Assessment of Item Parameter Drift of Known Items in a

University Placement Exam

by

Janet Krause

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Approved February 2012 by the
Graduate Supervisory Committee:

Roy Levy, Co-Chair
Marilyn Thompson, Co-Chair
Joanna Gorin

ARIZONA STATE UNIVERSITY

May 2012

ABSTRACT

This study investigated the possibility of item parameter drift (IPD) in a calculus placement examination administered to approximately 3,000 students at a large university in the United States. A single form of the exam was administered continuously for a period of two years, possibly allowing later examinees to have prior knowledge of specific items on the exam. An analysis of IPD was conducted to explore evidence of possible item exposure. Two assumptions concerning items exposure were made: 1) item recall and item exposure are positively correlated, and 2) item exposure results in the items becoming easier over time. Special consideration was given to two contextual item characteristics: 1) item location within the test, specifically items at the beginning and end of the exam, and 2) the use of an associated diagram. The hypotheses stated that these item characteristics would make the items easier to recall and, therefore, more likely to be exposed, resulting in item drift. BILOG-MG 3 was used to calibrate the items and assess for IPD. No evidence was found to support the hypotheses that the items located at the beginning of the test or with an associated diagram drifted as a result of item exposure. Three items among the last ten on the exam drifted significantly and became easier, consistent with item exposure. However, in this study, the possible effects of item exposure could not be separated from the effects of other potential factors such as speededness, curriculum changes, better test preparation on the part of subsequent examinees, or guessing.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

INTRODUCTION

Purpose of the Study

The purpose of this study was to examine the possibility of item parameter drift (IPD) in a calculus placement examination administered to approximately 3,000 university students over a period of two years at a large university in the southwestern United States. The same form of the exam was in continuous use during this time period. As a result, items could have been exposed, allowing later examinees to have prior knowledge of specific items on the exam. This study conducted an analysis of IPD as evidence of possible item exposure. Special consideration was given to two contextual item characteristics: item location within the test and the use of an associated diagram, to evaluate whether these characteristics contributed to the speed and/or severity of exposure effects on the performance of these items. BILOG-MG 3 (Zimkowski, Murake, Mislevy, & Bock, 2003) was used to calibrate the items and assess for IPD over the two years.

Item level data from a 60-item, paper-based, multiple choice exam were used in this study. The data may be somewhat unique within the context of current testing practice as the same form of the exam was used on a continuous basis over a two year period of time. With widespread access to the World Wide Web and the ability to easily share information between large numbers of students, a single form of an exam is seldom used on a consistent basis over an extended period of time. In addition, students were allowed to re-test as many times as desired with the limitation that a student could test only one time per day. Security of the exam was maintained by storing exam materials in a controlled location, administering the exam in a proctored environment, preventing examinees from taking any scratch paper or other materials from the testing room, and providing no feedback on performance other than total score. No further procedures were in place to prevent examinees from recalling specific test items and sharing that information with future examinees or course instructors.

Hypotheses

This study was intended to test three hypotheses in an item response theory (IRT) framework:

1. Evidence of IPD was expected to be exhibited by items on this exam due to item exposure. Specifically, it was anticipated that the items became easier over time.
2. When compared to items in the middle of the exam, IPD was expected to be more pronounced in the first and last items of the exam. Consistent with the theory of serial position effect as posited by Hermann Ebbinghaus and further developed by Glanzer and Peters (1962), it is expected that examinees are more likely to recall items at the beginning and end of the exam. Item recall is essential in order for examinees to share information with subsequent examinees. Also, the exam may have had some component of speededness which was expected to become less pronounced over time as more items were exposed. As a result, it was

expected that items at the end of the test became easier over time as more candidates had the opportunity to complete these last items.

3. Despite instructions not to write in the test booklet, many examinees drew on the diagram associated with item 15. Given that so many examinees demonstrated direct interaction with this specific item, it was anticipated that candidates might have better recall of this item. Again, it was expected that item 15 would exhibit IPD associated with exposure more quickly or more profoundly than similar items with a diagram on which few examinees wrote.

Chapter 2

LITERATURE REVIEW

Item Parameter Drift

Measuring educational attainment or achievement over time involves certain measurement problems. Goldstein (1983) stated:

> There is a clear duality between attributing change in an item parameter value to a change in the population response or in the characteristics of the item. The point, however, is whether the item should be regarded an equally fair assessment of the education system at each time, and it is here that judgment as well as empirical evidence is needed (p. 374).

Differential, or systematic, change in item parameter values over time is known as IPD. As an example, an analysis of the College Board Physics Achievement Test administered five times over a 10 year period found considerable changes in item parameters (Bock, Muraki, & Pfeiffenberger, 1988). The authors hypothesized that various factors could contribute to IPD: terminology may become outdated, concepts may be covered more actively in the mainstream media, curriculum emphases may change, or specific test items may become exposed to future examinees.

IRT models estimate the log-odds of the probability of a correct response for any given examinee (*i*) on a particular item (*j*) as a function of the latent trait

(θ) and one or more item parameters. A three parameter logistic (3PL) model for a dichotomously scored item is expressed as:

$$P(X_{ij}=1 \mid \theta_i, b_j, a_j, c_j) = c_j + (1 - c_j)\{1 + \exp[-a_j\ (\theta_i - b_j)]\}^{-1} \qquad (1)$$

where X_{ij} is the scored value (1 correct, 0 incorrect) for the response from examinee *i* to item *j*, θ_i is the latent trait of examinee *i, b*$_j$ is the difficulty parameter for item *j, a*$_j$ is the discrimination of item *j*, and c_j is the lower asymptote of item *j*. This equation creates a sigmoidal curve, bounded by *c* and 1, that is symmetrical around the point of inflection. In dichotomously scored items, the *b* parameter is the θ value at the point of inflection. The *a* parameter is proportional to the slope of the item response curve at the point of inflection (slope = .25 a_j [1 - c_j]). The *c* parameter is the lower asymptote of the curve which may not approach zero at the lowest levels of θ because of guessing or chance. Special cases of this model include the two parameter logistic (2PL) model, in which for all items c_j = 0, and the one parameter logistic (1PL) model, in which c_j = 0 and the a_j are equal for all items.

From a theoretical perspective, item parameter drift is problematic because item invariance is a key assumption in IRT, differentiating it from classical test theory (Hambleton, Swarminathan, & Rogers, 1991). Item parameter invariance means that estimates for the difficulty (*b*) parameter, the discrimination parameter (*a*), and the lower asymptote (*c*) for an item administered to two groups of examinees from the same population should be the same after allowances for sampling errors and scaling adjustments. This quality is essential to many IRT applications such as computer adaptive testing and test equating. If item parameter estimates vary across groups of examinees, the item is deemed to be performing differently and thus providing different information from these groups of individuals. As summarized by Donogue & Isham, “IPD threatens the validity of scores by introducing trait-irrelevant differences over time” (1998, p. 49).

Exposed items are problematic because later examinees would have specific, prior knowledge of test content. The item may no longer measure the intended construct, but rather an examinee’s ability to use outside resources to determine a correct response and/or the examinee’s ability to remember the correct response when the item is presented in the test. Because this item is no longer measuring the intended construct, the item response curve and its associated item parameters, as determined in an IRT framework, may change. Exposed items are likely to become less difficult, which is reflected in reduction in the item’s *b* parameter over time (Bock, Muraki, & Pfeiffenberger, 1988). In addition, if lower ability examinees capitalize on an exposed item at a disproportionately higher rate than other examinees, they have an increased probability of a correct response. It is expected that these higher likelihoods of a correct response on the part of the lower ability examinees would be reflected in higher *c* parameter values over time as well.

Techniques to Identify Item Parameter Drift

IPD is a special case of differential item functioning (DIF). In both instances, an item is not performing the same way across groups of examinees. DIF concerns differences in item performance in different examinee sub-groups and typically refers to group distinctions at a single point in time (e.g. ethnicity or gender); IPD refers to changes in item performance at different points in time. Many of the techniques originally proposed to identify DIF can be applied to problems of IPD as well by treating examinees at different time points as the subgroups in a DIF paradigm (Donoghue & Isham, 1998).

Several approaches for the identification of DIF have been studied for application to the problem of IPD in an IRT framework. Donoghue and Isham (1998) assessed a 3PL model over two time periods in a Monte Carlo study designed to simulate a large-scale assessment scenario. The goal of the study was to link the results from the two administrations of the assessment in the presence of IPD. Measures were compared on their ability to effectively identify items that were exhibiting drift. They compared several different measures including: 1) Lord's χ^2 statistic (Lord, 1980), 2) a number of methods devised by Kim and Cohen, and by Raju to measure the area between two item response curves (Kim & Cohen, 1991; Raju, 1988; Raju, 1990), 3) statistics based on Mantel-Haenszel (Holland & Thayer, 1986), and 4) several χ^2 measures generated by the NAEP BILOG/Parscale program. Donoghue and Isham found Lord's χ^2 measure to be most effective in identifying the exhibited drift, but only in situations where the lower asymptote, *c*, was held constant.

The Donoghue and Isham study was limited to comparing two groups, but situations involving measurement over time often call for comparisons of more than two groups. In their 1988 study, Bock, Muraki, and Pfeiffenberger studied the item parameter drift in a national physics exam administered over a ten year period of time. They proposed a method for detecting, estimating, and minimizing the effects of item parameter drift in exam item pools for long-term testing programs. The authors found evidence to support several conclusions about IPD in this particular study that may generalize to other situations (pp. 284-285):

1. Differential drift can occur over a period of years in a nationally administered educational test.
2. Drift will affect item locations (difficulties) much more strongly than item slopes (validities).
3. Differential drift of item locations are relatively steady in large populations and are describable as a linear function of time.
4. A linearly time-dependent item response model can describe educational test data accurately enough to support an IRT-based system for maintaining consistent scales of measurement over an extended period as items are retired and replaced in the item pool.

This study by Bock, Muraki, and Pfeiffenberger helped to provide the foundation for the procedures implemented in the BILOG-MG program. In modeling item parameter drift, the program makes certain assumptions regarding IPD, including:1) drift will be evidenced first in the difficulty parameter, 2) only the item x time interaction is considered drift, and 3) this interaction can be expressed as a low degree polynomial. In the BILOG-MG drift model, the discrimination and lower asymptotes are held constant over time while the difficulty parameter is allowed to vary based on a polynomial trend model in which the degree may take on a value up to one less than the number of time points.

In a 3PL model, linear location drift over time is expressed as follows:

$$P(X_{ij}=1 \mid \theta_i, b_j, a_j, c_j, \delta_j, t_k) = c_j + (1 - c_j)\{1 + \exp[-a_j\,(\theta_i - b_j - \delta_j t_k)]\}^{-1} \quad (2)$$

where t_k is time at occasion k measured from an arbitrary origin (the time period selected as the reference) and δ_j is a coefficient for the item capturing linear drift in the item's difficulty over time. The values for δ_j are constrained, as follows:

$$\sum_{j=1}^{n} \delta_j = 0 \quad (3)$$

Quadratic location drift over time is expressed as follows:

$$P(X_{ij}=1 \mid \theta_i, b_j, a_j, c_j, \delta_{1j}, \delta_{2j}, t_k) = c_j + (1 - c_j)\{1 + \exp[-a_j\,(\theta_i - b_j - \delta_{1j}t_k - \delta_{2j}t_k^2)]\}^{-1} \quad (4)$$

where δ_{1j} and δ_{2j} are coefficients for the item capturing linear and quadratic drift in the item's difficulty over time, respectively. Again, the values of the linear drift coefficients are constrained, as follows:

$$\sum_{j=1}^{n} \delta_{1j} = 0 \quad (5)$$

and δ_{2j} is unconstrained. Higher order models would be similarly expanded from equation (4).

