Stability of Item Parameters in Equating Items

Yuan-Ling Liaw

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Education

University of Washington

2012

Committee:

Catherine Taylor

Min Li

Program Authorized to Offer Degrees:

College of Education

University of Washington

ABSTRACT

Stability of Item Parameters in Equating Items

Yuan-Ling Liaw

Chair of the Supervisory Committee:

Professor Catherine Taylor

Area of Measurement, Statistics, and Research Design

Department of Educational Psychology

College of Education

The paper investigates the item factors that may cause item parameter instability. The primary concern of standardized tests is the accuracy of test score interpretation and the appropriateness of test score use across multiple tests. Equating is essential for any testing program. Equating refers to a statistical process used to adjust scores on test forms and establish comparability between alternate forms. Regarding the assumption of item parameter invariance, the statistical properties of the common items are stable across forms. Content and context effects on item parameter estimates appear most likely to violate the assumption will be discussed. Context effects, such as the item type, position adjacency to different kinds of items, wording or appearance and arrangement, as well as content effects, such as instructional and curricular emphasis, have been found to impact item parameter estimates. The data for this study came from the state-level Washington Assessment of Student Learning (WASL) tenth grade mathematics exams administered from 1999 to 2004. Item factors were labeled first. After

labeling item characteristics, the process of test equating and the suspect item identifications was conducted. Two methods, Robust Z-statistics and the signed area between item characteristics curves, were used to detect items that demonstrate item parameter drift. The thesis presents the results of the analyses. Patterns regarding the features of unstable items are described and suggestions for future item development or for selection of anchor items are made.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I: Introduction

Overview

Standardized tests are increasingly used in large scale educational achievement testing programs. The results of testing affect what is taught, how it is taught, what is learned, and how it is learned (Madaus, 1988). In the United States, standardized achievement test results are often linked to high-stakes accountability for students, teachers, and schools. The most important question regarding any achievement testing program concerns the accuracy of test score interpretation and the appropriateness of test score use across multiple tests.

Operational test items are sometimes required to be released after, and sometimes before, test administrations, so that teachers and students have a better understanding of the content covered by the test as well as the test's general format. Disclosure of test materials, when used inappropriately, may cause test security violations that will then result in test score pollution (Haladyna, Nolen, & Haas, 1991) and test score inflation (Koretz, 2003). When there is test score pollution or test score inflation, increases in test scores may not reflect similar real increases in student learning and achievement. For this reason, using multiple test forms across years is recommended in order to ensure that teachers teach to the content standards rather than to the content specifically addressed on the tests (Eignor, Council of Chief State School Officers, & Association of Test Publishers, 2010).

Direct comparisons between examinees who took different forms are not fair unless the scores are adjusted to take into account the statistical properties and context effects for the tests. Meanwhile, for purposes of continuity, a scoring system is needed whereby scores on newer versions can be compared directly with scores on earlier versions (Skaggs & Lissitz, 1986).

Equating is a statistical process used to establish comparability between alternate forms of a test built to the same content and statistical specifications by placing scores on a common scale (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999). Thus, allowing interchangeable use of scores on these forms. Equating allows test users to assess performance change over a year and it allows test developers to compare item parameter estimates from different samples or statistical procedures. For some state assessment programs, the assessments across test administrations, are calibrated using Item Response Theory (IRT), and are equated using the common-item nonequivalent groups design with equating constants for year-to-year equating (Kolen & Brennan, 2004).This design involves the use of a subset of test items ("anchors") in each of the tests forms to be equated. Each testing population takes only one test, but each test form shares a common set of test questions. For standards-based tests, by equating the current year's state test to state tests given in previous years, the performance standard can be maintained over time.

A key assumption made when an equating is performed under the common-item design is

that the statistical properties of the common items that operate as anchors are stable across forms; when two groups respond to two alternate forms, the common items must function similarly in both forms. If two groups of examinees respond differently to the same item, then that item might not be appropriate to be included in the equating process. However, due to random or systematic errors and Item Parameter Drift (IPD) (Goldstein, 1983), the parameters are not always the same as the ones in the item bank when using multiple test forms. Any differences between two populations' overall results could be attributed to the students being different, the test items being different, or both. A primary goal of a common item equating design is to eliminate the effects on scores of these unintended differences in item parameters. When common items are differentially difficult for two groups, and the differences cannot be accounted for by differences in ability of examinees, using them to generate an equating transformation is questionable. Test equating is necessary to be fair to examinees taking different test forms and to provide score-users with scores that mean the same thing, regardless of the items appearing in different test years and taken by different examinees.

The purpose of this study is to investigate the item factors that may cause item parameter instability.

CHAPTER II: Literature Review

Item Response Theory (IRT)

Today, Item Response Theory (IRT) is used by large test publishers and state departments of education to investigate item bias, to equate tests, and to report test score information. The purpose of IRT is to provide a basis for making predictions, estimates, or inferences about examinees' latent abilities or traits measured by a test using an equal interval scale. The IRT model specifies the relationship between a latent trait and observed performance on the test that is designed to measure that trait (Hambleton, Swaminathan, & Rogers, 1991). With IRT, each item is described by a set of parameters that can be used to graphically depict the relationship between an item and a latent trait through use of an Item Characteristic Curve (ICC) (Holland, 2007). The curve is defined by three parameters: a, which is proportional to the slope of the curve at the point of inflection and represents the discriminating power of the item; b, the location of the item on the underlying scale; θ, at the point of inflection and representing the difficulty of the item in relation to the underlying scale; and c, the lower asymptote — alternatively called the pseudo-guessing parameter — corresponding to the probability that a person lacking in proficiency will answer the item correctly.

One basic advantage of IRT methods is that, if data fit the model reasonably well, it is possible to obtain invariance of item and ability parameters. Parameter invariance, the idea that

IRT item parameter estimates remain unchanged across various groups of examinees and that ability estimates from items drawn from the same calibrated pool, will remain invariant, gives IRT its applicability and usefulness. That is, for a set of calibrated items, an examinee will be expected to obtain the same ability estimate from any subset of items. Moreover, for any subsample of examinees, item parameter estimates will be the same. In other words, when an IRT model fits the data, the same ICC is obtained for the test item regardless of the distribution of ability in the group of examinees used to estimate the item parameters. Hence, the ICC is invariant across the two samples from the same population. The parameters that characterize an item do not depend on the ability distribution of the examinees. The parameter that characterizes an examinee does not depend on the set of test items, either.

According to the invariance property of IRT, item parameters estimated from different samples of the same population are supposed to be invariant, even over different measurement occasions (Wells, Subkoviak, & Serlin, 2002). Items that behave consistently over multiple administrations with equivalent samples from the same population are appropriate for use in the test equating process.

Challenges of Test Equating - Item Parameter Drift (IPD)

An important property of IRT is that item parameters calibrated on samples from the same population will be invariant, even over different testing administrations. Clearly, parameter

invariance gives IRT its applicability and usefulness (Linn, 1993; Hambleton et al., 1991). However, item parameters do not always remain invariant over time and occasions, due to factors other than sampling error. For example, it is likely that examinees from two consecutive years of an achievement testing program are not equivalent groups. As educators adjust curriculum and instruction to align with tested curriculum standards, student abilities from one year to the next year will not be equivalent. In fact, the researchers might be concerned about standards-based testing programs if performances did not improve. To address these differences, equating methodologies use the parameter values of stable equating items to anchor the IRT scale. When items are not stable (i.e., invariance does not hold), the item is considered to be drifting from its original parameter value. More precisely, Goldstein (1983) defined it as Item Parameter Drift (IPD) which refers to the differential change of item parameters over subsequent testing occasions.

In practice, there can be other challenges to a successful equating. Holland (2007) pointed out that one of the primary problems with test equating is “security breaches”. Typically, questions on standardized tests are kept secure and hidden from students and teachers until the time of testing. Statistical validity and reliability estimates depend on the test being secret, and, when a test is kept secret, the items can undoubtedly be used again. However, research during the 1980s and 1990s (Haladyna et al., 1991) showed that “teaching tested content” (which differs

from teaching tested standards) was widespread when tests were used for high stakes decisions. As a result, item exposure makes it difficult for test developers to examine test score validity and interpret test scores. The validity of standardized test scores is compromised and undermined when improper test preparation strategies are employed based on disclosure of test materials.

For test security and test disclosure reasons, many testing programs use multiple test forms. These forms are constructed based on the same specifications so that they are similar to each other in content and statistical characteristics. Updating of items in tests at regular intervals is desirable – both to keep the content abreast of changing social contexts and to protect the test items from overexposure. Test score equating is basically the process of determining the relationship between raw or scaled scores on two or more test forms (Linn, 1993; Kolen and Brennan, 2004), especially when tests measure the same construct in the same way, in the same structure, timing, item types, and formats, so that scores on different forms can be used interchangeably (i.e., parallel tests).

Taylor and Lee (2010) also mentioned that, item releases and multiple forms are two factors that create challenges for test developers in terms of managing the equating and comparability of the content of different test forms. Testing programs may be required to regularly release items after, and sometimes before, test administrations, so that teachers and students have a better understanding of the content covered by the test as well as the test's general format. However,

using items more than once in operational administrations may be a concern, particularly, when some examinees have access to the questions in advance. Test scores may systematically increase without improvement of actual performance on attributes being assessed.

Furthermore, to insure validity of scores across tests, equating items, which are referred as common items or anchor items, need to be used more than once. It is prevalent to use the same items or close analogues in different forms to provide strong support for the validity of the newest form. In practice, testing programs use multiple test forms but each form involves an anchor item set. Therefore, common items provide the statistical means for equating test forms and making scores from different administrations of the same testing program comparable.

