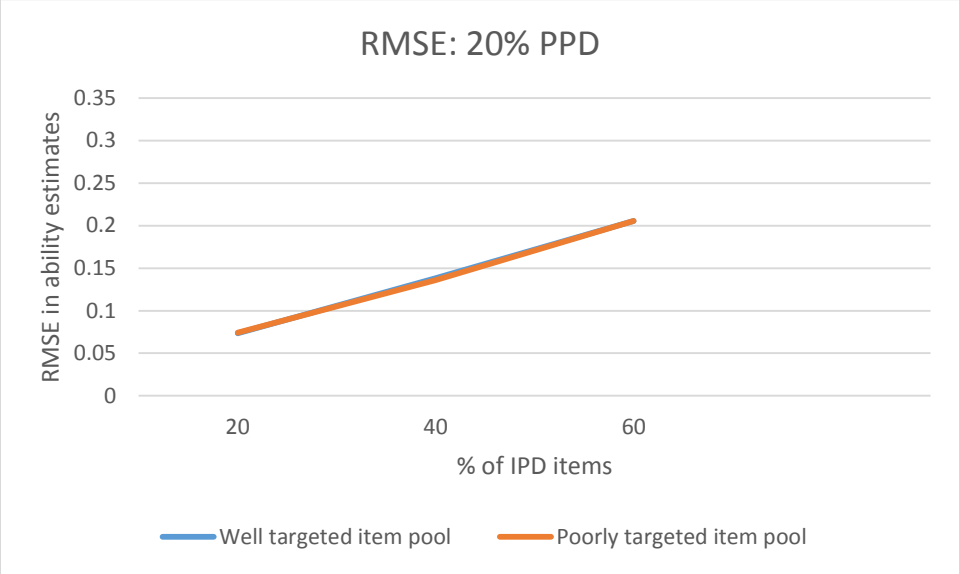


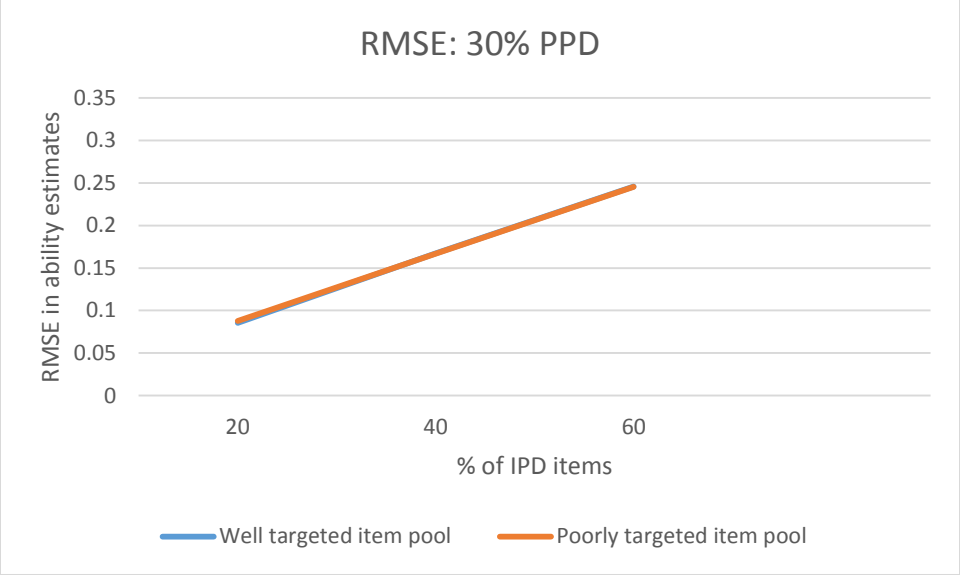


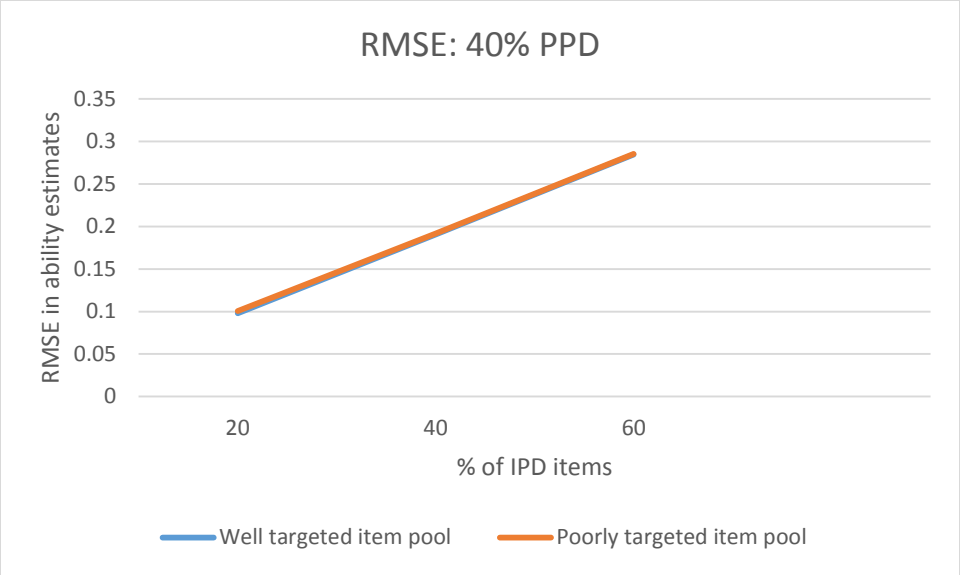

RMSE: 60% IPD
RMSE in ability estimates
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
20
30
40
50
% of PPD examinees
Well targeted item pool
Poorly targeted item pool

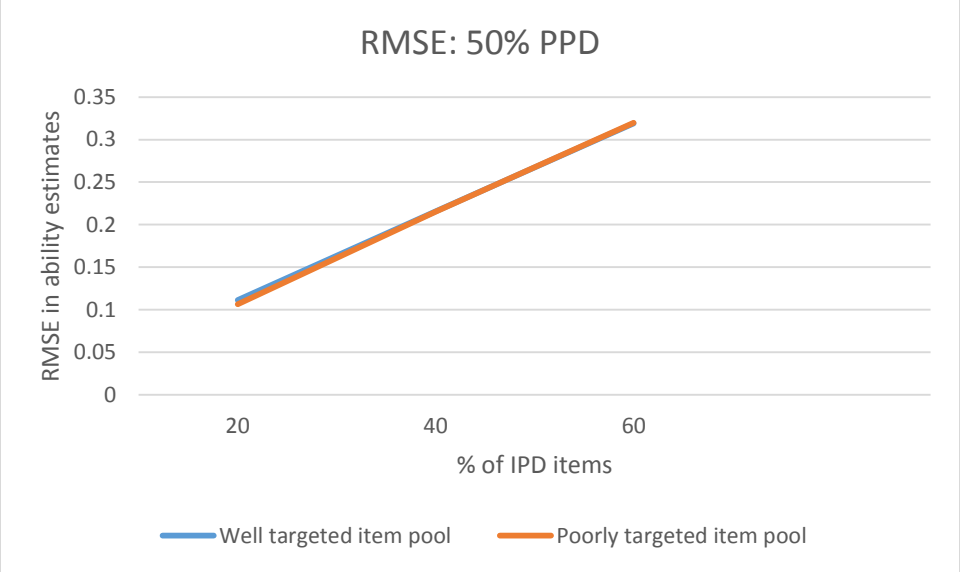
APPENDIX B

RMSE PLOTS FOR FOUR PPD% CONDITIONS






RMSE: 40% PPD
RMSE in ability estimates
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
20
40
60
% of IPD items
Well targeted item pool
Poorly targeted item pool


RMSE: 50% PPD
RMSE in ability estimates
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
20
40
60
% of IPD items
Well targeted item pool
Poorly targeted item pool

CITED LITERATURE

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Angoff, W. H., & Huddleston, E. M. (1958). *The multi-level experiment: A study of a two-level test system for the College Board Scholastic Aptitude Test.* (Statistical Report No. 58-21). Princeton, NJ: Educational Testing Service.

Ash, K. (2008, November, 19). Adjusting to test takers. *Education Week.* Retrieved from http://www.edweek.org/ew/articles/2008/11/19/13tech_ep.h28.html

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

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

Becker, K. (2013). CAT Simulator. Chicago, IL: Promissor.

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

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.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

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

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26,* 381-409.

Bugbee, Jr., A. C., & Bernt, F. M. (1990). Testing by computer: Findings in six years of use 1982-1988. *Journal of Research on Computing in Education, 23,* 87-100.

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.

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika, 80,* 1-20.