In the drift model, the discrimination (*a*) and lower asymptote (*c*) parameters are held constant across the different time periods. The drift parameters estimated by BILOG-MG are the coefficients of the time parameter, notated by δ, and, conceptually, represent the adjustment to the location parameter (*b*). The constraint on the first power of the drift parameter (see Equations (3) and (5)), prevents drift in a single direction of all items on an exam. In other words, the model does not allow all the items to become easier, nor may they all become more difficult.

In her study, DeMars (2004) compared three methods of detecting IPD: the procedure used in BILOG-MG, the CUSUM procedure implemented by Veerkamp and Glas (2000) and a modification of Kim, Cohen, and Park's (1995) χ^2 test. The study simulated item parameters for a 3PL model for a test of 100 items in which 10% of the items exhibited drift and 90% did not. Various patterns and degrees of parameter drift were assessed. Five time points were simulated and IPD was assessed at the third, fourth, and fifth time periods. DeMars found that BILOG-MG and the modified Kim, Cohen, and Park (KPC) procedure effectively detected the drift items while maintaining Type I error rates very near the nominal alpha. BILOG-MG held the discrimination factor constant; only changes in the difficulty parameter were modeled. Whereas the modified KPC χ^2 test detected changes in both the discrimination and difficulty parameters, the process was significantly more complex. DeMars expressed the opinion that this extra complexity might be worthwhile if the KPC χ^2 had increased the power or decreased the error rate when compared to BILOG-MG. However, she did not find that to be the case.

Item Exposure

The literature suggests that item exposure is one of the components that can lead to IPD. To avoid the undesirable effects of item exposure, testing professionals and educators have traditionally been proactive in their attempts to minimize the possibility of its occurrence. A number of steps are typically taken to protect the integrity of items on a test: the secure shipment and storage of materials, the administration of exams in proctored environments, required confidentiality agreements from examinees, algorithms to control for item exposure, and the implementation of a regular, systematic process of writing, testing, and calibrating new items to replace older ones. All of these costly measures are intended to avoid item exposure or prior knowledge of items on the part of future examinees. As a result, several studies have focused on specific methods to detect exposed items and measure the effects of this exposure. Empirical studies to detect exposed items, however, have obtained mixed results and have not consistently demonstrated the ability to identify and measure evidence of item exposure.

One such study examined the Rasch estimates of ability and difficulty for two fixed form exams that were administered continuously for six months (Hertz & Chin, 2003). The study compared the means and standard deviations for each 2-month period in the 6-month testing cycles and found only slight differences in the difficulty estimates. The results suggested that candidates who took the exam later in the cycle did not benefit from information they may have received from candidates who tested earlier in the same cycle.

Another study investigated the impact of “braindump” sites on the performance of six items on an IT certification exam (Smith, 2004). A braindump was described as a website which allows candidates to exchange advice and information about an exam. While most of these sites instruct candidates not to post live test items, there are generally no controls in place to prevent such postings. In this study, six research items were added to six live IT exams (two research items per exam) and were posted, with correct responses, to a braindump site on the day the live test forms were released. Results for the exams were then monitored for a period of 18 months after the live exam was released. It was expected that all test items would become easier over time. Smith also hypothesized that the six exposed items would become noticeably easier, at a faster rate, than the other items on the exam, while controlling for candidate ability.

The test appeared to become easier as evidenced by a pass rate that went from 34.6% in the beta version of the exam to 65.5% in the operational test after only 10 months. The pass rate stabilized at this approximate level for the remaining 8 months of the study. Smith used the software program Winsteps, which utilized a modification of the Mantel-Haenszel DIF method, to detect drift. The research items, however, did not perform as expected. The study found that three of the research items became significantly more difficult based on the drift analysis while only one item became easier. Overall, the experimental items did not perform differently than the scored items on the test. A search of free braindump sites during the 18 month period the exam was in use showed that part of the item bank was compromised only three weeks after the exam was released, and the exam was almost entirely exposed, with a very high degree of accuracy, after 8 months. Smith asserted that the widespread exposure of items helped to explain the marked improvement in test scores and the lack of relative differences in the performance of the research items and the other test items.

An Internet search in April 2011 for information on the calculus placement exam used in this study yielded very little. Searches were conducted using the following keywords: placement tests, calculus placement tests, the actual name of the placement test, and the name of the university. Very few results were found, and the references that were found generally dealt with the perceived ease or difficulty of the exam. One posting on "Yelp.com" requested answers to a different version of calculus placement exam published by the same testing organization, but no responses to that request were found. No item level information about the exam was found. These limited results may be due, in part, to the three year time span from the time this exam was actually used and the date of the search.

Giordano and Subhiyah (2005) evaluated a take-home recertification exam in the medical field for item exposure. Sixty of 300 total items were presented on each of three consecutive administrations of the exam. The exams were administered every six months with candidates having three months to complete and submit their responses. Using Winsteps, IPD was assessed for the 60 repeated items. Results showed that only 12 of the 60 repeated items had significant IPD, and only six of these 12 items became easier over time as would be expected for exposed items. A comparison of the *p*-values for the repeated items and the new items found no difference between the performance of the items. Giordano and Subhiyah concluded that there was no evidence of exposure effects for the repeated items.

Another recent study focused on the responses of 130 candidates who retested on an exam that contained 36 repeated items (Wood, 2009). The exam administrations were four months apart. The study compared ability estimates for the candidates based on both the new and repeated items. While the ability estimates improved from the first test administration to the second, the increase was not significantly different for the repeat examinees on the reused versus new items.

Although these studies generally failed to demonstrate effects of prior knowledge on item function, they varied somewhat from the present study. In most of these studies, a relatively small percentage of the total test items were previously presented or exposed and, in most cases, the order or placement of specific items did not remain consistent from one test form or administration to another. In the present study, all items were used throughout the two year testing period. Even the specific order in which the items were presented remained consistent throughout the entire time. The fixed form of this exam allowed an analysis of item drift in which the specific item characteristic of location within the test was considered as a possible contributing factor to recall, and therefore potentially IPD.

Serial Position Effect and Item Recall

When the calculus placement exam was administered, specific steps were taken to actively protect the security of exam materials. As a result, the primary means by which items could become exposed or known to subsequent examinees was for the items to be recalled after the test was completed. Candidates who recalled items could research the problems for a solution prior to re-testing or could share specific item content with friends and acquaintances who needed to take the exam at a later date.

This study analyzed the item level data to determine if the first and last items in the exam exhibited IPD more markedly than other items which were placed in the middle of the exam. This hypothesis is consistent with the theory of serial position effect as espoused by Hermann Ebbinghaus and later revisited by Glanzer and Peters (1962). Generally, this theory suggests that, given a list of items, individuals will be more likely to recall the first items (primacy effect) and the last items (recency effect) on the list as opposed to items located in the middle. In a similar vein, this study considered whether examinees were more likely to recall the items that appeared toward the beginning and end of the exam.

Speededness

Another aspect of the exam that may have contributed to IPD in the last items was the speededness of the exam. The exam was timed at 90 minutes, and candidates were informed that their scores were based on the number of correct items. Candidates also were told it was to their benefit to answer all questions and were given a verbal warning when only five minutes of the testing time period remained. Some examinees did not complete the exam and it is reasonable to assume that others may have been guessing or answering items in a random manner at the end of the exam because they had insufficient time to complete the exam. Studies have demonstrated that the *a* and *b* parameters for items at the end of speeded exams tend to be overestimated (Oshima, 1994; Wollack, Cohen, & Wells, 2003). In this instance, if items on the exam were exposed, it may be reasonable to assume that a larger portion of the examinees completed the items at the end of the exam during later test administrations. As a result, these items might exhibit drift and appear to become easier compared to items in other portions of the exam.

In the case of a speeded exam, one might argue that fewer of the examinees had sufficient time to complete and cognitively process the items at the end of the test. As a result, they may have been less able to recall and share these items with future examinees. However, students who took this placement exam were aware that they had the right to re-test and by the end of the test may have had some general idea of their performance on the exam. If students felt less confident about the likelihood of achieving the required placement score on this exam, they may have taken some time at the end of the exam to review and consider these items in an attempt to remember them for a future attempt on the exam.

Focused Candidate Interaction with Item

This study also focused on item 15 on the exam. Examinees were instructed not to write in the test booklets as they were intended for re-use. At the completion of each test administration the proctor(s) had to inspect every page of each test booklet used in the exam administration. Stray marks had to be fully erased. If the proctor determined that it was not possible to fully erase all markings, the test booklet was retired. I served as a proctor for approximately one-half of all the administrations of this exam and in the process of inspecting the test booklets noted that item 15 was the one that was most commonly marked. Students who disregarded the testing rules and wrote in the test booklet most commonly wrote on item 15. The item included an associated diagram in which a shaded figure was bound on one side by a diagonal line, and examinees were asked to calculate the area of this figure. Many examinees made extraneous marks in an attempt to create a figure with 90 degree angles. Although there were 13 other items that included diagrams, item 15 seemed to elicit the need to draw on the part of the exam candidates. Given that a larger proportion of the examinees actively processed this item, the hypothesis is that students were able to better recall it. Again, exposure of test items is assumed to be positively correlated with the ability of earlier examinees to recall the item.

Chapter 3

METHODOLOGY

This study used the item level results from a 60-item calculus placement exam administered over the course of two years in a university environment. A single version of this paper-based exam, timed at 90 minutes, was administered to approximately 3,000 students over a two year time period. Based on reasoning that will be discussed in the coming sections, the decision was made to use results only for the students who were taking the calculus placement exam in order to register for a calculus course in the fall semesters.

Students were permitted to take the exam as many times as they desired, with the limitation that they could test only one time per day. In the present study, only the responses from a candidate's first attempt were utilized. As such, candidates who had prior knowledge of the test items would have received this information from external sources rather than prior testing experience with this particular assessment. Removing the candidates who were registering for the spring term and candidates who were re-testing reduced the number of test administrations, as well as the number of candidates to 2,787.

Grouping of Data

This study divided the data into six groups based on the chronological order in which the exam was administered. The groups were defined on three bases:

1. Administrations for any particular month remained together in one group. The specific month in which each exam was administered could be determined, but the specific date was not known. As a result, exams for any particular month were not split into separate groups.
2. Months were grouped together in an attempt to create adequate sample sizes for calibration and IPD analyses. The six groups, discussed next, varied in size from 238 to 691 examinees.
3. The test administrations were divided into three testing time windows for each of the two years in which the exam was administered, yielding a total of six groups. This division provided three pairs of time periods in which the characteristics of the examinee population were expected to be better matched and more consistent. A rationale for these groupings follows.

The groupings combined the months of 1) February, March, and April, 2) May and June, and 3) July and August. Each of these combinations of months provided one set of data for 2007 and another for 2008, resulting in six separate groups. The groupings were created in an attempt to better match student populations for comparison purposes. When frequencies of correct responses change over time, it may be due to IPD or differences in proficiency of examinee groups. To conclude that there is IPD, one must rule out the alternative explanation of changes in population characteristics. This study attempted to better match population characteristics in order to render changes in response tendencies attributable to IPD.

Candidate Characteristics by Group

The three groups in each of the two years were generally defined as Early, Regular, and Late examinees. Different general characteristics of these populations were observed by the researcher in her capacity as an administrator of the exam. For example, the Early examinees (February, March, April) tended to be eager, self- or parent-driven, and more confident in their skills. The Late examinees, on the other hand, tended to be less organized and exhibited signs of greater stress. A 2001 study analyzed academic performance and retention for students who registered in a community college on an Early, Regular, or Late basis (Street, Smith, & Olivarez, 2001). The results of the study showed that new students who registered late withdrew from a higher number of course credits and were less likely to persist by returning the following semester. Likewise, returning students who registered late had a lower GPA for the semester in which they registered late, successfully completed fewer credit hours, withdrew from more credit hours, and were less likely to persist to the next semester. These data would suggest that student populations vary in certain characteristics depending on the time frame in which they register for courses. Completing necessary placement tests, such as this calculus placement exam, is often an important, initial step in the registration process.