According to IRT, each item has only one true set of item parameter values. Holland (2007) suggested for equating design to be effective, equating items must maintain their statistical properties across the old and new forms. Only when measurement invariance does hold over two or more occasions, will the observed scores be directly comparable. Therefore, it is critical to maintain stable item parameters so that scores remain comparable across administrations and test forms.

Sources for Item Parameter Drift (IPD)

Item parameter drift can occur for a number of reasons. Bock, Murakl, and Pfeiffenberger (1988) have noted that IPD effect could be a result of educational, technological, or cultural

changes. These changes could cause some items to become more or less difficult relative to other items over the period in which these items are used. Therefore, during the equating process, the common items between the two tests being equated are checked for IPD.

Several studies have been conducted to explore the threats to item parameter invariance existing at the item level. Context effects, such as the item type (Taylor & Lee, 2010), position within the test (Whitely & Dawis, 1976; Haertel, 2004; Meyers, Miller & Way, 2009), adjacency to different kinds of items (Yen, 1980; Kingston & Dorans, 1984; Eignor, 1985), wording (Gierl & Khaliq, 2001) or appearance and arrangement (Cizek, 1994). Further, content effects, such as instructional variation (Stone & Lane, 1991; Smith, 1991; Haladyna et al., 1991) and curricular emphasis (Yen, Green, & Burket, 1987; Traub, 1983), have been found to impact item parameter estimates. Concretely, items are constructed with specific item features, item formats, content strands, and different levels of content complexity in order to assess the different levels of ability. Thus, characteristics of each item could also affect its item parameter values.

Researchers have found that item format can affect the stability of item difficulty. Multiple-choice items and constructed-response items are widely used and the nature of the stimuli and responses to them are different. Taylor and Lee (2010) used statewide, standards-based reading and mathematics tests for grades 4, 7, and 10 from 1997 to 2001 and found that the parameters of constructed-response items, which were used as anchor items in an

equating process, were less stable than the parameters of dichotomous items. Practice on remembered anchor items, or on items similar to the anchor items, or on the skills required by constructed-response items might lead to more instability in item parameters for polytomous items than for dichotomous items.

The context in which items are presented influences the estimates of item parameters. "A context effect occurs when a change in the test or item setting affects student performance" (National Research Council, 1999, p. 34). Numerous studies have examined the effects of changes in item order or context on test performance (Whitely & Dawis, 1976; Haertel, 2004; Meyers et al., 2009). Whitely and Dawis (1976) found significantly different Rasch item difficulty estimates for 6 of 15 core items that differed only in item position across forms. Haertel (2004) found that differences in linking item position between the year 1 and year 2 tests caused anomalous behavior in linking items. Meyers et al. (2009) suggested that item position effects can vary depending on factors such as how much the item position changes, the direction of change (i.e., toward the beginning of the test versus toward the end of the test), and the ability levels of the test-takers.

In addition to impacting item difficulty, change in item location and appearance has been shown to affect equating results (Yen, 1980; Kingston & Dorans, 1984; Eignor, 1985). Yen (1980) investigated item location on the California Achievement Test (CAT) and found some effects on

parameter estimates as well as an impact on equating results. Similarly, Kingston and Dorans (1984) found that equating coefficients were often sensitive to item location effects. The equating results were often poor, particularly with the Quantitative and Verbal sections of the GRE, when common items were not in the same operational location. Eignor (1985) also studied the effects of item location on pre-equating of the SAT. Results often showed large differences between pre-equating and operational equating transformations, again particularly with the Quantitative section of the test.

As for other detailed item formats, Gierl and Khaliq (2001) identified sources of differential item and bundle functioning on translated achievement tests. They assumed that differences in punctuation, capitalization, item structure, typeface, and other formatting usages are likely to affect the performance for one group of examinees. This source of differences is comparable to the source described by Allalouf, Hambleto, and Sireci (1999) as "changes in format" because both sources focus on item format differences. According to their results, translation DIF associated with format differences, while rare, was reliably identified and predicted by the translations. Both studies concluded that item format must be consistent across language forms, and this consistency should be monitored during test development.

Additionally, anchor items shown in different appearance and arrangement on the page of test between two forms can change item parameter values as well. Anchor items are expected to

be presented on the page in the same way each year of testing. Also, the response alternates for multiple-choice items should appear in the same order in the old and new forms. Cizek (1994) found that if a common item is associated with stimulus materials that were used with a set of items in the old form, then the entire set of items associated with those stimuli should be included on the new form to avoid context effects.

Apart from item context effects, researchers also considered instructional effects (Stone & Lane, 1991; Smith, 1991; Haladyna et al., 1991). Stone and Lane (1991) stated that different instructional approaches, different emphases on instructional strands, and different sequencing of instructional content can all affect the stability of item parameters over time. The tests have to be changed to reflect the current curricular innovations. Smith (1991) observed an unfortunate alignment of testing and teaching. She found that pressure to improve students' test scores caused some teachers to neglect untested materials. Haladyna et al. (1991) suggested that any practice that improves test performance without concurrently increasing actual mastery of the content tested produces "score pollution". The score does not represent actual academic achievement because it is "polluted" by unrelated factors. For example, practicing test items before a test produces score pollution because the scores no longer measure generalized mastery but only ability to memorize specific familiar item.

Yen et al. (1987) noted that different local curricular and instructional characteristics

influenced parameter estimates. They presented a case where the IRT difficulty parameter estimates in the national calibration of a mathematics test changed systematically at the local level. They compared the item difficulty value obtained for the national norm group with those obtained for a local educational agency for grade 5 Mathematics Concepts and Applications. In general, the items in measurement strand were relatively more difficult, while the items in numeration strand were relatively easier.

Generally, Traub (1983) concluded that not only content considerations but also additional factors such as instructional differences, individual differences in learning, and differences in test taking behaviors can all contribute to IPD in a particular group of test items.

Methods of detecting Item Parameter Drift (IPD)

According to Donoghue and Isham (1998), with only two time points, the drift problem is the same as Differential Item Functioning (DIF). Studies have examined changes in item parameters over repeated administrations using two time points to compare two subgroups – the two nonequivalent samples. In research by Donoghue and Isham (1998), they compared the results of a number of DIF measures for detecting IPD which covered two time points with one year apart. This strategy also used by Wells et al. (2002) as well as Stone and Lane (1991). The problem of IPD is formally identical to that of DIF: Does the item function the same for two groups of examinees? When DIF analyses examine whether items function differently in

examinee subgroups (e.g., males and females), IPD is particularly relevant to time of testing; however, the underlying question is the same. Therefore, researchers suggest that any DIF procedure could be utilized to examine IPD (Bock et al., 1988; Donoghue and Isham, 1998).

Many approaches to detecting DIF have been proposed to examine IPD. The first method is the use of Z-score. Within the context of IRT, Lord (1980) noted a procedure for testing the differences in item parameters between two groups. To test the hypothesis that a single parameter, for example b_t (where t is the time of testing), differs between two groups, Lord proposed comparing, for a given item, the difference between the estimated b_t divided by the estimated standard error

$$d_i = \frac{(b_1 - b_2)}{SE(b_1 - b_2)}$$

$$SE(b_1 - b_2) = \sqrt{var(b_1) + var(b_2)}.$$

The sample sizes involved in IRT calibration are usually so large that the degrees of freedom for these estimates are effectively infinite, so that the probability statements could be made by referring d_i to tables of the standard normal distribution.

$$d_i \cong Z_i$$

However, the Z-score method is not always reliable because the mean and standard deviation are influenced by the outliers. The Robust-z statistic (Tenenbaum et al., 2001) has emerged as particularly promising for use in standardized testing programs. The Robust-z

requires setting item parameters free to be estimated in a new (at testing time 2) and comparing the freely estimated difficulties with their anchor item parameters:

$$\text{Robust} - Z_i = \frac{\left(b_{iF} - \widehat{b_{iE}}\right) - M_d}{IQR_d \times 0.74}$$

where b_{iF} stands for the fixed anchor item difficulty, and $\widehat{b_{iE}}$ denotes the estimated item difficulty for item i at the second times of testing. The median and inter-quartile range of the differences between the fixed values and the item difficulty estimates across all the items in the anchor set are represented by M_d and IQR_d respectively. For the normal distribution, the IQR is equal to $1.35 \times SD$ or $SD = 0.74 \times IQR$. With the quantity $0.74 \times IQR$ emulating the standard deviation, a robust version of the traditional z statistic can be taken as the ratio in the formula above. If the absolute value of the robust-z for an item is equal to or larger than 1.96, the item is typically flagged as drifting.

Another way to look at DIF is to estimate two sets of item characteristic curve (ICC) for each item. ICC provides the probability of examinees answering an item correctly for examinees at different points on the ability scale. With the one-parameter logistic (1PL) or Rasch model, the item characteristic curves vary only in their difficulty and differ only by a translation along the ability scale. With the two-parameter logistic (2PL) model, item characteristic curves vary in both slope and translation along the ability scale. For example, some curves increase more rapidly than others when the corresponding test items are more discriminating than others.

Finally, with the three-parameter logistic (3PL) model, curves may differ in slope, translation and lower asymptote. The ICCs are completely determined by their corresponding item parameters, thus DIF can be identified by comparing the item parameters to tell if two ICCs are different or not. Rudner, Getson, and Knight (1980) further proposed to compare the ICCs by evaluating the signed area between two ICCs. After placing the parameter estimates on a common scale, if the ICCs are identical, then the area between them should be zero. On the other hand, when the area between ICCs is not zero, DIF is present.