Chen, Q. (2013). Remove or keep: Linking items showing item parameter drift. Retrieved from ProQuest Digital Dissertations. (AAT 3560290)

Choppin, B. H. L. (1985). Lessons for psychometrics from thermometry. *Internal Journal of Educational Research, 9*(1), 9-12.

Choi, S. W., Grady, M. W., & Dodd, B. G. (2011). A New Stopping Rule for Computerized Adaptive Testing. *Educational and Psychological Measurement, 71*(1), 37-53.

Clark, A. (2013, September). *Review of Parameter Drift Methodology and Implications for Operational Testing.* Paper presented at the annual meeting of the National Conference of Bar Examiners, Chicago, IL.

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.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Davey, T., & Parshall, C. G. (1995, April). *New Algorithms for Item Selection and Exposure Control with Computerized Adaptive Testing.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Davey, T., & Nering, M. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 165–191*).* Mahwah, NJ: Lawrence Erlbaum.

Davis, M. R. (2012, October 15). Adaptive testing evolves to assess common-core skills. *Education Week.* Retrieved from http://www.edweek.org/dd/articles/2012/10/17/01adaptive.h06.html

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

DeMars, C. E. (2004b, April). *Item parameter drift: The impact of curricular area.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Deng, H., & Melican, G. (2010). *An investigation of scale drift for arithmetic assessment of ACCUPLACER®* (Report No. 2010-2). New York, NY: The College Board.

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

Flaugher, R. (2000). Item pools. In Wainer, H. (2000). *Computerized adaptive testing: A primer* (pp. 37-59). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Georgiadou, E. G., Triantafillou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment, 5*(8), 5-31.

Gershon, R. C., & Bergstrom, B. (1991, April). *Individual Differences in Computer Adaptive Testing: Anxiety, Computer Literacy, and Satisfaction.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Gershon, R. C., & Bergstrom, B. A. (1995, April). *Does cheating on CAT pay: NOT!* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Gershon, R. C. (2005). Understanding Rasch measurement: Computer adaptive testing. *Journal of Applied Measurement, 6*(1), 109-127.

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

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347-360.

Guernsey, L. (2000, August 6). An ever-changing course: taking admission tests on computer. *The New York Times.* Retrieved from http://www.nytimes.com/2000/08/06/education/an-ever-changing-course-taking-admissions-tests-on-computer.html?pagewanted=1

Guo, F., & Wang, L. (2005). *Evaluating scale stability of a computer adaptive testing system* (Report No. 05-12). Reston, VA: Graduate Management Admission Council®.

Guo, F. (2009, February). *Quantifying impact of compromised items in CAT.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200). New York, NY: Macmillan.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

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

Han, K. T., & Guo, F. (2011). *Potential impact of item parameter drift due to practice and curriculum change on item calibration in computerized adaptive testing* (Report No. 11-02). Reston, VA: Graduate Management Admission Council®.

Han, K., Wells, C. S., & Sireci, S. G. (2012). The impact of multidirectional item parameter drift on IRT scaling coefficients and proficiency estimates. *Applied Measurement in Education, 25,* 97-117.

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.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141-144). Washington DC: American Psychological Association.

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning.* Hillsdale, NJ: Erlbaum.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*(3), 16-25.

Huynh, H., & Meyer, P. (2010). Use of robust z in detecting unstable items in item response theory models. *Practical Assessment, Research, & Evaluation, 15*(2), 1-8.

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

Kelley, T. L. (1927). *Interpretation of educational measurements.* Yonkers, NY: World Book Company.

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,* 2-27.

Krathwohl, D. R., & Huyser, R. J. (1956). The sequential item test (SIT). *American Psychologist, 2,* 419.

Kreitzberg, C. B., & Jones, D. H. (1980). *An empirical study of the broad range tailored test of verbal ability* (Report No. 80-5). Princeton, NJ: Educational Testing Service.

Lange, R. (2007). Binary items and beyond: A simulation of computer adaptive testing using the Rasch partial credit model. In E. V.,Smith, Jr. & R. M. Smith, (Eds.), *Rasch*

Measurement: Advanced and Specialized Applications (pp. 148-181). Maple Grove, MN: JAM Press.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. (2000). *Computer-adaptive testing: A methodology whose time has come* (Report No. 89). Chicago, IL: MESA. Retrieved from http://www.rasch.org/memo69.pdf

Linacre, J. M. (2005). WINSTEPS Rasch measurement software. Available at www.WINSTEPS.com.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs, 7*, 1-84.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1,* 95-100.