Selected Cases

The number of students testing for placement into spring semester calculus-based courses was significantly smaller than those testing for placement into fall courses. In addition, the testing period for the spring semester was somewhat more condensed than the testing period for the fall. Dividing the spring semester test administrations into three groups based on the student characteristics outlined in the preceding section may have resulted in groups that were too small for drift analysis. As a result, the decision was made to use only those cases for students who were testing for placement into fall semester courses.

The total numbers of subjects in the three paired comparisons in this study were 1,118, 1,117 and 552. Kirisci, Hsu, and Yu (2001) suggested that a test length of more 20 items and a sample size of more than 250 may be necessary to effectively estimate item parameters, but test lengths of 40 items and sample sizes of 1,000 may not be necessary. Likewise, Rupp (2003) provided a general guideline for tests with 15 to 50 items of approximately 250 subjects for the 1PL and 2PL models, while the 3PL requires 500 to as many as 1,000 examinees. The sample sizes proposed in this study were consistent with the suggested ranges in these studies.

Unidimensionality

One of the key assumptions of the IRT models studied here is unidimensionality. Exploratory factor analyses using M*plus* with weighted least squares with mean and variance adjustment (WLSMV) estimation were run on all 60 items for each of the six groups and the group of all examinees combined to determine if the data were consistent with the assumption of unidimensionality. Three basic criteria were considered in assessing the factor structure of the data for each group: 1) scree plots, 2) goodness-fit-indices, and 3) the actual factor loadings. Positive factor loading indicators included: item factor loadings of at least 0.30, limited cross-loadings for models with more than one factor, and at least 4-5 items loading on each factor for models with more than one factor.

Drift Analysis

The first drift analysis was conducted for all six groups using the Early 2007 group as the reference to see if any general trends could be identified. This analysis was followed by three separate drift analyses in which each group in 2007 was compared to its counterpart in 2008. BILOG-MG 3 was again used to conduct these analyses (see Appendix for sample BILOG code). The paired comparisons were intended to isolate the drift that may be attributable to item characteristics (i.e., exposure) as opposed to changes in the attributes of the population of examinees.

The statistics which were used to determine drift in this analysis were the polynomial (or linear) drift coefficients and standard errors. Similar to the technique used by DeMars (2004) in her drift study, an item was deemed to exhibit drift if the absolute value of the ratio of the polynomial coefficient to its standard error exceeded 1.96 ($\alpha = .05$). The drift coefficients were used to make the various comparisons across the six time periods and between the paired comparisons. The drift behavior of the items of interest was compared to that of the other items from the middle of the test.

Chapter 4

RESULTS

The analysis began by looking at the exam results from a classical test perspective to consider group differences relative to the observed scores. Table 1 reports sample sizes, mean test scores, standard deviations, and the percentages of examinees who successfully placed into calculus by achieving the minimum required placement score of 36. The overall one-way ANOVA comparing the mean test scores for the six groups was statistically significant ($F(5, 2781) = 7.372$, $p < .01$). The results from the post hoc pairwise comparisons, using the Bonferroni method to control family-wise Type I error, are reported in Table 2. As the results indicate, the mean score for Late examinees in 2007 was significantly lower than the mean score for all other groups. Of particular interest in this analysis, the mean score for the Late 2007 registrants was lower than both Early and Regular examinees in 2007 and lower than the mean score for Late examinees in 2008. No other pairwise mean differences were statistically significant.

Table 1

Mean Scores on Calculus Placement Exam by Group and Year

	2007				2008			
Exam Group	n	M	SD	% Placed	n	M	SD	% Placed
Early	593	37.13	11.35	55.5%	525	38.58	11.90	58.5%
Regular	426	36.97	11.45	54.9%	691	36.92	11.45	55.9%
Late	238	33.21	10.58	41.2%	314	36.41	11.42	51.6%
Total Year	1,257	36.33	11.34	52.6%	1,530	37.39	11.63	55.9%
Total - All	2,787	36.91	11.51	54.4%				

Table 2

Mean Score Differences between All Examinee Groups

	Regular 2007	Late 2007	Early 2008	Regular 2008	Late 2008
Early 2007	0.16	3.92*	-1.45	0.20	0.72
Regular 2007		3.76*	-1.61	0.05	0.56
Late 2007			-5.37*	-3.72*	-3.20*
Early 2008				1.65	2.17
Regular 2008					0.52

Note: Mean differences calculated as row minus column.

In an effort to determine if speededness may have affected results for items at the end of the test, the percentage of candidates who attempted each item was determined and the mean attempt rates were calculated for the first ten items, the last ten items, and the remaining items in the test, excluding item 15. Table 3 shows that the attempt rates were very high for all item locations within the test, for all groups of registrants, and for both years. The exam instructions included a statement indicating that scores were based on the total number of correct items.

Examinees were further informed that it was to their advantage to answer all questions. Whereas the high overall response rates would suggest that the vast majority of students attempted all questions, the response rates also show that examinees in all three registration groups responded to questions at the beginning of the test at slightly higher rates than the items at the end of the test. In the context of this study, no attempts were made to determine if an examinee was carefully responding to the items or guessing. Although various methods to address this issue exist, they are beyond the scope of the current study.

Table 3

Average Percentage of Candidates who Attempted Items by Examinee Category and Year

	Early		Regular		Late	
	2007	2008	2007	2008	2007	2008
1-10	99.1%	99.6%	99.2%	99.2%	99.1%	99.1%
15	98.3%	98.3%	97.4%	97.4%	97.9%	97.5%
51-60	95.2%	97.1%	96.8%	96.2%	93.2%	94.9%
All Other	98.3%	99.1%	98.2%	98.6%	97.7%	98.2%

An exploratory factor analysis was conducted using the software program MPlus 6.11 (Muthén &Muthén, 1998-2010) with WLSMV estimation on each of the six groups of first-time examinees and the total sample of examinees to determine if the data were consistent with the assumption of unidimensionality. The WLSMV estimation method uses tetrachoric correlations that are more appropriate for use with dichotomous data. Models with one to five factors, as well as a scree plot, were requested for each of the groups. The scree plots appear in Figures 1 through 7. Table 4 provides the following goodness of fit indices for each of the examinee groups: χ^2 , the root mean square error of approximation (RMSEA), the Tucker-Lewis Index (TLI), and the standardized root mean residual (SRMR). Hu and Bentler (1999) examined various cutoffs for many of the goodness of fit measures and their data suggested that Type I and Type II errors could be minimized by combining a relative fit index such as the TLI with the RMSEA or SRMR. Generally, these measures suggest good model fit at the following values: the χ^2 p-value is greater than 0.05, RMSEA is less than 0.06, TLI is greater than 0.95, and/or the SRMR is less than 0.08.