In computing the signed area, A_i, the numerical procedure involves (a) dividing the ability range into k intervals, (b) constructing rectangles centered around the midpoint of each level, (c) obtaining the values of the ICCs (the probabilities) at the midpoint of each interval, (d) taking the absolute value of the differences between the probabilities, and (e) multiplying the difference by the interval width and summing (Hambleton et al., 1991). This procedure can be expressed for item i as

$$A_i = \sum_{\theta=r}^{s} |P_{i1}(\theta) - P_{i2}(\theta)| \, \Delta\theta$$

where $\Delta\theta$ is the width of the interval and is chosen to be as small as possible (e.g., 0.01). The values r and s indicate the ability range over which the area is to be calculated. The choice for the ability range should ensure that the area is calculated over the ability range in which nearly all

examinees fall. The values are suggested to the range from three standard deviations below the lower group mean ability to three standard deviations above the upper group mean ability.

Purpose of the study

The purpose of this study was to investigate the item factors that may cause item parameter instability. In order to minimize the threats to item parameter invariance existing at the item level, testing programs measure the same constructs and are usually built to the same test specifications or test blueprint. However, different editions or forms of a test almost always differ somewhat in their statistical properties. Few studies have examined the stability of anchor items after controlling for the context effects between two test forms (as much as possible) and compared the characteristics of stable items and unstable items. Even though it may be difficult to identify content or context effects that have influenced item performance with certainty, considering a range of plausible explanations can help to decide whether a misbehaving item should be discarded from the common-item pool or not.

This study explored the issues that need to be addressed to better understand how anchor test design can maintain common items' statistical properties across multiple forms and make test equating effective.

Research Questions

1. The first question set relates to item characteristics. Do *Content* and *Process Strands* affect

the stability of item parameters? Are the parameters of constructed response items less stable than the parameters of multiple choice items? Are the parameters of routine items less stable than the parameters of *Comprehension* and *Nonroutine/Insightful* items? Are the parameters of *Story/Concrete* items less stable than the parameters of *Text/Abstract* items?

2. The second question relates to item structure which consists of stem, stimulus, and options in the multiple-choice item and the required answer in constructed-response item. Do any of their characteristics affect the stability of item parameters?
3. The third question considers the changes in common item over years. Do changes in wording affect the stability of item parameters? Do changes in context affect the stability of item parameters?

CHAPTER III: Methodology

This study used real data to investigate the potential causes of IPD. First, Rasch equating procedures were replicated and items with parameters that have drifted more than expected amount ($Z > 1.96$) were identified statistically by Robust Z-statistics. As a second strategy, the area between a pair of common item response functions was computed to investigate the difference involving item difficulty, discrimination and guessing parameters. Next, possible causes of parameter drift were investigated. Once unstable items have been identified using the Robust Z-statistics and area measure methods, the characteristics of unstable items were compared with those of stable items. Through this analysis, patterns in terms of the item characteristics may be causing item parameter drift were identified.

Subjects

The data for this study came from the state-level Washington Assessment of Student Learning (WASL) tenth grade mathematics tests. Six administrations from 1999 to 2004 were inculded. A total of 30,000 students randomly selected were included in this study. In these data sets, there were 6,000 examinees sampled in each year of testing from 1999 through 2004.

Instruments

The Washington Assessment of Student Learning (WASL) mathematics test was a large-scale statewide mandated assessment. The test was used from spring 1997 to summer 2009.

All test forms followed the same clearly specified standardized procedures for each administration.

The WASL aimed to measure the level of mathematics proficiency that Washington students had achieved, according to the Essential Academic Learning Requirements (EALRs) established by the Washington Commission on Student Learning (CSL). The EALRs for mathematics consist of mathematics concepts and procedures and four fundamental processes (solves problems, reasons logically, communicates understanding, and makes connections.)

The WASL mathematics test included five content strands: Number Sense, Measurement, Geometric Sense, Probability/Statistics, and Algebraic Sense; and three process strands: Solves Problems and Reasons Logically, Communicates Understanding, and Makes Connections. These content and process strands represented different mathematical knowledge and skills but were correlated to some degree. In keeping with the EALRs Technical Manual (Washington State CSL, 2001), the EALRs in mathematics (the content and the process strands) were viewed as an integrated whole. However, each item was identified as to the primary content or process strand it is assessing.

The WASL used three types of items on the mathematics test: multiple choice, short answer, and extended response. For each multiple-choice item, students selected the one best answer from four choices provided. Each multiple-choice item was worth one point. These items were

machine scanned and scored. The other two "open-ended" item types were short answer and extended response which required students to solve multistep mathematics problems, draw diagrams or graphs, write explanations of procedures, show the procedures used, etc. Short-answer items were worth two points (scored 0, 1, or 2) and extended-response items were worth four points (scored 0, 1, 2, 3, or 4). These items required hand-scoring by well-trained professional scorers. According to WASL, only multiple-choice (MC) items and two point short-answer (SA) items were used as anchor items for purpose of equating. Table 1 shows the numbers of anchor item by item strand and item type for each paired year. It is evident that many more multiple-choice were used for equating purposes than short-answer items.

In order to identify common items that failed to meet the IRT parameter invariance assumption in test equating, the procedure conducted by Office of Superintendent of Public Instruction (OSPI) was as follows. Each year, a new version of the WASL mathematics test was created by sampling from a large pool of questions. OSPI used common item, non-equivalent groups equating procedures to maintain the same performance standard from year to year and to provide longitudinal comparisons across years, even though different questions were used, for example, with the year 2000 test being linked back to the 1999 test.

Operationally, the Rasch model was used to calibrate item parameters with the BIGSTEP computer program. Then, the two stage procedure for test equating was used (see Taylor & Lee,

2010, for a description). First, all candidate anchor items on the operational test were subjected to a stability analysis to determine the final anchor item set in the year-to-year common item equating. In practice, item difficulty estimates for the anchor items within each test were obtained for each year. The means of the item difficulties for the anchor items was computed separately for each test year. The difference between the means was defined as the "equating constant". In other words, based on the Rasch model, the equating constant was the difference between the mean of the difficulty parameters in a specific year and the mean of the difficulty parameters in the previous year. Item difficulty parameters from the previous year were also called the "bank values" for the items. The "adjusted item difficulty" was then computed by adding the equating constant to the item difficulty of each anchor item from the second year of testing. The anchor item difficulties from the first year and the adjusted anchor item difficulties from the second year were expected to be equal. The adjusted anchor item difficulties in the second year were compared with the bank values. Any item with an absolute difference greater than .30 between the bank value and the adjusted item difficulty in the second year was dropped from use as an anchor item.

The numbers of items administered ranged from 44 to 46 in each test year. In terms of each test form, 12 to 15 linking items were embedded in tests for equating to test scores of other forms. Common-item nonequivalent groups design was used in the WASL for data collection. To

eliminate item position effects, equated items were administered in the same or almost the same position in each year's test.

Table 1

Distribution of equating items by item strand and format

Exam Date (Years)	99-00		00-01		01-02		02-03		03-04		Total	
	MC	SA	MC	SA	MC	SA	MC	SA	MC	SA	MC	SA
Content Strand												
Number sense (NS)	2	0	1	0	2	0	0	0	3	1	8	1
Measurement (ME)	4	0	2	0	2	1	1	0	1	0	10	1
Algebraic sense (AS)	1	1	2	1	1	0	2	1	4	0	10	3
Geometric sense (GS)	2	1	2	0	2	0	0	0	2	0	8	1
Probability and statistics (PS)	2	1	2	0	1	1	3	0	0	1	8	3
Process Strand												
Solve problems/Reason logically (SR)	1	0	0	1	1	0	1	1	1	0	4	2
Communicate understanding (CU)	0	0	0	0	0	1	2	1	1	0	4	1
Make Connections (MC)	0	0	1	0	0	0	0	0	0	0	1	0
Subtotal	12	3	10	2	9	3	9	3	7	2	47	13
Total	15		12		12		12		11		60	

Evidence for the validity and reliability of scores from the WASL mathematics test was presented in the annual technical reports (see, e.g., Taylor, 1999a, 1999b, 1999c). Validity studies included correlations among scores from the WASL tests and scores from national standardized achievement tests, factor analyses, and regression analyses. The data presented in the technical reports provided strong support for the validity of WASL mathematics scores. Annually, correlations among mathematics strand scores were moderately strong. Factor analyses of WASL mathematics and reading strand scores showed two distinct factors with strand scores loading as expected on these two factors. Factor analyses of WASL on reading and mathematics strand scores and reading and mathematics subtest scores from the Iowa Test of Educational Development (ITED) showed that WASL mathematics subtest scores and ITED mathematics subtest scores loaded together on one factor. Scores from WASL reading strands and ITED reading subtests loaded on a second factor.

Reliability evidence was derived from three sources: inter-judge agreement for scores on constructed-response items, correlations between total scores resulting from different scorers, and alpha coefficients. Alpha coefficients for reading and mathematics across grade levels ranged from 0.89 to 0.92. Exact agreement on item scores for constructed response item between pairs of raters ranged from 70%-94% across grade levels, content areas, and years. Correlations between total scores from different scorers ranged from 0.98 to 0.99 across the five years of

testing.