Lord, F. M. (1980). Some how and which for practical tailored testing. In L. J. T. van der Kamp, W. F. Langerak & D. N. M. de Gruijter (Eds): *Psychometrics for educational debates* (pp. 189-206). New York: John Wiley and Sons.

Lord, F. M. (1981). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Luecht, R. L., & Sireci, S. (2011). *A review of models for computer-based testing* (Report No. 2011-2012). New York: The College Board.

Lunz, M.E. & Bergstrom, B. A. (1991). Comparability of decisions for computer adaptive and written examinations. *Journal of Allied Health, 20*, 15-23.

Masters, J. S., Muckle, T. J., & Bontempo, B. (2009, April). *Comparing Methods to Recalibrate Drifting Items in Computerized Adaptive Testing.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

McBride, J. R. (1988, August). *A Computerized Adaptive Version of the Differential Aptitude Tests*. Paper presented at the meeting of the American Psychological Association, Atlanta, GA.

McBride, J. R., & Sympson, J. B. (1985). The computerized adaptive testing system development project. In D. J. Weiss (Ed.), *Proceedings of the 1982 item response theory and computerized adaptive testing conference* (pp. 342-349). Minneapolis: University of Minnesota, Department of Psychology.

McCoy, K. M. (2010). *Impact of item parameter drift on examinee ability measures in a computer adaptive environment.* Retrieved from ProQuest Digital Dissertations. (AAT 3417367)

Meng, H., Steinkamp, S., & Matthews-Lopez, J. (2011, October). *Practitioner's approach to identify item drift in CAT*. Paper presented at the annual conference of the International Association for Computerized Adaptive Testing, Pacific Grove, CA.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*, (3rd ed.). Washington, D.C.: American Council on Education.

Meyers, J. L, Miller, G. E., & Way, W.D. (2009). Item Position and Item Difficulty Change in an IRT-based Common Equating Design. *Applied Measurement in Education, 22*(1), 38– 60.

Meyers, J. L, Murphy, S., Goodman, J., & Turhan, A. (2012, April). *The impact of item position change on item parameters and common equating results under the 3PL model.* Paper

presented at the annual meeting of the National Council on Measurement in Education, Vancouver, B. C.

Mislevy, R. J. (1982, March). *Five Steps toward Controlling Item Parameter Drift.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

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

National Council of State Boards of Nursing, Inc. (1991, July). *A psychometric comparison of computerized and paper-and-pencil versions of the national RN licensure examination*. Chicago IL: Author, Unpublished report.

National Research Council. (2007). Lessons learned about testing: Ten years of work at the National Research Council. Washington, D.C.

Nering, M. L., Davey, T., and Thompson, T. (1998, June). *A Hybrid Method for Controlling Item Exposure in Computerized Adaptive Testing.* Paper presented at the annual meeting of the Psychometric Society, Urbana, IL.

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

Pine, S.M. (1977). Applications of item characteristics curve theory to the problem of test bias. In D. J. Weiss (Ed.), *Applications of computerized adaptive testing: Proceedings of a symposium presented at the 18th annual convention of the Military Testing Association*

(Research Report No. 77-1. pp. 37-43). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research. (Expanded edition, 1980). Chicago: University of Chicago Press.

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

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice, 8*(3), 11-15.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York: Springer Science+Business Media.

Risk, M. C. (2015). *The impact of item parameter drift in computer adaptive testing (CAT).* (Unpublished doctoral dissertation). University of Illinois at Chicago, Chicago, IL.

Rost, J. (1990). Rasch models in latent class analysis: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14,* 271-282.

Rudner, L. (1998). An On-line, Interactive, Computer Adaptive Testing Mini-Tutorial. ERIC Clearinghouse on Assessment and Evaluation. Retrieved from http://echo.edres.org:8080/scripts/cat/catdemo.htm

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

Sands, W. A., & Gade, P. A. (1983). An application of computerized adaptive testing in U. S. Army recruiting. *Journal of Computer-Based Instruction, 10,* 87-89.

Segall, D. (1996). Multidimensional adaptive testing. *Psychometrika, 61,* 331-354.

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.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210.

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

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.