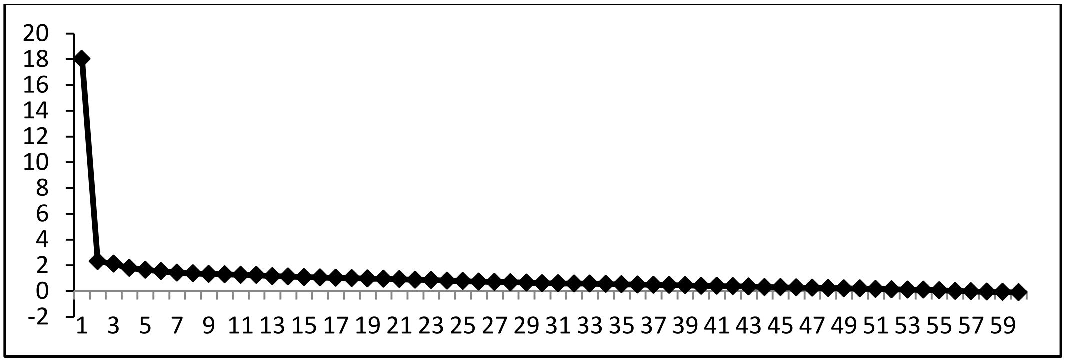


Figure 1. Scree plot of factor eigenvalues for Early 2007 examinees

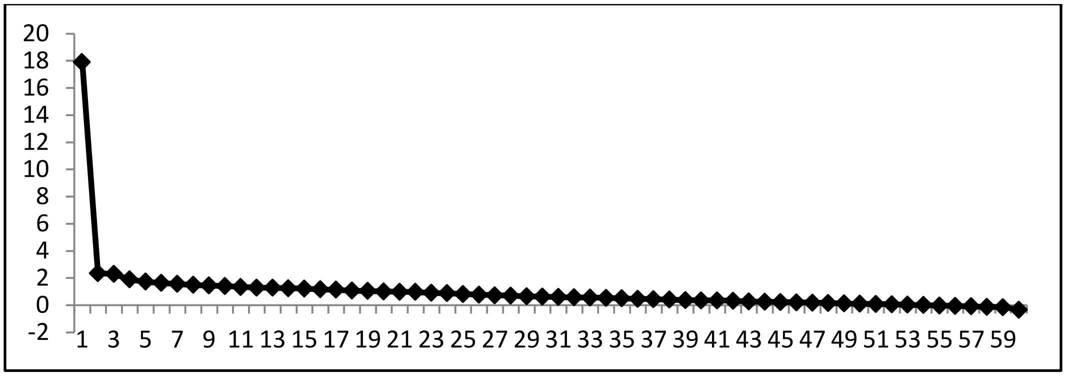


Figure 2. Scree plot of factor eigenvalues for Regular 2007 examinees

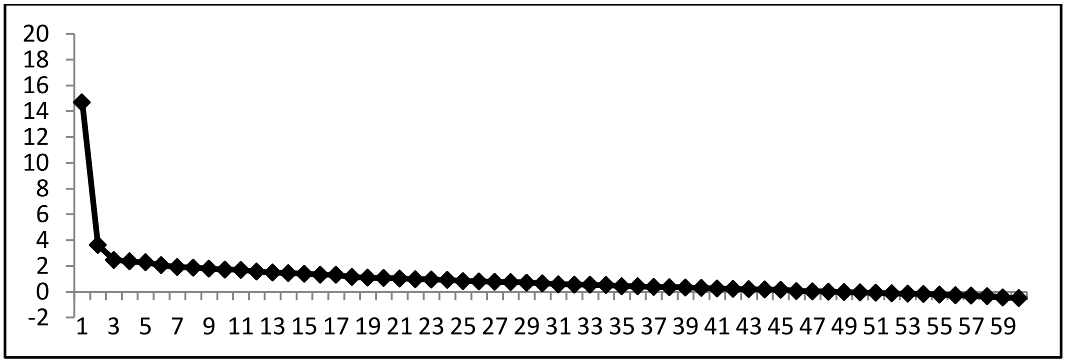


Figure 3. Scree plot of factor eigenvalues for Late 2007 examinees

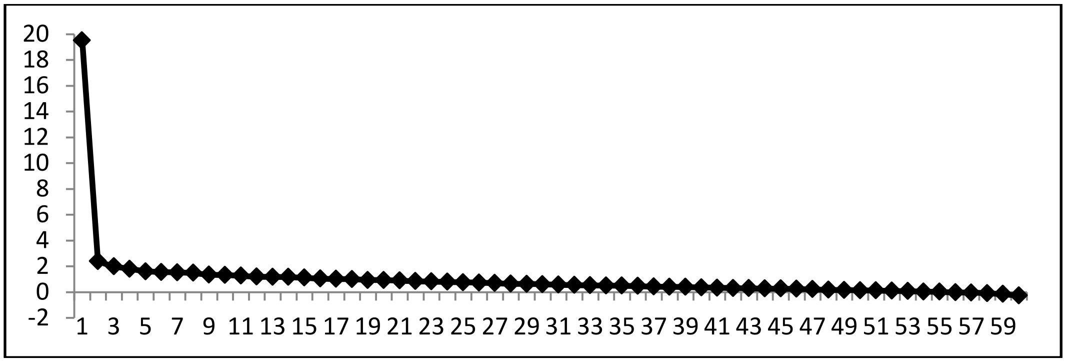


Figure 4. Scree plot of factor eigenvalues for Early 2008 examinees

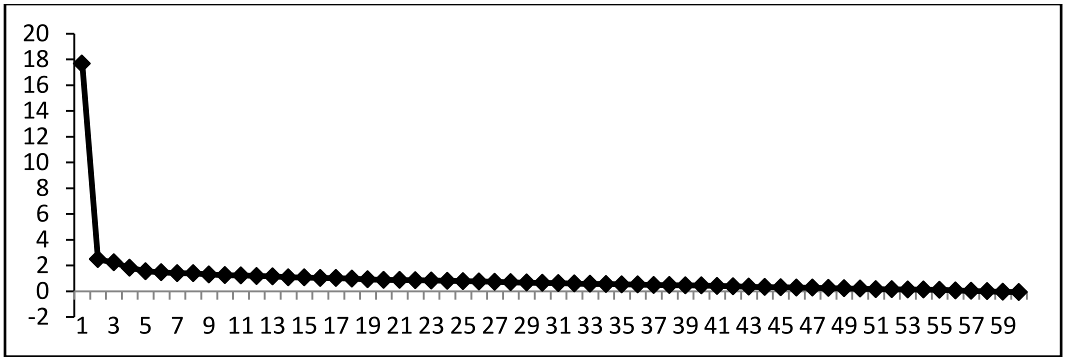


Figure 5. Scree plot of factor eigenvalues for Regular 2008 examinees

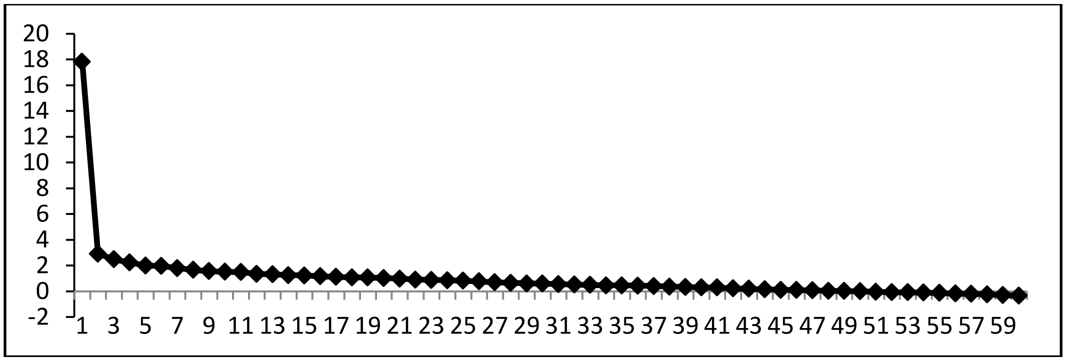


Figure 6. Scree plot of factor eigenvalues for Late 2008 examinees

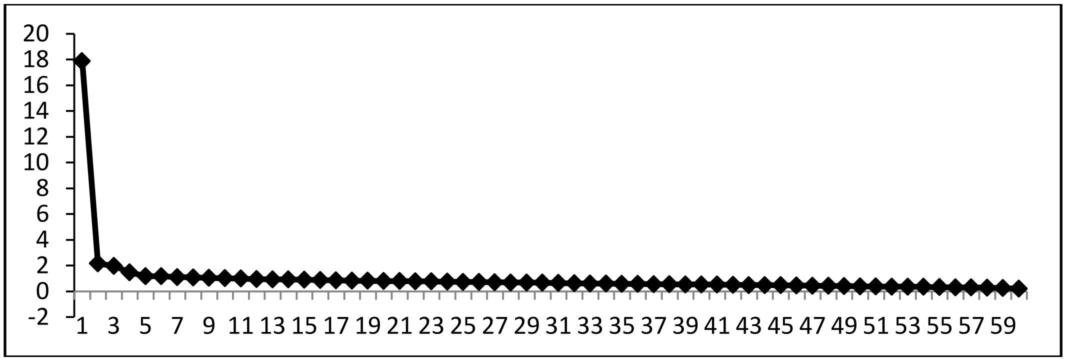


Figure 7. Scree plot of factor eigenvalues for all examinees

Table 4

Goodness of Fit Indices for One Factor Models

Group	χ^2	RMSEA	TLI	SRMR
Early 2007	1,945.908*	0.015	0.980	0.068
Regular 2007	1,836.934*	0.013	0.983	0.078
Late 2007	1,865.927*	0.020	0.939*	0.107*
Early 2008	1,929.218*	0.016	0.981	0.071
Regular 2008	2,130.925*	0.019	0.968	0.067
Late 2008	1,923.397*	0.020	0.961	0.092*
All	3,890.004*	0.021	0.967	0.045

Note: Degrees of freedom for χ^2 is 1,710.
* Statistics that fall outside of the generally accepted cutoff values.

The scree plots suggested that items were unidimensional for each of the groups of examinees, except the Late 2007 examinees. The items for the group of all examinees together, regardless of testing time frame, appeared to be unidimensional, as well. The goodness of fit indices also suggested that a unidimensional model fit the data relatively well for all of the groups except the Late 2007 examinees. Finally, the rotated factor loadings for the unidimensional models showed virtually all items loading well onto the single factor for all groups except the Late 2007 examinee group. Using a factor loading of 0.300 or higher as a cutoff, all groups other than the Late 2007 examinee group had only 2 or 3 items that did not load strongly onto the single factor (see Table 5). The item loadings for a single factor model for these groups were stronger than they were for any of the other models with two to five factors. There were two items (48 and 57) that failed to load on the factors for several of the groups.

Table 5

Factor Loadings for One Factor Models

Item	Early 2007	Regular 2007	Late 2007	Early 2008	Regular 2008	Late 2008	All Groups
1	0.260*	0.348	0.482	0.343	0.345	0.398	0.345
2	0.532	0.637	0.654	0.647	0.525	0.654	0.588
3	0.668	0.741	0.624	0.669	0.736	0.722	0.686
4	0.411	0.356	0.286	0.529	0.455	0.289	0.412
5	0.587	0.556	0.483	0.575	0.521	0.393	0.531
6	0.541	0.545	0.382	0.555	0.519	0.464	0.513
7	0.469	0.452	0.195*	0.446	0.459	0.391	0.423
8	0.424	0.464	0.425	0.435	0.353	0.417	0.411
9	0.627	0.610	0.455	0.604	0.571	0.543	0.578
10	0.604	0.636	0.617	0.572	0.544	0.684	0.594
11	0.437	0.497	0.305	0.478	0.480	0.362	0.446
12	0.416	0.525	0.332	0.494	0.405	0.402	0.435
13	0.475	0.468	0.430	0.520	0.398	0.524	0.461
14	0.320	0.371	0.401	0.419	0.444	0.290*	0.382
15	0.673	0.673	0.549	0.706	0.663	0.604	0.658
16	0.516	0.525	0.536	0.573	0.514	0.663	0.544
17	0.457	0.448	0.406	0.308	0.319	0.395	0.381
18	0.526	0.478	0.466	0.546	0.460	0.589	0.504
19	0.493	0.533	0.466	0.466	0.545	0.501	0.504
20	0.644	0.601	0.735	0.618	0.674	0.682	0.652
21	0.492	0.528	0.524	0.479	0.513	0.356	0.482
22	0.676	0.565	0.559	0.608	0.649	0.634	0.620
23	0.626	0.656	0.603	0.682	0.653	0.627	0.640
24	0.324	0.513	0.422	0.396	0.365	0.334	0.380
25	0.612	0.570	0.629	0.680	0.534	0.664	0.607
26	0.609	0.556	0.572	0.656	0.654	0.605	0.615
27	0.574	0.501	0.330	0.361	0.470	0.341	0.450
28	0.479	0.493	0.448	0.529	0.473	0.582	0.495
29	0.417	0.529	0.341	0.553	0.532	0.593	0.501
30	0.667	0.599	0.646	0.690	0.661	0.674	0.657
31	0.586	0.547	0.464	0.647	0.652	0.552	0.598
32	0.451	0.511	0.362	0.512	0.424	0.402	0.447
33	0.541	0.531	0.567	0.607	0.532	0.549	0.552
34	0.499	0.480	0.594	0.665	0.506	0.659	0.555
35	0.569	0.569	0.520	0.685	0.539	0.542	0.580
36	0.480	0.584	0.596	0.662	0.521	0.491	0.552

Item	Early 2007	Regular 2007	Late 2007	Early 2008	Regular 2008	Late 2008	All Groups
37	0.302	0.233*	0.168*	0.309	0.314	0.324	0.293*
38	0.583	0.532	0.433	0.633	0.541	0.677	0.577
39	0.505	0.551	0.436	0.489	0.536	0.542	0.519
40	0.616	0.597	0.492	0.608	0.609	0.473	0.581
41	0.579	0.624	0.483	0.522	0.505	0.559	0.547
42	0.636	0.522	0.550	0.641	0.643	0.564	0.605
43	0.683	0.757	0.577	0.701	0.722	0.685	0.697
44	0.683	0.627	0.549	0.673	0.611	0.692	0.645
45	0.661	0.611	0.517	0.615	0.622	0.536	0.604
46	0.622	0.622	0.680	0.659	0.675	0.664	0.650
47	0.541	0.638	0.552	0.601	0.591	0.526	0.579
48	0.237*	0.266*	0.145*	0.212*	0.208*	0.318	0.232*
49	0.730	0.727	0.618	0.739	0.682	0.645	0.700
50	0.701	0.646	0.631	0.670	0.637	0.875	0.679
51	0.567	0.557	0.573	0.618	0.517	0.637	0.574
52	0.554	0.516	0.408	0.496	0.480	0.529	0.504
53	0.536	0.472	0.438	0.580	0.548	0.471	0.526
54	0.692	0.598	0.532	0.711	0.676	0.612	0.655
55	0.347	0.448	0.374	0.382	0.393	0.337	0.382
56	0.541	0.434	0.383	0.549	0.532	0.475	0.506
57	0.355	0.407	0.168*	0.271*	0.271*	0.232*	0.299*
58	0.337	0.296*	0.255*	0.436	0.321	0.328	0.340
59	0.406	0.327	0.263*	0.382	0.421	0.419	0.383
60	0.644	0.623	0.663	0.676	0.710	0.757	0.676

* Item did not load onto the single factor at a level of 0.30 or higher.

The drift analysis began by looking at the trends of the six groups in chronological order using the first group as the reference to see if any discernible drift patterns could be identified. Although it is theoretically possible to create a drift model with a maximum power level of *G*-1, where *G* is the number of groups, BILOG-MG 3 did not converge for this data on any model with a power level higher than three. Successful IPD analyses were achieved for a linear and a quadratic drift model across all six groups, and the drift parameters for both models are provided in Table 6. Although the program provided results for a drift analysis at the third power, the resulting output was nonsensical with estimated *b* parameters in the hundreds and thousands, and all items drifted significantly. Consequently, these results were deemed invalid.