Procedure

Item coding

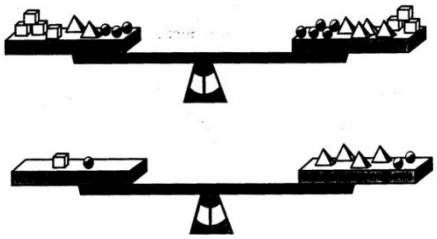
To examine potential cause of IPD, items were analyzed by their content and features. First, all equating items were categorized based on their *Item Strand, Item Format, Cognitive Complexity,* and *Item Concreteness*. *Item Stand* follows WASL's mathematics test and item specifications which include items designed to assess students' knowledge about mathematical content strands (number sense, measurement, geometric sense, probability/statistics, algebraic sense) and mathematical process strands (solve problems, communicate understanding, make connections). *Item Format* is either multiple-choice (MC) or short answer item (SA).

As for *Cognitive Complexity*, each item was coded into three categories, *Routine, Comprehension* and *Non-routine/Insightful*. Routine items required examinees to recall factual knowledge and/or perform mathematical manipulation; *Comprehension* items required examinees to solve problems that demonstrated comprehension of mathematical ideas and/or concepts, and *Non-routine/Insightful* items required novel approaches or careful analysis of the problem situation. Examples are shown in Table 2. First, "Measure of angle in parallelogram" item is a *Routine* item which asks the idea of the sum of adjacent angles in a parallelogram. Based on the properties of parallelograms, examinees are expected to recall factual knowledge

and find the solution. Second, “The greatest rectangular garden area” is an instance of *Comprehension* item. This item asks the relationship between perimeter, side lengths and area in a rectangle. A given perimeter is able to enclose different rectangles. Examinees were expected to observe patterns and the relationship between length and width that allowed them to find the largest area. And then, “Objects balanced on scale” item belongs to *Nonroutine/Insightful* item. The balanced scales imply an equality of measure which is represented in pictures. The weight of various solids was specially designated at the stage of item development. No fixed knowledge of the weights of the sphere, cube and triangular pyramid weight existed. Examinees were expected to use any formal algebraic procedures to solve the problem and translate graphical representation to equations. However, students could approach the problem in whatever way they understood as well.

Table 2

Examples of Cognitive Complexity Coding

Category	Example
Routine	"Measure of angle in parallelogram" in 2000 (24) and 2001 (26): **24** In parallelogram *PQRS*, the measures of angle *P* and angle *R* are each 146°. What is the measure of angle *Q*? ○ **A.** 146° ○ **B.** 112° ○ **C.** 68° ○ **D.** 34°
Comprehension	'The greatest rectangular garden area" in 1999 (4) and 2000 (4): **4** A gardener has twenty feet of fencing to surround a new **rectangular** garden. Which of the following is the greatest area that can be enclosed? ○ **A.** 20 square feet ○ **B.** 24 square feet ○ **C.** 25 square feet ○ **D.** 28 square feet
Nonroutine/Insightful	"Objects balanced on scale" in 1999 (32) and 2000 (32): **32** Kent is using the scale to compare the weight of various solids.  How many spheres will balance one cube? ○ **A.** 2 spheres ○ **B.** 3 spheres ○ **C.** 4 spheres ○ **D.** 5 spheres

Item Concreteness was coded into three categories, *Abstract, Contextual and Concrete/Story*. *Abstract* items referred to symbolic problems with no story-based content. Their symbolic expressions with variable quantities allowed examinees to ignore text to solve the problems. *Concrete/Story* items were word problems that required examinees to read the story situation carefully and also required the additional step of translating words to symbols. Problems written in verbal formats, including productive story contexts, required examinees' informal understanding of verbal scenarios. *Contextual* items were also word problems. Their variable quantities were written in text. However, there was no redundant information relevant to solving the problems that the story scenario added. The examples are shown in Table 3.

First, "Factor of an expression" item is a typical *Abstract* item. Numerical and symbolic approaches are offered only. In other words, the item is shown in concise and general representation. Next, "Original cost price before discount" item is categorized as *Concrete/Story* item. The item creates a natural environment for understanding its context and for communicating its solution, i.e., shopping experience. It emphasizes the concrete application and the connection between arithmetic and other domains of everyday life. Here, the prices and discount are logical and reasonable. At the same time, higher language comprehension demands are required. Third, "Volume of cylindrical storage tank" item is characterized as *Contextual* item. The variables are represented in the text (i.e., height and radius); however, the problem lacks an

action or a situated background.

Table 3

Examples of Item Concreteness Coding

Category	Example
Abstract	"Factor of an expression" in 2001 (7) and 2002 (7): **7** Which term is a factor of $3a^2 + 12a$? ○ **A.** $3a$ ○ **B.** $4a$ ○ **C.** $3a^2$ ○ **D.** $4a^2$
Concrete/Story	"Original cost price before discount" in 2002 (11) and 2003 (11): **11** Annette bought a coat from a department store at 35% off the regular price. The sale price was \$130. Which of the following questions would help Annette to find the original cost of her coat? ○ **A.** What is 65% of \$130? ○ **B.** How much is \$130 + 35% of \$130? ○ **C.** \$130 is 35% of what number? ○ **D.** \$130 is 65% of what number?
Contextual	"Volume of cylindrical storage tank" in 2000 (34) and 2001 (36): **34** If the height of a cylindrical storage tank is 11 m and the radius is 10 m, what would be its volume? ($\pi \approx 3.14$) ○ **A.** 314 m^3 ○ **B.** 691 m^3 ○ **C.** 1,100 m^3 ○ **D.** 3,454 m^3

Second, further item specifications were coded, i.e., the structure of item, the stimulus used in the question, and the stimuli used in the options in the multiple-choice item or the format requested in the short-answer item. The structure of item compares the relative location between stimulus and stem. It was coded into four categories, Stimulus after Stem, Stimulus in the middle of Stem, Stimulus before Stem, or No Stimulus.

Next, the stimulus used in the question was coded as well. *Expression* includes using numbers, variables, and equation. *Graphic or Chart* refers to items where statistics data were displayed in the form of a histogram, line graph or table. *Geometric Figure* refers to items that used figures to illustrate geometric concepts including triangle, circle, angle, quadrilateral, and polyhedral. *Grid* refers to items that included a grid with a X-axis, Y-axis, origin, coordinates, points and lines. *Picture of Phenomena* refers to the visual information that is required in order to answer the question. *Object Cue* refers to the visual display used in the item is not necessary for solving the problem; however, it reminds the examinee about the objects in the item. Finally, *Text* refers to items that used paragraph to describe problem situation. Examples of items that used *Picture of Phenomena* as stimuli are shown in Table 4. The "Objects balanced on scale" item was previously shown as an example for *Nonroutine/Insightful* item in the previous paragraph. The relationship of the weight of three kinds of solids, sphere, cube and triangular pyramid, is presented in the picture. "Probability of sum on spinners" item comprises two different spinners.

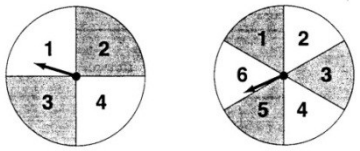
According to the picture, one spinner is divided by four sections; however, the other was divided by six. The “Explanation of “Drama” in diagram” has a Venn diagram with three circles to present the possible logical relations among three subjects, Music, Art, and Drama. The diagram comprises overlapping circles with different sizes as well. The picture in the “Area of carpet in triangular showroom” gives the information to the examinees that the triangle described in the text is a right triangle.

Table 4

Examples of "Picture of Phenomena" Stimulus Coding

"Probability of sum on spinners" in 2001 (3) and 2002 (3):

3 In a certain carnival game, a player gets to spin each of the following spinners once. What is the probability of getting two numbers that have a sum of 7?



○ **A.** $\frac{1}{4}$

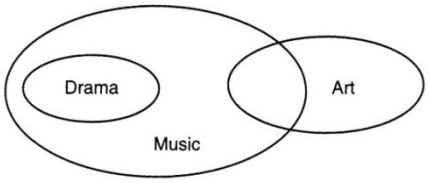
○ **B.** $\frac{1}{6}$

○ **C.** $\frac{5}{12}$

○ **D.** $\frac{7}{24}$

"Explanation of "Drama" in diagram" in 2001 (14) and 2002 (14):

14 Which of the following would be a reasonable explanation for the placement of the oval labeled "Drama" in the diagram below?



○ **A.** Students must complete drama before they may take music.

○ **B.** Students must complete music before they may take drama.

○ **C.** Students must complete both of the other courses before they may take drama.

○ **D.** Students must complete drama before they may take either of the other two courses.

"Area of carpet in triangular showroom" in 2002 (37) and 2003 (37):

37 Luis wants to put carpet in the triangular showroom shown below. He knows the width of the room is 3 feet more than $\frac{1}{3}$ the length of the room. If the length of the room is 21 feet, how many square feet of carpet does he need?

○ **A.** 84 square feet

○ **B.** 105 square feet

○ **C.** 168 square feet

○ **D.** 210 square feet

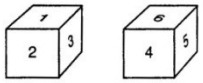
Those items that used “Object Cue” as stimuli are summarized in Table 5. In the “Picture of Phenomena” items, for which the main information is presented in the pictures, examinees are unable to ignore the visual representation to solve the problems. Instead, the pictures presented in “Object Cue” items images assist examinees in understanding text. For example, as for “Sum of two rolled number cubes” and “Possible outcomes of cube and coin” items, if students aren’t allowed to play with dice (number cubes), the picture would “cue” them to the object being discussed in the item. Further, in terms of “Shadow length of hoop from ratio”, the picture provides the image of shadows of person and basketball hoop system behind the verbal representation.