Store, D. (2013). Item parameter changes and equating: An examination of the effects of lack of item parameter invariance on equating and score accuracy for different proficiency levels. ProQuest Digital Dissertations. (AAT 3568922)

Swaminathan, H., & Gifford, J. A. (1983). Estimation in the three-parameter latent trait model. In D. Weiss (Ed.), *New Horizons in Testing* (pp. 13-29). New York: Academic Press.

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.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing*

Association, (pp. 973-977). San Diego CA: Navy Personnel Research and Development Center.

Taherbhai, H., & Seo, D. (2013). The philosophical aspects of IRT equating: Modeling drift to evaluate cohort growth in large-scale assessments. *Educational Measurement: Issues and Practice, 32,* 2-14.

The Code of Fair Testing Practices in Education. (1988). Washington, D.C.: Joint Committee on Testing Practices.

Vale, C. D. (1981). Design and implementation of a microcomputer-based adaptive testing system. *Behavior Research Methods and Instrumentation, 13,* 399-406.

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

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.

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 Associates.

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement, 38,* 19–49.

Ward, W. C. (1988). The College Board computerized placement tests: An application of computerized adaptive testing. *Machine-Mediated Learning, 2,* 271-282.

Way, W., Twing, J., Camara, W., Sweeney, K., Lazer, S., & Mazzeo, J. (2010). *Some considerations related to the use of adaptive testing for the Common Core assessments*. Retrieved from http://www.ets.org/Media/Home/pdf/AdaptiveTesting.pdf

Wei, E. X. (2013, August). *Impacts of item parameter drift on person ability estimation in multistage testing.* American Institute of CPAs Technical Report. Retrieved from http://www.aicpa.org/BecomeACPA/CPAExam/PsychometricsandScoring/TechnicalReports/DownloadableDocuments/Wei_Drift_2013.pdf

Weiss, D. J. (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Weiss, D. J., & Kingsbury, G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 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,* 77–87.

Wells, C. S., Hambleton, K. R., Kirkpatrick, R., & Meng, Y. (2014). An examination of two procedures for identifying consequential item parameter drift. *Applied Measurement in Education, 27*(3), 214-231.

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. (1997, March). *Overview of Practical Issues in a CAT Program.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL.

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.

Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement, 40,* 307-330.

Wollack, J. A., Sung, J. H., & Kang, T. (2006, April). *The impact of compounding item parameter drift on ability estimation.* Paper presented at the annual meeting of the National Council of the Measurement in Education, San Francisco, CA.

Wright, B. D., & Stone, M. (1978). *Best Test Design.* Chicago: MESA Press.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis.* Chicago: MESA Press.

Wright, B. D., & Bell, S. R. (1984) Item banks: What, why, how. *Journal of Educational Measurement, 21,* 331-345.

Ye, M., & Xin, T. (2014). Effects of item parameter drift on vertical scaling with the nonequivalent groups with anchor (NEAT) design. *Educational and Psychological Measurement, 74*(2), 227-235.

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.

Zhang, J. (2014). A sequential procedure for detecting compromised items in the item pool of a CAT system. *Applied Psychological Measurement, 38*(2), 87-104.

VITA

Beyza Aksu Dunya
University of Illinois at Chicago
Email: baksu2@uic.edu

EDUCATION

2017 (Expected)	**University of Illinois at Chicago** Ph.D. in Educational Psychology Emphasis: Measurement, Evaluation, Statistics, and Assessment (MESA) Advisor: Everett Smith, Ph.D. GPA: 3.83
2012	**Boston College** Master of Education in Educational Research, Measurement, and Evaluation Advisor: Larry Ludlow, Ph.D. GPA: 3.88
2008	**Istanbul University** Bachelor of Science in Mathematics

RELEVANT ADDITIONAL EDUCATION

2015	**University of Illinois at Chicago, Chicago, IL, USA** Graduate Certificate in Interdepartmental Concentration in Survey Research Methodology

RESEARCH EXPERIENCE

2017-	**Visiting Clinical Assistant Professor / Assessment Coordinator** University of Illinois at Chicago University Library
2012 – 2017	**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. 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. Assist with developing research designs for both faculty and student research projects.
2012 – 2015	**Research Assistant** University of Illinois at Chicago, Chicago, IL Institute of Educational Sciences (IES) Grant of Everett Smith Investigate psychometric properties of a nonverbal accuracy measure used in evaluating social emotional learning of children.

2010 – 2011 **Research Assistant**
Boston College, Chestnut Hill, MA
Teaching American History (TAH) Project
Open-code interview data of the "Teaching American History (TAH)" workshop series for the fourth year technical report. Track and analyze key data of participating teachers in Lowell, MA.