Table 6

Drift Parameters and Drift Statistics from the Linear and Quadratic IPD Analysis of All Six Groups of Examinees Combined

	Linear Analysis		Quadratic Analysis		
Item	Drift Parameter	Drift Statistic	First Power Drift Parameter	Second Power Drift Parameter	Drift Statistic
1	-0.032	-0.678	0.073	-0.016	0.316
2	-0.040	-1.877	0.051	-0.014	0.469
3	0.035	1.155	-0.201	0.035	-1.339
4	0.036	0.776	0.332	-0.044	1.423
5	0.051	1.964*	0.221	-0.026	1.669
6	0.043	1.636	0.101	-0.009	0.766
7	-0.011	-0.254	-0.048	0.005	-0.229
8	-0.023	-0.905	-0.055	0.005	-0.419
9	0.003	0.182	0.021	-0.003	0.229
10	0.056	2.709*	0.060	-0.001	0.580
11	0.016	0.594	-0.154	0.026	-1.144
12	-0.013	-0.455	0.321	-0.051	2.181*
13	0.071	1.802	-0.052	0.018	-0.261
14	0.052	1.715	0.122	-0.011	0.779
15	0.015	0.892	0.013	0.000	0.150
16	-0.005	-0.226	0.268	-0.041	2.224*
17	0.023	0.404	-0.211	0.035	-0.775
18	-0.036	-1.265	0.237	-0.041	1.625
19	0.017	0.617	0.033	0.000	0.239
20	-0.034	-1.970*	-0.112	0.012	-1.277
21	0.042	1.292	-0.061	0.015	-0.365
22	0.002	0.103	0.075	-0.011	0.766
23	-0.015	-0.736	-0.056	0.006	-0.547
24	-0.035	-1.141	0.004	-0.006	0.024
25	0.011	0.575	0.148	-0.021	1.519
26	-0.015	-0.813	0.050	-0.010	0.530

	Linear Analysis		Quadratic Analysis		
	Drift	Drift	First Power	Second Power	Drift
Item	Parameter	Statistic	Drift Parameter	Drift Parameter	Statistic
27	0.022	0.847	0.063	-0.006	0.494
28	-0.080	-3.131*	0.032	-0.017	0.243
29	0.010	0.366	0.017	-0.001	0.121
30	-0.025	-1.581	0.084	-0.016	1.060
31	0.003	0.196	0.063	-0.009	0.707
32	0.065	2.603*	-0.120	0.028	-0.945
33	0.046	2.374*	-0.119	0.025	-1.192
34	-0.022	-1.147	0.056	-0.012	0.567
35	-0.021	-1.108	0.092	-0.017	0.941
36	-0.014	-0.653	-0.105	0.014	-0.959
37	0.000	0.001	-0.034	0.005	-0.163
38	0.000	-0.026	0.105	-0.016	1.252
39	-0.033	-1.440	-0.017	-0.002	-0.149
40	0.051	2.481*	-0.072	0.019	-0.695
41	-0.006	-0.286	0.044	-0.008	0.384
42	0.025	1.246	0.070	-0.007	0.682
43	0.051	3.397*	-0.059	0.016	-0.765
44	-0.020	-1.149	-0.045	0.004	-0.503
45	-0.046	-2.135*	-0.283	0.035	-2.557*
46	-0.003	-0.161	-0.006	0.000	-0.064
47	-0.032	-1.636	-0.021	-0.002	-0.213
48	-0.015	-0.392	0.247	-0.040	1.293
49	0.016	1.112	0.076	-0.009	1.035
50	-0.027	-1.600	0.056	-0.013	0.679
51	-0.042	-2.120*	-0.112	0.011	-1.103
52	0.042	1.721	0.027	0.002	0.215
53	0.028	1.623	-0.103	0.020	-1.130
54	-0.021	-1.336	-0.066	0.007	-0.813
55	-0.048	-1.696	-0.151	0.015	-1.045
56	-0.017	-0.726	-0.087	0.011	-0.752
57	0.008	0.313	-0.091	0.015	-0.670
58	-0.021	-0.831	-0.262	0.037	-1.957
59	-0.011	-0.388	0.048	-0.009	0.348
60	-0.029	-1.696	-0.178	0.022	-2.044*

Note: Parameter drift statistic was calculated by dividing the first power drift parameter by its standard error.

* Statistically significant results, $p < 0.05$.

In the linear drift model with all six examinee groups, ten items were marked as drifting (items 5, 10, 20, 28, 32, 33, 40, 43, 45, 51). The drift parameters for four of these items (20, 28, 45, 51) were negative, meaning that the items were becoming easier. In the quadratic drift model for all six examinee groups, four items exhibited significant drift (items 12, 16, 45, 60). By the nature of a quadratic trend, the directionality of the drift changes and does not move consistently in one direction during the time period under consideration. For example, if the coefficient for the quadratic term is positive the trend will start downward and then become an upward trend (concave curve). Conversely, a negative coefficient will start as an upward trend and then reverse becoming a negative trend during the period of interest (convex curve).

Next, a drift analysis was run for each of the paired comparisons. For each group of examinees (Early, Regular, and Late), a drift analysis was run comparing 2007 to 2008, with each of the 2007 groups serving as the reference. The paired comparisons involved only two time points so the drift model was linear. Although these paired comparisons seemed less complex than the linear and quadratic drift analyses of all six groups, BILOG-MG 3 would not converge using a 3PL model. In an effort to get the drift models to successfully converge, various criteria were changed and implemented in various combinations: float was disallowed, the maximum EM cycles were increased from the default of 20, and the criterion for convergence was increased from the default of 0.01. Ultimately the models would not converge until the number of parameters in the model was decreased from 3PL to 2PL and float disallowed. The default maximum of 20 EM cycles and a criterion for convergence of 0.01 were used for the analyses. (See Appendix 1 for an example of the BILOG-MG syntax.)

As stated previously, the drift statistic for each item was calculated as the ratio of the drift parameter to its standard error. The results were deemed significant at the .05 level if the absolute value of the ratio exceeded 1.96. The analysis identified several drift parameters that were significant in the paired comparisons (see Table 7).

Table 7

Drift Parameters and Drift Statistics from the Paired Comparisons of Examinee Groups

	Drift Parameters			Drift Statistics		
Item	Early	Regular	Late	Early	Regular	Late
1	0.142	-0.261	-0.070	0.492	-1.084	-0.254
2	0.047	-0.063	-0.156	0.374	-0.520	-1.060
3	-0.090	0.007	0.280	-0.649	0.064	1.529
4	0.304	0.186	-0.257	1.533	0.821	-0.616
5	0.302	0.191	-0.162	2.239*	1.374	-0.566
6	0.278	-0.157	0.243	1.976*	-1.181	0.957
7	-0.025	-0.123	0.055	-0.127	-0.698	0.142
8	-0.121	0.059	-0.062	-0.707	0.326	-0.234
9	0.120	-0.028	0.088	1.063	-0.248	0.434
10	0.281	0.205	0.399	2.286*	1.752	2.704*
11	-0.089	-0.022	0.215	-0.535	-0.149	0.648
12	0.216	-0.026	-0.566	1.252	-0.159	-1.833
13	0.266	-0.098	0.599	1.481	-0.477	2.263*
14	0.375	0.163	-0.146	1.848	0.909	-0.442
15	0.112	0.046	0.135	1.108	0.472	0.772
16	0.161	0.107	-0.244	1.190	0.797	-1.470
17	-0.063	-0.012	0.044	-0.231	-0.042	0.132
18	0.049	-0.176	-0.266	0.327	-1.133	-1.325

	Drift Parameters			Drift Statistics		
Item	Early	Regular	Late	Early	Regular	Late
19	0.087	0.043	0.074	0.516	0.309	0.324
20	-0.274	-0.071	-0.214	-2.516*	-0.715	-1.577
21	0.136	-0.054	0.275	0.753	-0.372	1.002
22	0.121	-0.133	0.208	1.110	-1.205	1.296
23	0.042	0.024	0.146	0.382	0.236	0.889
24	-0.187	0.072	-0.178	-0.896	0.420	-0.591
25	0.036	0.111	-0.238	0.327	0.880	-1.537
26	0.032	0.050	-0.029	0.276	0.450	-0.165
27	-0.130	0.216	-0.266	-0.814	1.520	-0.852
28	-0.218	-0.184	-0.331	-1.446	-1.146	-1.507
29	-0.189	0.164	-0.286	-1.103	1.186	-1.339
30	0.066	-0.078	0.005	0.609	-0.729	0.029
31	0.153	0.118	-0.197	1.335	1.127	-0.944
32	0.276	0.283	0.659	1.784	1.818	2.359*
33	0.091	0.331	0.338	0.721	2.518*	1.688
34	0.011	-0.122	0.125	0.085	-0.839	0.747
35	-0.048	0.073	-0.401	-0.380	0.507	-1.713
36	-0.177	-0.064	-0.050	-1.347	-0.472	-0.228
37	-0.155	-0.345	-0.115	-0.642	-1.323	-0.299
38	0.039	-0.003	-0.138	0.349	-0.022	-0.747
39	-0.267	-0.181	-0.376	-1.707	-1.481	-1.784
40	0.064	0.330	0.217	0.562	3.034*	0.992
41	-0.147	0.070	-0.282	-1.090	0.571	-1.450
42	0.116	-0.050	0.039	1.032	-0.445	0.210
43	0.129	0.060	0.459	1.320	0.694	2.709*
44	-0.132	-0.034	-0.059	-1.277	-0.306	-0.380
45	-0.299	-0.181	0.144	-2.611*	-1.664	0.722
46	0.001	-0.090	0.015	0.013	-0.910	0.100
47	-0.119	-0.130	-0.258	-0.945	-1.196	-1.288
48	0.099	-0.507	-0.119	0.343	-1.732	-0.289
49	0.083	0.169	-0.088	0.835	1.654	-0.474
50	-0.105	-0.059	0.189	-0.859	-0.444	1.198
51	-0.345	0.032	-0.414	-2.735*	0.237	-2.273*
52	0.063	0.070	0.018	0.448	0.505	0.078
53	-0.044	0.092	0.162	-0.352	0.740	0.710
54	-0.089	-0.026	-0.067	-0.858	-0.255	-0.354
55	-0.255	-0.056	-0.188	-1.247	-0.331	-0.637
56	-0.024	-0.040	0.025	-0.167	-0.291	0.095

	Drift Parameters			Drift Statistics		
Item	Early	Regular	Late	Early	Regular	Late
57	-0.520	-0.456	-0.033	-2.223*	-2.032*	-0.072
58	-0.314	0.119	0.161	-1.696	0.576	0.502
59	-0.055	-0.122	0.121	-0.286	-0.659	0.387
60	-0.239	-0.182	0.218	-1.977*	-1.694	1.396

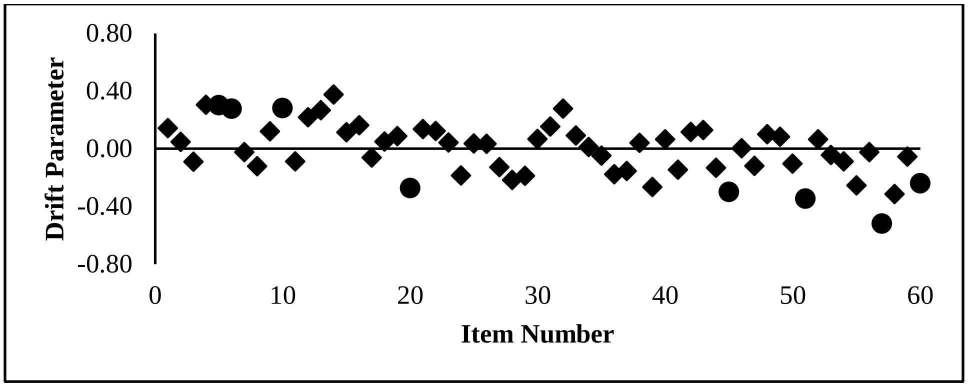
Note: Parameter statistic was calculated by dividing the drift parameter by its standard error.
* Statistically significant results, *p*<.05

Among the 60 items, thirteen items were deemed to have drifted significantly for at least one category of examinees. Of these thirteen drifting items, the drift parameter was positive for eight items, indicating they became more difficult for the subsequent group of examinees. The five remaining items that exhibited significant drift became easier.