Table 5

Examples of "Object Cue" Stimulus Coding

"Sum of two rolled number cubes" in 2000 (29) and 2001 (29):

29 Akio and Tamera are playing a game with two number cubes, each labeled 1–6. On each turn, the person rolls both number cubes. The sum of the two numbers on top of the number cubes tells how many spaces that person can move the game piece.

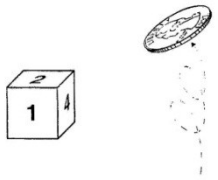


Akio needs to move nine spaces to win the game. What is the probability that Akio will roll a sum of at least nine on his next turn?

- ○ **A.** $\frac{1}{10}$
- ○ **B.** $\frac{1}{4}$
- ○ **C.** $\frac{5}{18}$
- ○ **D.** $\frac{4}{11}$

"Possible outcomes of cube and coin" in 2003 (9) and 2004 (9):

9 Joey and Rochelle made up a game in their mathematics class. To earn points in the game, each player rolls a six-sided cube with numbers 1 through 6 on the sides and then flips a coin. If the coin lands "tails up," the player gets a total number of points equal to the number at the top of the cube. If the coin lands "heads up," the player's points are doubled for that turn.

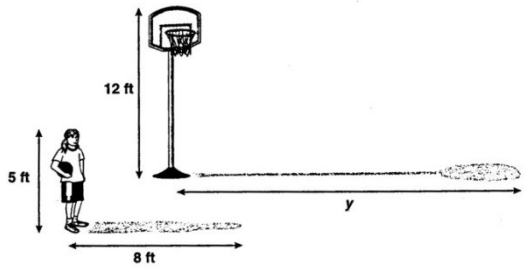


In the box below, list all the possible outcomes for each turn. Then indicate the probability of a player getting 6 points in one turn.

Table 5

Examples of "Object Cue" Stimulus Coding (Continued)

"Shadow length of hoop from ratio" in 2000 (2) and 2001 (2):



2 A 5-foot-tall person casts an 8-foot shadow. If a vertical pole that supports a basketball hoop is 12 feet high, how long is its shadow?

- ○ **A.** 7.5 feet
- ○ **B.** 15 feet
- ○ **C.** 19.2 feet
- ○ **D.** 25 feet

The codes above focused on the item itself. Since equating items were used twice, items were coded in terms of any changes and variations in the characteristics (i.e., item number change and its distance between two years, the length of item based on the numbers of page, the position of item appearance and arrangement of the page, the use of keyword, and other specifications) between the two years.

The change in *Item Location* indicated that the labeled item number was moved forward, moved backward or stayed on the same location. The change in *Number of Page* indicated that the layout on the test page was extended from half page to full page, shrunk from full page to half page, or kept on full page or half page. The change in *Item Position* indicated that the item was moved from left page to right page on the booklet, or was moved the contrary way, or stayed on the right or left page.

As for those items that were on a half-page, the characteristics of their accompaniments were considered as well. For example, the equating items may have moved from bottom to top or moved from top to bottom. The item formats of two items on the same page were either identical or different which were coded.

The overall coding plan is summarized in Table 6.

Table 6

Item Coding and its Hypothesis

Item Characteristic	The Hypothesis of Item Difficulty
Item Strand Number Sense Measurement Algebraic Sense Geometric Sense Probability And Statistics Solves Problem/Reason Logically Communicates Understanding Makes Connections	The different item strands will affect the stability of item parameters
Item Format Multiple-Choice Short-Answer	The parameters of SA items will be less stable than MC items
Cognitive Complexity Routine Comprehension Nonroutine/Insightful	The parameters of Nonroutine/ Insightful items will be less stable than Routine and Comprehension items
Item Concreteness Abstract Contextual Story/Concrete	The parameters of Story/Concrete items will be less stable than Text/Abstract items
Item Structure The relation between the stimulus and stem	The parameters of items with stimulus will be less stable than the items without stimulus. For items with stimulus, the parameters of items with stimulus in the middle of stem will be relatively stable.
Stimulus used in the question Expression Graph or Chart Geometric Figure Grid Picture of Phenomena Object Cue Text	The parameters of graphing items will be less stable than the parameters of expression and text items.

Table 6

Item Coding and its Hypothesis (Continued)

Item Characteristic	The Hypothesis of Item Difficulty
Change in Item Appearance	
Item location number change and its distance between two years	The changes will affect the stability of item parameters
The length of item based on the numbers of page	
The position of item appearance and arrangement of the page	
The use of keyword	

Statistical Analysis

After labeling item characteristics, the process of test equating and the suspect item identifications were conducted. The two separate sets of parameter estimates for the anchor items in the two groups were generated: the bank values and freely calibrated parameter estimates.

Two methods were used to compare the performance on suspect items over two years. First, anchor items were checked for stability by examining whether adjusted item difficulties in each successive year were nearly the same as the original bank values. Robust Z-statistics were used to detect item parameter drift. If the absolute value of the Robust-Z for an item was larger than 1.96, the item was flagged as drifting or showing DIF between test years. Second, using an area model for detecting item parameter drift, for each item, the ICC was computed for each group, and the two ICCs were placed on the same scale, and then compared. If the area between two ICCs was larger than a determined cut-off score, the item was seen as demonstrating item parameter drift. The criteria of cut-off score were determined by the simulated data and differ across each paired years. The criterion of the item difficulty difference between two years was the product of 1.96 and SD_d (the standard deviation of the difference between bank value and adjust value). The values of the ICCs (the probabilities) were then constructed using the item difficulty as 0.00 and the item difficulty as the criterion of the item difficulty difference. Then taking the absolute value of the differences between the probabilities and multiplying the

difference by the interval width and summing, the criterion of cut-off score was generated. Finally, once unstable items were identified, the patterns in terms of item characteristics that caused item parameter drift were described. Chapter 4 presented the results of the analyses.

CHAPTER IV: Results

This chapter presents the results of the analyses of equating items. First, item coding results are summarized. The characteristics of items are presented including changes from one year to the next. Next, items with IPD are identified. Finally, patterns regarding the features of unstable items are described.

Item Coding Results

Sixty items were identified as equating items from 1999 to 2004. The characteristics identified include item strand, item format, cognitive complexity, and item concreteness are summarized in Table 7. As for item strand, these 60 common items contained a variety of items so that all strands or EALRs could be addressed during equating. Each of the operational test forms contained a larger number of items in *Content Strand* than in *Process Strand*. The total equating items used in this study followed the approximate proportions of items for each strand in the operational forms. Most of the common items were related to *Content Strand*; the most frequent content strand items were *Algebraic Sense* (n=12, 20%), *Measurement* (n=11, 18.3%) and *Probability and Statistics* (n=11, 18.3%). Eleven items were targeted in *Process Strand*: 5 items were related to *Problems/Reasons Logically* (8.3%), 5 items were *Makes Connections* (8.3%) and 1 item was *Communicates Understanding* (1.7%). In terms of *Item Format*, forty-seven (78.3%) items were multiple-choice items and thirteen (21.7%) items were

constructed-response items (i.e., short-answer items). The proportions of multiple-choice and constructed-response used as common items followed the design of each operational form. As for *Cognitive Complexity*, 27 items (45%) were *Routine* items, 30 items (50%) were *Comprehension* items and 3 items (5%) were *Non-routine/Insightful* items. Apparently, *Non-routine/Insightful* items were not often selected as equating items. Regarding to *Item Concreteness*, 17 items (28.3%) were *Abstract* items, 21 items (35%) were *Contextual* items, and 22 items (36.7%) were *Concrete/Story* items.

Table 7

Items by Strands, Format, Cognitive complexity, and Text/Story

	N	Percentage (%)
Item Strands		
Number sense (NS)	6	10.0
Measurement (ME)	11	18.3
Algebraic sense (AS)	12	20.0
Geometric sense (GS)	9	15.0
Probability and statistics (PS)	11	18.3
Solve problems/Reason logically (SR)	5	8.3
Communicate understanding (CU)	1	1.7
Makes Connections (MC)	5	8.3
Item Format		
Multiple – choice (MC)	47	78.3
Short– answer (SA)	13	21.7
Cognitive Complexity		
Routine	27	45.0
Comprehension	30	50.0
Nonroutine/insightful	3	5.0
Item Concreteness		
Abstract	17	28.3
Contextual	21	35.0
Concrete/Story	22	36.7

Considering the structure of items, specifically the relative position of stimulus and stem located in the items, 24 (40.0%) of the items' stimuli were in the middle of the stem, 14 (23.3%) of the items' stimuli were after the stem, and only one item's stimulus was before the stem. Twenty-one items (35.0%) didn't have any stimuli. For the 39 items with stimuli, the majority coded as *Graph or Chart* (n=12, 20%), and 13.3% (n=8) were coded as *Geometric Figure*. Generally, when item stimuli were classified, item content determined the type of visual display in the question. For example, *Geometric Sense* items were also coded as *Geometric Figure* or *Grid*; sometimes, *Measurement* items were coded as *Geometric Figure* as well; *Probability and Statistics* items were often coded as *Graph or Chart*; the functions in *Algebraic Sense* items sometimes were coded as *Graph or Chart*. Items under *Process Strand* inclined to use *Picture of Phenomena* and *Object Cue* as stimuli.