TEACHING EXPERIENCE

2008-2009 **Mathematics Teacher**
Pertevniyal High School, Istanbul, Turkey

Assist senior mathematics teachers as a student teacher by participating classroom instructions. Develop formative and summative assessments.

PUBLICATIONS

Kostovich, C., **Dunya, A. B.,** Schmidt, L., & Collins, E. (2016). A Rasch Rating Scale Analysis of the Presence of Nursing Scale-RN. *Journal of Applied Measurement, 7(4).*

Kartal, O., **Dunya, A. B.,** Diefes-dux, H., A., Zawojewski, J. S. (2016). The relationship between students' performance on conventional standardized mathematics assessments and complex mathematical modeling problems. *International Journal of Research in Undergraduate Mathematics Education, 2(1),* 239-252.

INTERNATIONAL CONFERENCE PRESENTATIONS

Dunya, A. B., Smith, E., & Mckown, C. (2017, April). Dimensionality and Item Hierarchy of the 12-item Theory of Mind (ToM) Scale. Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

Dunya, A. B. (2016, April). *Practice Differences and Item Parameter Drift in Computer Adaptive Testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Dunya, A. B. (2015, November). *Item Parameter Drift in Computer Adaptive Testing.* Paper presented at the Ideas in Testing Research Seminar, Chicago, IL.

Dunya, A. B. (2015, April). *Impact of Item Parameter Drift on an Adaptive Test*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Dunya, A. B., & Ozyer, K. K. (2015, April). *Validity of the Adapted Online Self-Regulated Learning Questionnaire.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Kartal, O, & Dunya, A. B. (2015, April*). An Alternative Mathematics Assessment Mode: Mathematical Modeling.* Paper presented at the annual meeting of the National Council of Teachers of Mathematics, Boston, MA.

Kostovich, C., Dunya, A. B., Schmidt, L., & Collins, E. (2015, April). *A Rasch Rating Scale Analysis of the Presence of Nursing Scale-RN.* Paper presented at the International Outcome Measurement Conference, Chicago, IL.

Dunya, A. B., Smith, E., Mckown, C., & Russo, N. (2014, April). *Applying DIF Detection Methods to a Nonverbal Accuracy Assessment.* Paper presented at the International Objective Measurement Conference, Philadelphia, PA.

Dunya, A. 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.

Ozyer, K. K., Aksu, B. (2013, June). *Computerized ToEFL Exam Acceptance.* Paper presented at the annual meeting of the European Association for Language Testing and Assessment, Istanbul, Turkey.

PROFESSIONAL SERVICE

2014-Present	**Proposal Reviewer** National Council on Measurement in Education Annual Conference
2015-Present	**Editorial Advisory Board** Handbook of Research on Training Evaluation in the Modern Workforce
2015-Present	**Manuscript Reviewer** UIC Interdisciplinary Undergraduate Research Journal
2016- Present	**Manuscript Reviewer** Journal of Measurement and Evaluation in Education and Psychology

UNIVERSITY/COLLEGE/DEPARTMENT SERVICE

2013-Present	**Graduate student representative** University of Illinois at Chicago, College of Education, Chicago, IL

PROFESSIONAL AFFILIATIONS

American Educational Research Association (AERA)
National Council on Measurement in Education (NCME)
International Association for Computerized Adaptive Testing (IACAT)

STATISTICAL SOFTWARE SKILLS

SAS, SPSS, HLM, R, WINSTEPS, FACETS, ConQuest, RUMM, BILOG-MG, STATA, Qualtrics, GENOVA, SimulCAT, MPlus, AMOS

RELEVANT DOCTORAL COURSEWORK

Computer Adaptive Testing by Dr. John Stahl
Item Response Theory by Dr. Lixiong Gu
Educational Measurement by Dr. Yue Yin
Rating Scale and Questionnaire Design and Analysis by Dr. Everett Smith

Approaches to Analyzing Rating Data by Dr. Carol Myford
Hierarchical Linear Models by Dr. George Karabatsos
Multivariate Analysis of Educational Data by Dr. Yue Yin
Simulation and Programming with R by Dr. Hakan Demirtas
Analyzing Survey Data by Dr. Timothy Johnson
Survey Nonresponse by Dr. Allison Holbrook

LANGUAGE PROFICIENCY

Turkish – Native
English – Advanced proficiency