The items of particular interest in these analyses were the first ten items, item 15, and the last ten items. Among the first ten items, three items (5, 6, and 10) drifted significantly. Items 5 and 6 drifted between 2007 and 2008 for Early examinees only. Item 10 drifted between 2007 and 2008 for both Early and Late examinees. In all of these cases, the drift parameter was positive, indicating that the items became more difficult for the subsequent group. For Regular examinees, none of the first ten items exhibited significant drift. In addition, Item 15 did not drift significantly in either direction for any of the three groups of examinees.

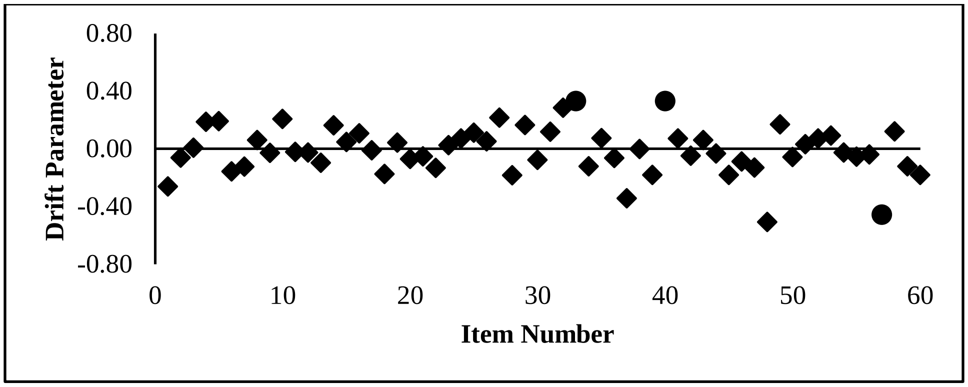
The last ten items also had three items that drifted significantly (items 51. 57, and 60). All three of these items had negative drift parameters indicating that the items became easier. Item 51 drifted for the Early and Late examinees, item 57 drifted for Early and Regular examinees, and item 60 drifted for Early examinees only. In the other portions of the exam (items 11 – 50, excluding item 15), seven items drifted significantly for at least one category of examinees. Only two of the seven items that drifted significantly in the middle portion of the exam became easier.

Figures 8 – 10 below illustrate the drift parameters by item number with the items in numerical order on the *y*-axis. Consistent with the constraint on the coefficients on the time variables (δ in Equation (3)), the figures exhibit a relatively even distribution of positive and negative drift parameters. Figure 8 suggests a slightly downward trend of drift parameters by item number for Early examinees. In other words, the trend would suggest that later items generally became easier, compared to earlier items. The regression equation (y = -0.006x + 0.164, R^2 = 0.279) suggests a slightly downward trend. Further, the R^2 of .28 indicates that item location accounted for approximately 28% of the variability in the item drift parameter for Early examinees. No trend in the drift parameter relative to item location was evidenced for the Regular and Late examinees (see Figures 9 and 10).



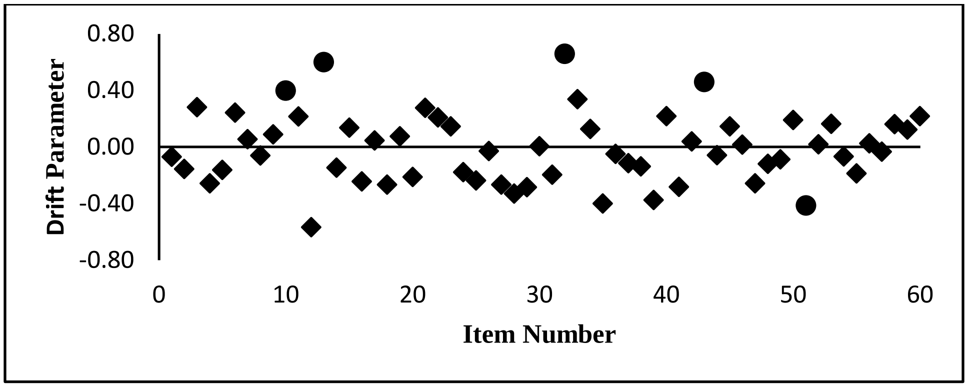
Note: The items which drifted significantly are marked with a circle.

Figure 8. Drift parameter as a function of item number for Early examinees.



Note: The items which drifted significantly are marked with a circle.

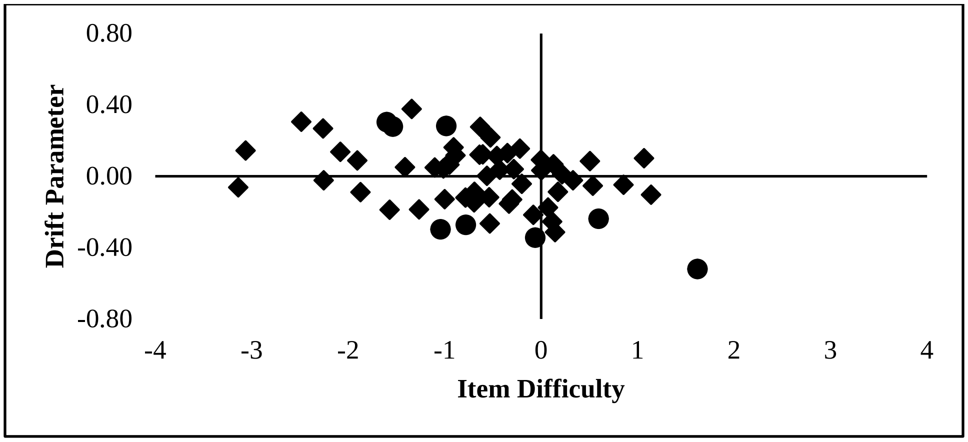
Figure 9. Drift parameter as a function of item number for Regular examinees.



Note: The items which drifted significantly are marked with a circle.

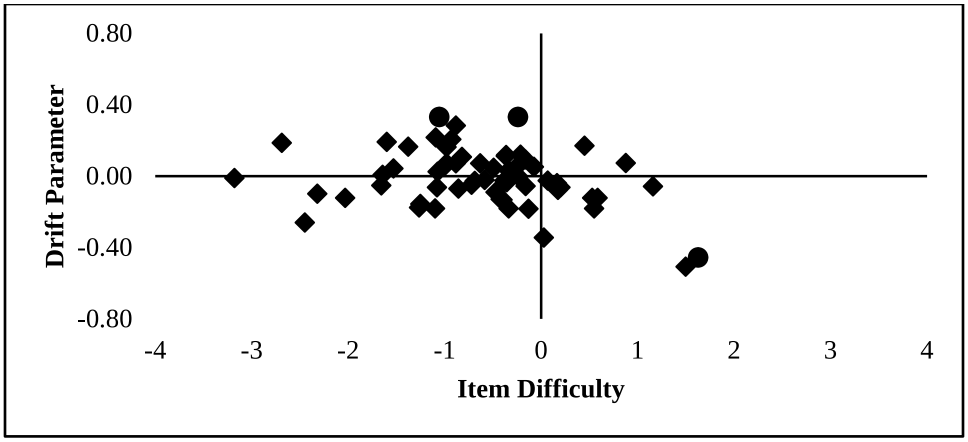
Figure 10. Drift parameter as a function of item number for Late examinees.

Figures 11-13 illustrate the relationship between item difficulty (*b*), as estimated for the reference group, and item drift. For all three examinee groups, easier items tended to drift in a positive direction, indicating they became more difficult while the harder items tended to drift in a negative direction indicating that they became easier. For Early examinees this relationship between item difficulty (*b*) and item drift was expressed by the regression equation of $y = -0.0761x - 0.0561$, $R^2 = 0.1584$. The regression equation for Regular examinees was $y = -0.0532x – 0.0441$, $R^2 = 0.0907$. Similarly, the relationship between item difficulty (*b*) and item drift for Late examinees was expressed as $y = -0.0433x – 0.0245$, $R^2 = 0.0354$.



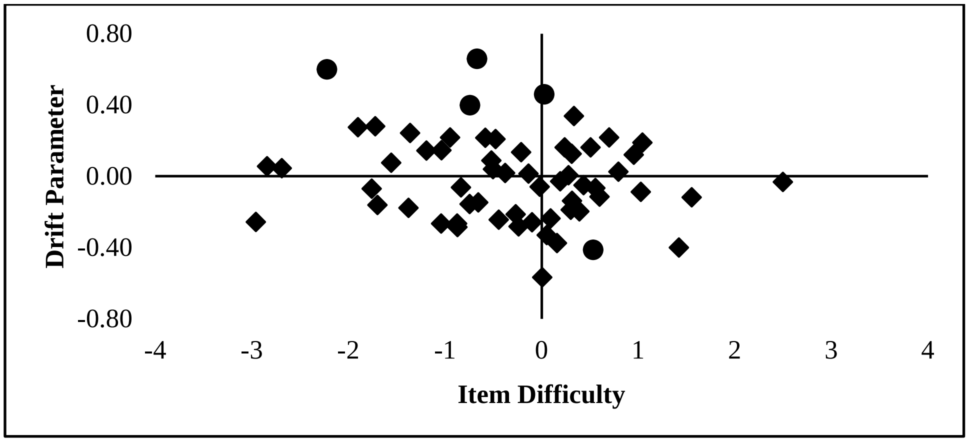
Note: The items which drifted significantly are marked with a circle.

Figure 11. Drift parameter as a function of item difficulty for Early examinees.



Note: The items which drifted significantly are marked with a circle.

Figure 12. Drift parameter as a function of item difficulty for Regular examinees.



Note: The items which drifted significantly are marked with a circle.

Figure 13. Drift parameter as a function of item difficulty for Late examinees.

Figure 14 shows the relationship between the estimated item drift and item difficulty (*b*) for each item that drifted significantly in any of the three paired comparisons. Three of the items (10, 51, and 57) drifted in two of the three paired comparisons. Each of these items is graphed twice to indicate the estimated drift on the basis of the estimated difficulty in each group in which the item was marked for significant drift. As a result, there are 16 points on this graph. The graph illustrates a negative correlation between item drift and item difficulty such that easier items were more likely to become more difficult and difficult items tended to become easier. The regression equation (y = -0.2578x – 0.0367, R^2 = 0.4688) indicated the moderately strong relationship in which item difficulty accounted for almost 47% of the variance in item drift.

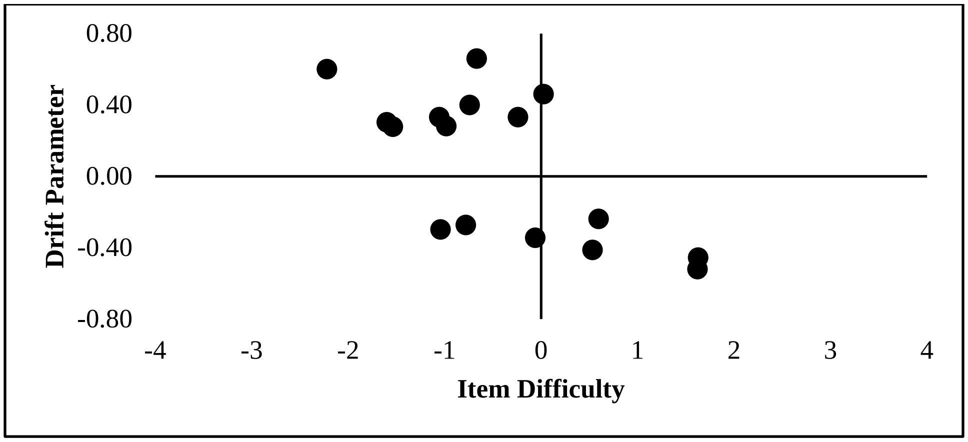


Figure 14. Drift parameters of significantly drifting items as a function of item difficulty.