Table 8

Items by Structure, Stem and Stimuli

	N	Percentage (%)
Structure		
Stimulus in the middle of stem	24	40.0
Stimulus after stem	14	23.3
Stimulus before stem	1	1.7
No stimulus	21	35.0
Stimulus [a]		
Expression	5	8.3
Graph or Chart	12	20.0
Geometric Figure	8	13.3
Grid	4	6.7
Picture of Phenomena	4	6.7
Object Cue	3	5.0
Text	3	5.0

Note. [a] The total number of items with stimulus was 39.

As for item location changes, the majority of equating items (n=46, 76.7%) were located in the same item locations, 7 items (11.7%) were moved backward 1 position, 5 items (8.3%) were moved backward 2 positions; two items were moved forward 1 position and 2 positions, respectively. In terms of the change in layout of an item, 19 items (31.7%) maintained the layout as full page and 24 items (40.0%) as half page, however, 8 items (13.3%) were expanded from half page to full page and 9 items (15.0%) were shrunk from full page to half page. Considering the common items and their adjacent items, 11 items (18.3%) were moved from the left page to the right page, 12 items (20.0%) were moved from the right page to the left page, 17 items (28.3%) stayed on the right page, and 20 items (33.3%) stayed on the left page. The distributions of changes in item location and length of item were summarized in Table 9.

Table 9

Items by Changes in Location, Numbers, and Position

	N	Percentage (%)
Item Locations		
move forward 2 positions (-2)	1	1.7
move forward 1 position (-1)	1	1.7
move backward 1 position (1)	7	11.7
move backward 2 positions (2)	5	8.3
stay on the same location (0)	46	76.7
Numbers of Page		
half page to full page	8	13.3
full page to half page	9	15.0
full page (no change)	19	31.7
half page (no change)	24	40.0
Item Positions		
move from left to right (LR)	11	18.3
move from right to left (RL)	12	20.0
stay on right pages (RR)	17	28.3
stay on left pages (LL)	20	33.3

For those 24 items that kept their layout as half page, 4 of them moved from the bottom to the top of the page, 5 items moved from top to bottom, 6 stayed at the bottom, and 9 stayed at the top. Considering the item type of their adjacent items, the majority of items (n=17) were associated with items of the same type over years, i.e., the multiple-choice item were paired with another multiple-choice item on the same page the next year. However, 2 of 24 items were paired with another item with same format in the first year but paired with an item with a different format in the second year; 4 items with a different format in the first year and an item with the same format in the second year. Only one item was paired with another item with different format in both years. The distributions of changes in item position and the item format of common items' adjacent item were summarized in Table 10.

Table 10

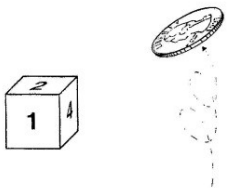
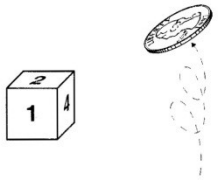
Half-Page Items by Changes in Position and Associated Item (N=24)

	N	Percentage (%)
Item Positions		
move from bottom to top (BT)	4	16.7
move from top to bottom (TB)	5	20.8
stay on the bottom pages (BB)	6	25.0
stay on the top pages (TT)	9	37.5
Accompaniment of item format		
same to different format over years	2	8.3
different to same format over years	4	16.7
both years with different formats	1	4.2
both years with same format	17	70.8

Although equating items were supposed to be the same over years, some variables were not controlled in different forms. There were even some slight changes in item content. In some items, names of persons were different from one year to the next. For example, for the "Pizza check split by four" item, "Alice, Bob, Carol, and Dan" named in 1999 (20) became "Alice, Bob, Farhana, and Jamal" in 2000 (20). For "Possible outcomes of cube and coin" item, "Joey and Rochelle" in 2003 (9) were changed into "Joseph and Cindy" in 2004 (9). These two items are shown in Table 11.

Table 11

Changes in Names of Persons

First Year	Second Year
"Pizza check split by four" in 1999(20) and 2000(20):	
20 **Alice, Bob, Carol, and Dan went out for pizza. When the bill came, they decided to split the check. Alice figured out what she owed by multiplying the bill by 0.25. Bob figured his share by finding 30% of the total. Carol figured out her amount by dividing the total by 3. To determine what he owed, Dan found 12% of the total. Who paid the most money?** ○ **A.** **Alice** ○ **B.** **Bob** ○ **C.** **Carol** ○ **D.** **Dan**	**20** Alice, Bob, Farhana, and Jamal went out for pizza. When the bill came, they decided to split the check. Alice figured out what she owed by multiplying the bill by 0.25. Bob figured his share by finding 30% of the total. Farhana figured out her amount by dividing the total by 3. To determine what he owed, Jamal found 12% of the total. Who paid the most money? ○ **A.** Alice ○ **B.** Bob ○ **C.** Farhana ○ **D.** Jamal
"Possible outcomes of cube and coin" in 2003(9) and 2004(9):	
9 Joey and Rochelle made up a game in their mathematics class. To earn points in the game, each player rolls a six-sided cube with numbers 1 through 6 on the sides and then flips a coin. If the coin lands "tails up," the player gets a total number of points equal to the number at the top of the cube. If the coin lands "heads up," the player's points are doubled for that turn.  In the box below, list all the possible outcomes for each turn. Then indicate the probability of a player getting 6 points in one turn.	**9** Joseph and Cindy made up a game in their mathematics class. To earn points in the game, each player rolls a six-sided cube with numbers 1 through 6 on the sides and then flips a coin. If the coin lands "tails up," the player gets a total number of points equal to the number at the top of the cube. If the coin lands "heads up," the player's points are doubled for that turn.  In the box below, list all the possible outcomes for each turn. Then indicate the probability of a player getting 6 points in one turn.

Furthermore, some overlooked changes happened in the context. The term “flipped” was used in the item entitled “Position of new coordinates of point” in 2000 (22) to explain “reflected;” however, the term was omitted in 2001 (20). For the item entitled “Numbers in increasing pattern,” used in 2000(33) and 2001(35), the phrase “Show your work” followed the question sentence and came before the stimulus in 2000 (33); however, it was presented after the stimulus and located before the answer space in 2001 (35). When the item entitled “Line graph of two temperatures scales” was used in 2001 (9), “title, scale labels, axis labels” were listed vertically, however, these terms were listed horizontally in 2002 (9). Further, the x-axis and y-axis were not labeled on the coordinate plane in the first year. In the item entitled “Width of strip on perpendicular sides,” which was used in 2001 (34) and 2002 (34), respectively, the phrase that asked examinees to describe their work was changed from “Show your work” to “Clearly show your work” between 2001 and 2002. For the item entitled “Total salary based on TV sold,” which was used in 2003 (33) and 2004 (33), the font size of the graph title was enlarged in the second year. The items described above are shown in Table 12.

Table 12

Changes in Random Details

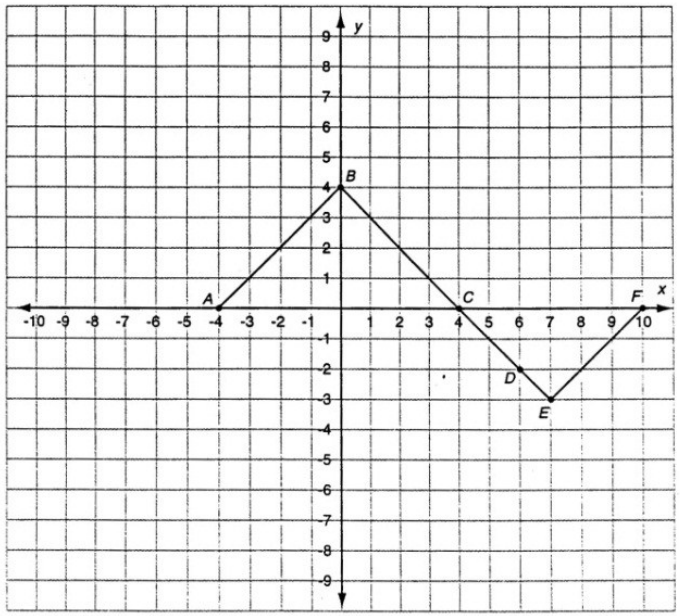
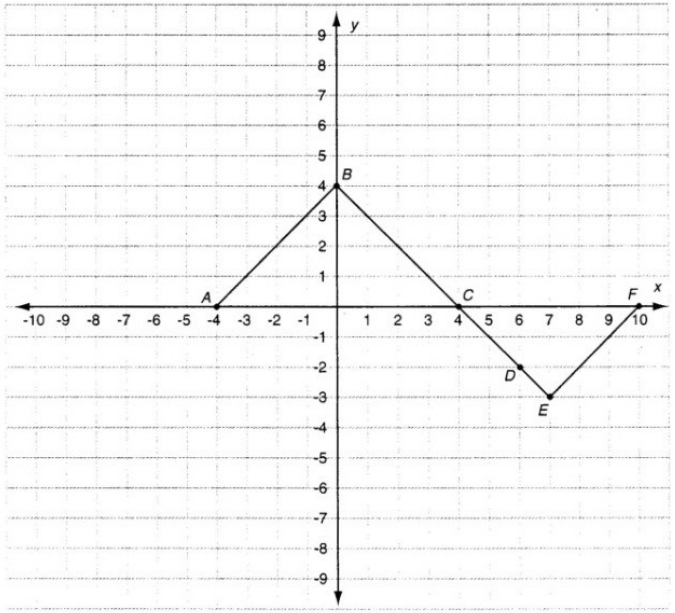
First Year	Second Year
"Position of new coordinates of point" in 2000(22) and 2001(20):	
22 If the figure below were reflected (flipped) over the *y*-axis, what would be the new coordinates of point *D*?  ○ **A.** (6, 2) ○ **B.** (−6, −2) ○ **C.** (2, 6) ○ **D.** (−2, 6)	**20** If the figure below were reflected over the *y*-axis, what would be the new coordinates of point *D*?  ○ **A.** (6, 2) ○ **B.** (−6, −2) ○ **C.** (2, 6) ○ **D.** (−2, 6)