The BILOG-MG results also provided mean ability estimates (θ values) for examinees in each of the groups. In each of the three paired comparisons, the group in 2007 was set as the reference group. As such, the mean and standard deviation of the θ values for these examinees were set by BILOG-MG to 0.00 and 1.00, respectively. Table 8 provides the mean and standard deviation as estimated by the BILOG-MG drift analyses. The mean θ values were significantly higher for both Early and Late examinees in 2008 compared to 2007.

Table 8

Mean Theta Estimates for Examinees by Group and Year

Registration	2007		2008	
Group	M	SD	M	SD
Early[1]	0.00	1.00	0.18	1.16
Regular	0.00	1.00	0.00	1.03
Late[2]	0.00	1.00	0.33	1.18

[1] Significant mean differences for 2007 and 2008, *t* (1,116) = 2.77, *p*< .01

[2] Significant mean differences for 2007 and 2008, *t* (550) = 3.54, *p*< .01

Chapter 5

DISCUSSION

This study began by looking at the data from a classical test theory perspective by comparing the mean scores for the six groups of examinees. The ANOVA analysis with post hoc pairwise comparisons found that the group of Late 2007 examinees was the only one with a statistically significant mean difference. This group was also the one with potential concerns about the assumption of unidimensionality on the basis of the exploratory factor analysis. This group of examinees was the smallest with 238 examinees which was a relatively small number of subjects compared to the number of items (60). A larger sample in this group may have provided a more stable factor analysis. Although the data for this one group yielded some evidence to question the IRT assumption of unidimensionality, the decision was made to proceed with the drift analysis because a unidimensional model seemed to adequately fit the data for all other groups including the combined sample of all examinees.

The results from the linear and quadratic drift analyses run on all six examinee groups combined were compared to those obtained from the three paired comparisons. The linear models were fairly consistent, identifying nine of the same items (5, 10, 20, 32, 33, 40, 43, 45, 51) as drifting significantly. Both models agreed, as well, on the directionality of each of these drifting items. The quadratic drift model for the six groups identified four drifting items, but only two of these items (45, 60) were consistent with the findings of the paired comparisons. The interpretability of the drift results from quadratic or higher polynomial models, however, seems problematic. As DeMars (2004) stated,

> A complication of using the highest order polynomial is that an item would be flagged if the difficulty unpredictably fluctuated in both directions, such fluctuations might be interesting but they could not be meaningfully interpreted as drift in the way of directional changes detected by the linear procedures (pp. 298-299).

While the linear drift analysis of the six examinee groups combined marked many of the same items for drift as the pairwise comparisons, this analysis was not consistent with the hypothesized differences in examinees groups. Consequently, the drift results from the pairwise comparisons were used to test the hypotheses in this study.

This study tested three IPD hypotheses in an IRT framework. The first of these hypotheses stated that evidence of IPD was expected to be exhibited by items on the calculus placement exam due to item exposure. Consistent with this hypothesis, items were expected to become easier over time. Of the 60 test items, ten items (5, 6, 13, 20, 32, 33, 40, 43, 45, and 60) exhibited drift in one of the three paired comparisons of Early, Regular, and Late examinees. Three additional items (10, 51, and 57) exhibited significant IPD in two of the three paired comparisons. Of the thirteen items that exhibited significant IPD, only five of the items became easier in the later time period. The results of this study failed to support this first hypothesis.

The second hypothesis stated that IPD would be more pronounced in the first and last items on the exam, in comparison to those in the middle of the exam (items 11-50, excluding 15). Three items in the first ten (5, 6, 10) exhibited significant IPD. However, contrary to the hypothesis, all three became more difficult rather than easier. Among the last ten items on the test, three items (51, 57, 60) exhibited significant drift. As hypothesized, each of these three items became easier. These results may be consistent with the hypothesis of item exposure for these last items on the exam.

Relative to the issue of speededness for this placement exam, Table 3 showed that a higher percentage of Early and Late examinees in 2008 attempted the last ten items on the exam compared to their counterparts in 2007. For these three items, in particular, the average response rates between 2007 and 2008 increased from 94.0% to 96.6% for Early examinees and 91.9% to 94.3% for Late examinees. Since items with no response were marked as incorrect, higher attempt rates are likely associated with higher rates of correct responses, if for no other reason than guessing. Therefore, the three items among the last ten on the exam that became significantly easier may have drifted, in part, due to the higher response rates for Early and Late exam candidates in 2008.

Three items among the first ten (30%) exhibited significant drift in at least one of the paired comparisons and three of the last ten items (30%) showed evidence of drift. Among the 39 other items in the middle of the exam, seven exhibited significant drift in one of the paired time comparisons (18%). Clearly, drift was exhibited at a higher rate by the items at the beginning and end of the exam, but only the items at the end of the exam became easier, consistent with a hypothesis of item exposure.

The third hypothesis suggested that item 15 would exhibit item parameter drift by becoming easier over time due to exposure. It was hypothesized that this item would be more memorable because significant numbers of candidates interacted with this particular item by marking or drawing on the diagram associated with this item. However, the drift analyses showed no evidence of item parameter drift for item 15.

When results for an item differ over time, the difference may be explained by changes in the performance of the item itself, by changes in candidate ability, or by a combination of the two. In the drift analysis, BILOG-MG estimates changes in each item over time in the form of the drift parameter and also provides an estimated mean θ value for the examinees at each time period.

As discussed previously, the drift parameter values for all items on an exam are constrained to sum to zero (Equations (3) and (5)). As a result, all the items on an exam cannot move in a single direction over time; items will not all become more difficult or easier. Further, the values of the discrimination (a) and lower asymptote (c) variables are held constant for each item at the levels established in the reference group. As a result, changes in the overall performance of examinees on the exam over time can be explained only by changes in the b parameter for the items or by the θ estimates for the examinees. To the extent that the overall performance of a group of examinees at a certain point in time is better or worse than the performance of the reference group, θ estimates may vary to account for these wholesale differences in performance.

In the paired comparisons of Early and Late examinees, the drift analyses estimated higher mean ability for candidates in 2008 compared to 2007. The statistically significant higher ability estimates in 2008 for Early and Late examinees may be attributable to a number of reasons in addition to the hypothesized item exposure. In this instance, higher ability estimates in 2008 may also be due to advance notice of the placement testing requirement, the availability of a practice exam/study guide, and/or changes in pre-calculus curriculum.

The calculus placement exam requirement was put into place on very short notice in 2007 and many candidates arrived at freshman orientation events to discover they had to take this placement exam prior to registering for a calculus course. In contrast, by 2008, the requirement had been in place for at least a year. Mailings and emails to incoming students had been updated to include this information, and a study guide and short practice exam were available online. This online information on the university website was available not only to incoming students, but also local high school teachers and community college instructors who may have altered the content of pre-calculus courses to help students successfully place into calculus courses at the university. As a result, students in 2008 may have had the advantage of being in the correct mindset for the exam and/or having had the opportunity to review prior to sitting for the exam. In other words, they may have had a higher degree of test-preparedness.

Item drift and item difficulty exhibited a moderately strong, negative correlation for the items that drifted significantly in the pairwise comparisons. Items with low difficulty tended to drift in a positive direction, becoming more difficult, and items with a higher item difficulty tended to drift in a negative direction, becoming easier. This relationship suggested that the significantly drifting items tended to drift toward more moderate difficulty levels rather than drifting to the extremes. This correlation may be an indication of the limitations of conducting a drift analysis utilizing two time points as the differences in sampling variability between the two groups may be more significant. These comparisons may result in some tendency for regression toward the mean.

Implications

These results did not support the hypotheses as outlined previously in this study. Rather, they tend to suggest that the items were performing relatively well over the two year period of time. While 13 of the 60 items (21.7% of all items) showed significant drift in at least one of the paired comparisons, eight of these items became more difficult while the others became easier. The evidence did not suggest that significant changes in item performance were adversely affecting the results from this single form of a placement exam even after two years of almost continual use. In and of themselves, these results would not constitute a compelling reason for the university to change or discontinue the use of this placement exam, nor would it suggest to the testing company that the integrity of the items on this exam form had been compromised.

Limitations

Several limitations in this study were identified. For example, the study found that three items among the last ten on the exam exhibited significant negative drift, becoming easier for a subsequent group. The limitations of this drift analysis, however, make it impossible to state conclusively that the items became easier because of item exposure. The current analysis did not provide a method to discern whether this drift occurred as a result of item exposure, curriculum changes, better preparation on the part of subsequent examinees, or guessing. The drift analysis marked certain items as potentially troublesome, but additional analysis would be appropriate in determining if these items should be retired and replaced.

Additional time periods would have been helpful in this analysis. The individual group comparisons were limited to two time periods. Although these comparisons marked certain items as drifting significantly, two time periods were not sufficient to establish trends. More time periods may have been beneficial in identifying IPD.

Also, the examinee group divisions used in this study may have been somewhat artificial for a number of reasons. Examinees were divided on the basis of the month in which each tested for the first time. Examinees could not be divided into groups based on finer-grained time intervals because information about the specific date on which each examinee tested was no longer readily available.

Examinees were grouped in an attempt to control for perceived differences in candidates based on the particular time in which the student was testing and attempting to register for a calculus course. Such a broad generalization of behavioral characteristics was inevitably imprecise, particularly in the absence of other behavioral measures.

It may be noteworthy to compare the level of significant item drift exhibited by each of the paired examinee groups. The Early examinee groups had significant drift in eight items, with five of these items becoming easier. This compares to three items with significant drift for the Regular examinees, with only one of these items becoming easier. Similarly, five items exhibited significant drift for the Late examinees, with only one item becoming easier. The Early examinees in 2007 were the first group at the University to take the placement exam. They had very limited opportunities to gain prior knowledge of item content from their peers. However, the Regular and Late examinees, even in 2007, may have had the opportunity to obtain information about the exam content from friends and acquaintances. Neither of these two groups may have been as “pure” a reference group as the Early 2007 examinees were.

Finally, this study also encountered limitations of the BILOG-MG program. Theoretically, parameter drift can be calculated to a power equal to the number of groups minus one. When all six groups were combined for the purpose of a drift analysis, the program could run the drift analysis successfully only in the first and second power. The program calculated meaningless values for the parameters in the third power and failed to converge in the fourth and fifth powers. Similarly, DeMars (2004) commented that she had trouble getting higher order trends to converge. In addition to convergence issues for higher order trends, BILOG-MG did not converge for the paired comparisons unless the model was reduced to 2PL and float was disallowed. Whereas the program is a powerful tool, these highly complex calculations cannot always be completed as anticipated or desired.

Additional Studies

An analysis of the predictive validity of the exam would be an appropriate follow-up to this drift analysis. The scores for examinees were generally higher in year two than in year one, and this drift analysis marked certain items as drifting significantly. A study to determine student success in courses after placement might help to determine if the placement test was performing as desired. If the overall performance of the placement test was declining, certain corrective actions could be taken such as replacing test items or adjusting the cut score. Although the university discontinued the use of this assessment for calculus placement at the end of the two year period in this analysis, the follow-up study of predictive validity could prove informative.