Table 12

Changes in Random Details (Continued)

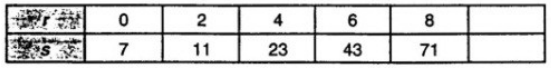
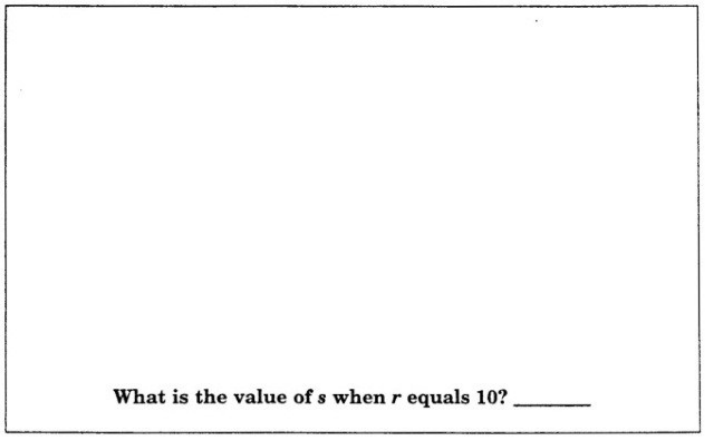
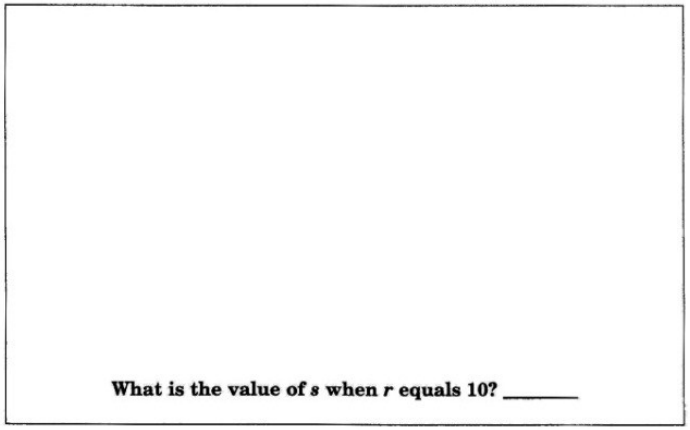
First Year	Second Year
"Numbers in increasing pattern" in 2000(33) and 2001(35):	
33 Study the pattern shown in the following table. What would be the value of s when r equals 10? Show your work.  	r	0	2	4	6	8	
s	7	11	23	43	71	|  **What is the value of s when r equals 10? _____**	**35** Study the pattern shown in the following table. What would be the value of s when r equals 10? | r	0	2	4	6	8	
s	7	11	23	43	71		 Show your work.  **What is the value of s when r equals 10? _____**

Table 12

Changes in Random Details (Continued)

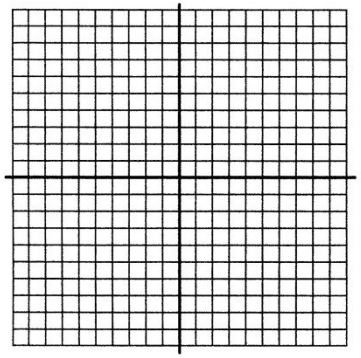
First Year	Second Year
"Line graph of two temperatures scales" in 2001 (9) and 2002 (9):	

9 The following table shows the relationship between some Fahrenheit temperatures and their Celsius equivalents.

° Fahrenheit	° Celsius
-13	- 25
- 4	- 20
5	- 15
23	- 5
32	0
50	10
68	20

Use this information to draw a line graph that shows Fahrenheit temperature on one axis and Celsius temperature on the other. Be sure your graph has the following:

- title
- scale labels
- axis labels



9 The following table shows the relationship between some Fahrenheit temperatures and their Celsius equivalents.

° Fahrenheit	° Celsius
–13	–25
–4	–20
5	–15
23	–5
32	0
50	10
68	20

Use this information to draw a line graph that shows Fahrenheit temperature on one axis and Celsius temperature on the other. Be sure your graph has the following:

• title • scale labels • axis labels

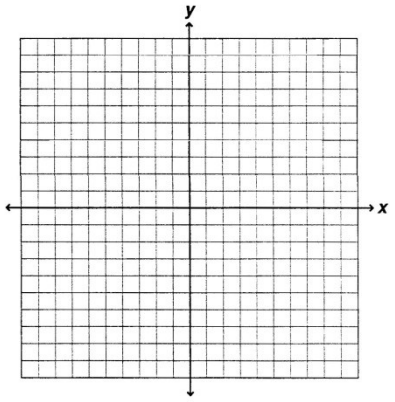


Table 12

Changes in Random Details (Continued)

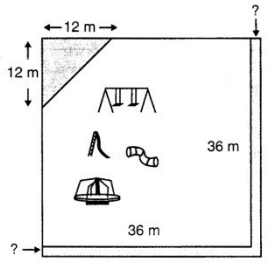
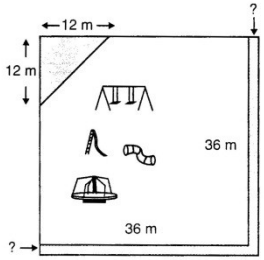
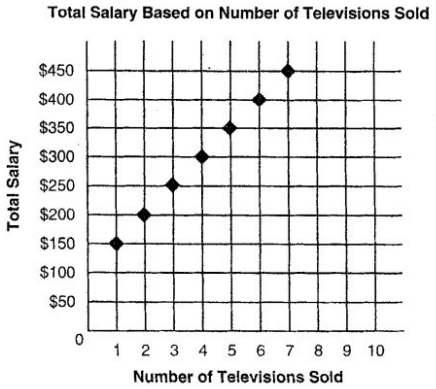
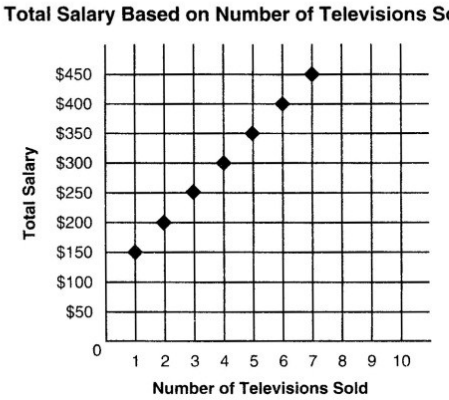
First Year	Second Year
"Width of strip on perpendicular sides" in 2001 (34) and 2002 (34):	
34 The school playground, a 36 m square, is going to lose part of one corner due to construction. The city will take a triangular section that measures 12 m and 12 m on the perpendicular sides.  The city has agreed to allow the school to expand the playground on the remaining two sides to replace the lost area. Find the approximate width of the strip that will be added to each of the two sides. Show your work. **Approximate width of the strip is _____**	**35** The school playground, a 36 m square, is going to lose part of one corner due to construction. The city will take a triangular section that measures 12 m and 12 m on the perpendicular sides.  The city has agreed to allow the school to expand the playground on the remaining two sides to replace the lost area. Find the approximate width of the strip that will be added to each of the two sides. Clearly show your work. **Approximate width of the strip is _____**

Table 12

Changes in random details (Continued)

First Year	Second Year
"Total salary based on TV sold" in 2003 (33) and 2004 (33):	
33 The chart below shows the amount of total salary (commission plus base salary) paid to employees of a store that specializes in big screen televisions.  Which equation best represents the total salary (T) that an employee makes for selling any (n) number of television sets? ○ **A.** $T = 50n + 100$ ○ **B.** $T = 100(n + 50)$ ○ **C.** $T = 100n + 50$ ○ **D.** $T = 50(n + 100)$	**33** The chart below shows the amount of total salary (commission plus base salary) paid to employees of a store that specializes in big screen televisions.  Which equation best represents the total salary (T) that an employee makes for selling any (n) number of television sets? ○ **A.** $T = 50n + 100$ ○ **B.** $T = 100(n + 50)$ ○ **C.** $T = 100n + 50$ ○ **D.** $T = 50(n + 100)$

Some items used words in bold to emphasize important information. However, three items emphasized keywords in the second year, but not in the first year, i.e., “at least” became bold-faced for the item “Sum of two rolled number cubes”, “percentage” became bold-faced for the item “Bar graph representing percentage”, and “least to greatest” became bold-face for the item “Boxes ordered least to greatest volume”. These items are shown in Table 13.

Items using keywords in bold for both years are summarized in Table 14. The terms included: “rectangular”, “all”, “greatest to least”, “median”, “remaining”, and “Explain in detail”. Expressions and equations were written in bold as well. In the “Winning percentage for the season” item, the term “more” in question “How many more wins are needed?” was underlined both years.