Item level responses from first-time examinees only were included in this analysis. Yet students who failed to achieve the minimum calculus placement score were permitted repeated attempts on the same form of the exam. The only restriction was that students could test only one time per day. Among all students who took the calculus placement exam at the university, 242 (8% of the total) took the exam more than one time. A separate analysis of IPD on this test could be conducted by comparing examinee responses from the first test administration to responses on a subsequent attempt. This analysis could again consider specific item characteristics such as item location.

The present study looked for evidence of item exposure across all levels of theta. However, a follow-up study might compare the relative effects of item drift conditional on proficiency. Items may have drifted in a different manner for examinees at the extremes of the proficiency continuum than they did for the other examinees.

Finally, the present study used the estimation method in BILOG MG to identify item parameter drift. This method limits the expression of drift to the difficulty (*b*) parameter. In addition, this drift analysis constrains all of the drift coefficients (δ_j) to sum to zero. This constraint prevents all items from becoming easier or all becoming more difficult. However, this method of drift identification may be somewhat inconsistent with hypotheses where drift is expected to be uni-directional, as in the case of item exposure. Future studies might use other methods to detect IPD such as modifications of the Mantel-Haenszel method (Holland & Thayer, 1986) or some of the methods devised to measure the area between two item response curves (Kim & Cohen, 1991; Raju, 1988; Raju, 1990).

Conclusions

For the most part, the empirical evidence in this study did not support the hypotheses. There was some evidence that a higher proportion of the last items drifted and became easier over time, consistent with the behavior of exposed items. In addition, the theta estimates suggested that the Early and Late examinees in 2008 had higher average ability. However, the results did not separate the possible effects of known items, better test preparation, curriculum changes, or speededness.

The drift analysis method utilized by BILOG-MG may not be the best method to assess for known items or to test other hypotheses that likely would result in uni-directional item drift. The constraint that requires the first power drift coefficients (δ_j) of all items on a test to sum to zero essentially precludes drift in only one direction.

REFERENCES

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. Retrieved from http://www.jstor.org.ezproxy1.lib.asu.edu /stable/1434961?seq=1

DeMars, C. E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education, 17,* 265-300. Retrieved from http://web.ebscohost.com.ezproxy1.lib.asu.edu/ehost /pdfviewer/pdfviewer?sid=593584bd-7117-4999-a79d-6a7d4b44d25e %40sessionmgr14&vid=2&hid=12

Donoghue, H. R., & Isham, S. P. (1998). A comparison of procedures to detect Item parameter drift. *Applied Psychological Measurement, 22*, 33-51. Retrieved from http://apm.sagepub.com.ezproxy1.lib.asu.edu/content /22/1/33.full.pdf+html

Giordano, C., Subhiyah, R., & Hess, B. (2005). An analysis of item exposure and item parameter drift on a take-home recertification exam. *Paper presented at the Annual Meeting of the American Educational Research Association*. Montreal, Canada. Retrieved from http://www.eric.ed.gov/PDFS /ED497708.pdf

Glanzer, M. & Peters, S. C. (1962). Re-examination of the serial position effect. *Journal of Experimental Psychology, 64,* 3, 258-266. Retrieved from http://search.proquest.com.ezproxy1.lib.asu.edu/docview/614250654/fulltextPDF/134B4E5F0DE53042FCF/9?accountid=4485

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, 20. Retrieved from http://www.jstor.org.ezproxy1.lib.asu.edu/stable /1434953?seq=1

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory, Newbury Park, CA: SAGE Publications, Inc.

Hertz, N. R. & Chinn, R. N. (2003). Effects of question exposure for conventional examinations in a continuous testing environment. *Paper presented at the 2003 Annual Meeting of the National Council on Measurement in Education*. Chicago, IL. Retrieved from http://www.eric.ed.gov/PDFS /ED476422.pdf

Holland, P.W. & Thayer, D. T. (1986). Differential item performance and the Mantel-Haenszel procedure. *Paper presented at the American Educational Research Association Annual Meeting.* San Francisco, California. Retrieved from http://www.eric.ed.gov/PDFS/ED272577.pdf

Hu, L., & Bentler, P.M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling,* 6(1), 1-55.

Kim, S.-H. & Cohen, A.S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement*, 15, 269-278. Retrieved from http://apm.sagepub.com.ezproxy1.lib.asu.edu/content/15/3/269

Kim, S.-H., Cohen, A.S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32,* 261-276. Retrieved from http://www.jstor.org.ezproxy1.lib.asu.edu/stable/10.2307/1435297?origin=api

Kirisci, L., Hsu, T,-C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*, 146-162. DOI: 10.1177/01466210122031975

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems, Hillsdale, NJ: Lawrence Erlbaum Associates.

Muthén, B.O. and Muthén, L.K. (1998-2010). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén. Retrieved from http://statmodel.com/download/usersguide/Mplus%20Users%20Guide%20v6.pdf

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*, 3,200-219. Retrieved from http://www.jstor.org.ezproxy1.lib.asu.edu/sici?sici=0022-0655%281994%2931%3A3%3C200%3ATEOSOP%3E2.0.CO%3B2-Q&origin=serialsolutions

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometricka*, 53, 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned area between two item response functions. *Applied Psychological Measurement*, 14, 197-207. Retrieved from http://apm.sagepub.com.ezproxy1.lib.asu.edu/content/14/2/197

Rupp, A. (2003). Item response modeling with BILOG-MG and MULTILOG for Windows. *International Journal of Testing, 3,* 365-384. Retrieved from http://web.ebscohost.com.ezproxy1.lib.asu.edu/ehost/pdfviewer/pdfviewer?sid=c4b27920-4146-440d-a2e8-140d6829622e%40sessionmgr14&vid=2&hid=7

Smith, R. W. (2004). The impact of braindump sites on item exposure and item parameter drift. *Paper presented at the Annual Meeting of the American Education Research Association*. San Diego, CA. Retrieved from http://www.caveon.com/SmithBD.pdf

Street, M. A., Smith, A. B., & Olivarez, A. (2001). The effects of early, regular, and late registration on community college student success: A case study. *Paper presented at the Annual Conference of the Aerican Assocation of Community Colleges.* Chicago, IL. Retrieved from http://www.eric.ed.gov/PDFS/ED454896.pdf

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. Retrieved from http://www.jstor.org.ezproxy1.lib.asu.edu/stable/10.2307/1165221?origin=api

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. Retrieved from http://onlinelibrary.wiley.com.ezproxy1.lib.asu.edu/doi/10.1111/j.1745-3984.2003.tb01149.x/pdf

Wood, T. J. (2009). The effect of reused questions on repeat examinees. *Advances in Health Science Education, 14*:465-473. DOI 10.1007/s10459-008-9129-z.

Zimowski, M.F., Murake, E., Mislevy, R.J. & Bock, R. D. (2003). BILOG-MG 3 (computer software). Lincolnwood, IL. Scientific Software International.

APPENDIX A

SAMPLE BILOG-MG3 SYNTAX

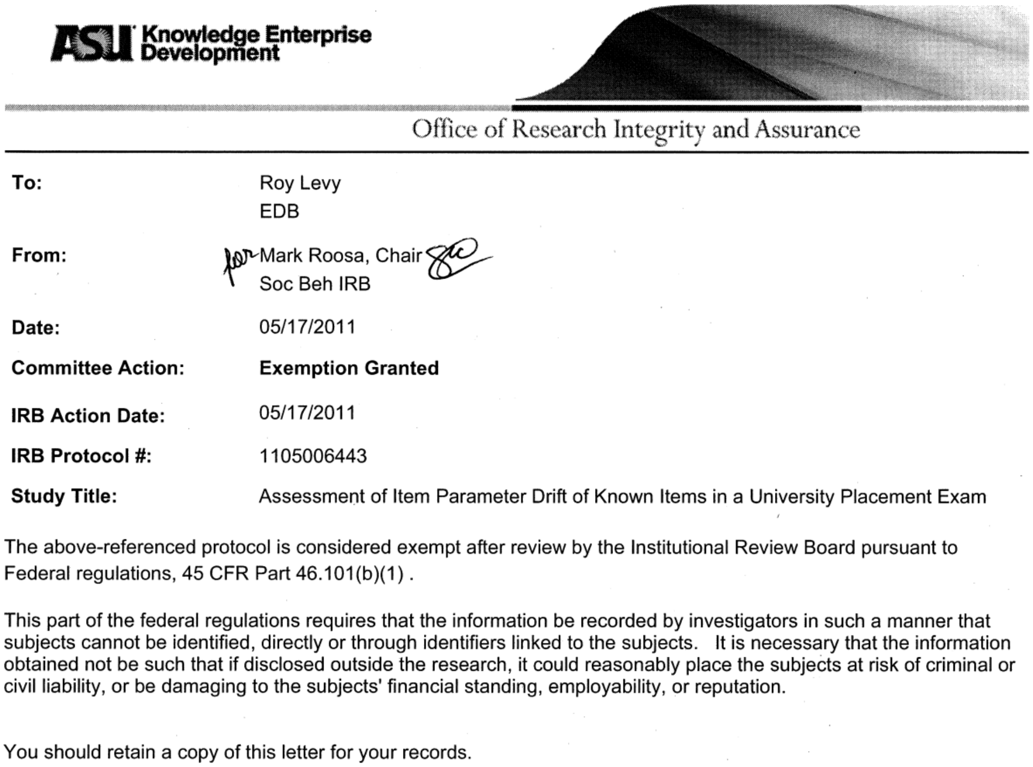
The following BILOG-MG 3 syntax is provided as an example. This particular syntax was used to compare the Early examinees in 2007 to those in 2008,

```
COMMENT DRIFT ANALYSIS FOR GROUPS 1 AND 4 Power 1, Reference 1,
Parm 2

>GLOBAL DFName = 'E:\Thesis data\ThesisMasterDataFile14.txt',
        NPArm = 2,
        LOGistic,
        SAVe;
>SAVE PARm = 'Thesis0911g14p1r1.PAR',
      SCOre = ' Thesis0911g14p1r1.SCO',
      DRIft = ' Thesis0911g14p1r1.DRI';
>LENGTH NITems = (60);
>INPUT NTOtal = 60,
       NIDchar = 4,
       NGRoup = 2,
       Nalt = 5,
       DRIft;
>ITEMS ;
>TEST1 TNAme = 'CALCPLAC',
       INUmber = (1(1)60);
>GROUP1 GNAme = 'GROUP001',
        LENgth = 60,
        INUmbers = (1(1)60);
>GROUP2 GNAme = 'GROUP004',
        LENgth = 60,
        INUmbers = (1(1)60);
>DRIFT MAXpower = 1;
(4A1, 1X, I1, 1X, 60A1)
>CALIB NQPt = 41, NOFLOAT,
       Reference = 1,
       PLOt = 1.0000;
>SCORE RSCtype = 3;
```

APPENDIX B

INSTITUTIONAL REVIEW BOARD EXEMPTION



ASU Knowledge Enterprise Development

Office of Research Integrity and Assurance

To:	Roy Levy EDB
From:	for Mark Roosa, Chair Soc Beh IRB
Date:	05/17/2011
Committee Action:	**Exemption Granted**
IRB Action Date:	05/17/2011
IRB Protocol #:	1105006443
Study Title:	Assessment of Item Parameter Drift of Known Items in a University Placement Exam

The above-referenced protocol is considered exempt after review by the Institutional Review Board pursuant to Federal regulations, 45 CFR Part 46.101(b)(1) .

This part of the federal regulations requires that the information be recorded by investigators in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects. It is necessary that the information obtained not be such that if disclosed outside the research, it could reasonably place the subjects at risk of criminal or civil liability, or be damaging to the subjects' financial standing, employability, or reputation.

You should retain a copy of this letter for your records.