Table 13

Items for which Keywords were Emphasized in the Second Year

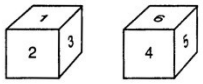
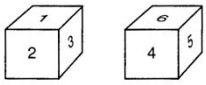
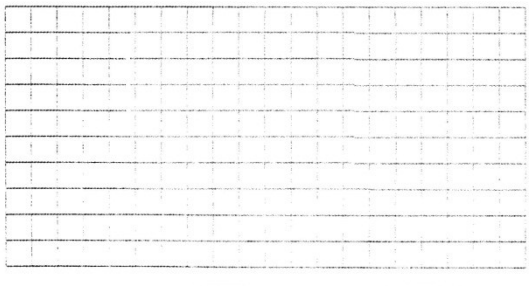
First Year	Second Year
"Sum of two rolled number cubes" in 2000 (29) and 2001 (29):	
29 Akio and Tamera are playing a game with two number cubes, each labeled 1–6. On each turn, the person rolls both number cubes. The sum of the two numbers on top of the number cubes tells how many spaces that person can move the game piece.  Akio needs to move nine spaces to win the game. What is the probability that Akio will roll a sum of at least nine on his next turn? ○ **A.** $\frac{1}{10}$ ○ **B.** $\frac{1}{4}$ ○ **C.** $\frac{5}{18}$ ○ **D.** $\frac{4}{11}$	**29** Akio and Tamera are playing a game with two number cubes, each labeled 1–6. On each turn, the person rolls both number cubes. The sum of the two numbers on top of the number cubes tells how many spaces that person can move the game piece.  Akio needs to move nine spaces to win the game. What is the probability that Akio will roll a sum of **at least** nine on his next turn? ○ **A.** $\frac{1}{10}$ ○ **B.** $\frac{1}{4}$ ○ **C.** $\frac{5}{18}$ ○ **D.** $\frac{4}{11}$
"Bar graph representing percentage" in 2001 (4) and 2002 (4):	
4 Mrs. Andrews is supervising an independent study course. Each of the students is required to complete 20 assignments. The list below shows how many assignments each student has completed. **Student** / **Number of Assignments Completed** Mike Cooper 10 Manuel Flores 15 Latasha Williams 11 Sondra Rao 10 Tam Chan 14 Use the grid to create a bar graph that shows the percentage of assignments completed by each student. Clearly label the scale and axes. **Percentage of Assignments Completed** 	**4** Mrs. Andrews is supervising an independent study course. Each of the students is required to complete 20 assignments. The list below shows how many assignments each student has completed. **Student** / **Number of Assignments Completed** Mike Cooper 10 Manuel Flores 15 Latasha Williams 11 Sondra Rao 10 Tam Chan 14 Use the grid to create a bar graph that shows the **percentage** of assignments completed by each student. Clearly label the scale and axes.

Table 13

Items for which Keywords were Emphasized in the Second Year (Continued)

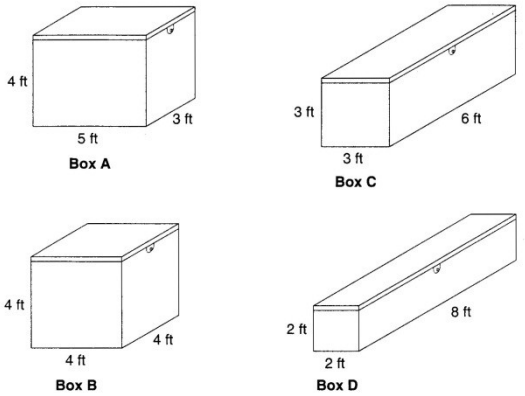
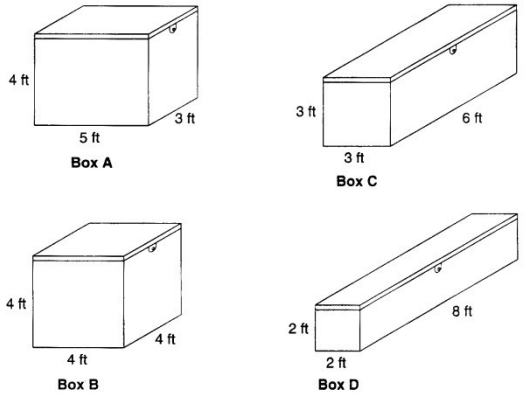
First Year	Second Year
"Boxes ordered least to greatest volume" in 2001 (24) and 2002 (25):	
24 Terry is designing a flyer to advertise storage boxes that he sells.  He wants to show the boxes from least to greatest volume. What is the correct order? ○ **A.** BACD ○ **B.** ABCD ○ **C.** DCBA ○ **D.** DCAB	**25** Terry is designing a flyer to advertise storage boxes that he sells.  He wants to show the boxes from **least to greatest** volume. What is the correct order? ○ **A.** BACD ○ **B.** ABCD ○ **C.** DCBA ○ **D.** DCAB

Table 14

Items for which Keywords were Bold-Faced or Underlined Both Years

Item	Keyword
"The greatest rectangular garden area" in 1999(4) and 2000(4) :	
4 A gardener has twenty feet of fencing to surround a new **rectangular** garden. Which of the following is the greatest area that can be enclosed? ○ **A.** 20 square feet ○ **B.** 24 square feet ○ **C.** 25 square feet ○ **D.** 28 square feet	rectangular
"Equation representing apartment rent" in 1999(21) and 2000(21):	
21 Lamont and Pete are sharing a two-bedroom apartment. Pete has offered to pay \$50 more in rent per month than Lamont, if Pete can have the larger bedroom. The total rent for the apartment is \$725. Write an equation to figure out how much rent Lamont will have to pay. Then solve the equation. **Let *r* = Lamont's rent.** **Lamont will have to pay \$_____.**	Let r = Lamont's rent
"Expression be divisible by 6" in 1999 (31) and 2000 (31):	
31 Which of these expressions is divisible by 6 for **all** whole number values of a and b? ○ **A.** $3(2a + 3)$ ○ **B.** $3b + 3$ ○ **C.** $3a(4b)$ ○ **D.** $3b(9a)$	all

Table 14

Items for which Keywords were Bold-Faced or Underlined Both Years (Continued)

Item	Keyword
"Properties of a and b" in 2001 (1) and 2002 (2):	
1 Audrey is given the following problem to solve. $\begin{array}{r} a\ b\ a \\ +\ a\ b \\ \hline a\ 7\ 7 \end{array}$ Audrey has to solve for a and b. Which of the following is **not** possible? ○ **A.** b is odd and greater than a. ○ **B.** a is even and smaller than 5. ○ **C.** a and b are both odd numbers. ○ **D.** a and b are both prime numbers.	$\begin{array}{r} a\ b\ a \\ +\ a\ b \\ \hline a\ 7\ 7 \end{array}$
"Factor of an expression" in 2001 (7) and 2002 (7):	$3a^2 + 12a$
7 Which term is a factor of $3a^2 + 12a$? ○ **A.** $3a$ ○ **B.** $4a$ ○ **C.** $3a^2$ ○ **D.** $4a^2$	
"Equation of car rental cost" in 2001 (12) and 2002 (12):	Let C=cost, M=miles, and D=days
12 Kesha is planning to rent a van for her trip to Mt. Rainier. Two of her friends each rented the same type of van from the same car rental company last week. This is what they told her: John: "The cost of my rental was $240. The company charged me a certain amount per day and a certain amount per mile. I had the rental for five days and I drove it 200 miles." Katie: "The cost of my rental was only $100. I drove it for 100 miles and had it for two days." Kesha plans to get the same type of van that John and Katie had from the same car rental company. Kesha estimated her trip would be 250 miles, and she would have the vehicle for four days. Which of the following equations could Kesha use to figure out how much her rental would cost? **Let C = cost, M = miles, and D = days** ○ **A.** $C = 40.00M + 0.20D$ ○ **B.** $C = 40.00D + 0.20M$ ○ **C.** $C = 20.00M + 0.40D$ ○ **D.** $C = 20.00D + 0.40M$	

Table 14

Items for which Keywords were Bold-Faced or Underlined Both Years (Continued)

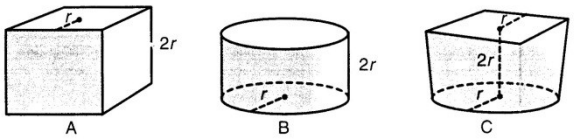
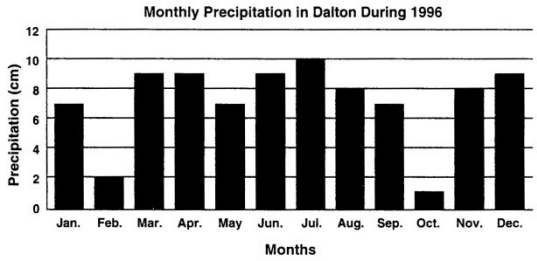
Item	Keyword
"Containers ordered greatest to least volume" in 2002 (10) and 2003 (10):	
10 Arrange the three containers in order from **greatest to least** volume. (Drawings are not to scale.) • Container A is a cube with edge $2r$. • Container B is a cylinder with radius r and height $2r$. • Container C has a circular base with radius r, a square top with side $2r$, and height $2r$.  ○ **A.** ABC ○ **B.** BCA ○ **C.** ACB ○ **D.** CBA	greatest to least
"Mediam amount of monthly precipitation" in 2002 (32) and 2003 (32):	median
32 The graph shows the approximate amount of precipitation that fell in Dalton each month during 1996.  What was the **median** amount of monthly precipitation in Dalton during 1996? ○ **A.** 7 cm ○ **B.** 8 cm ○ **C.** 9 cm ○ **D.** 10 cm	

Table 14

Items for which Keywords were Bold-Faced or Underlined Both Years (Continued)

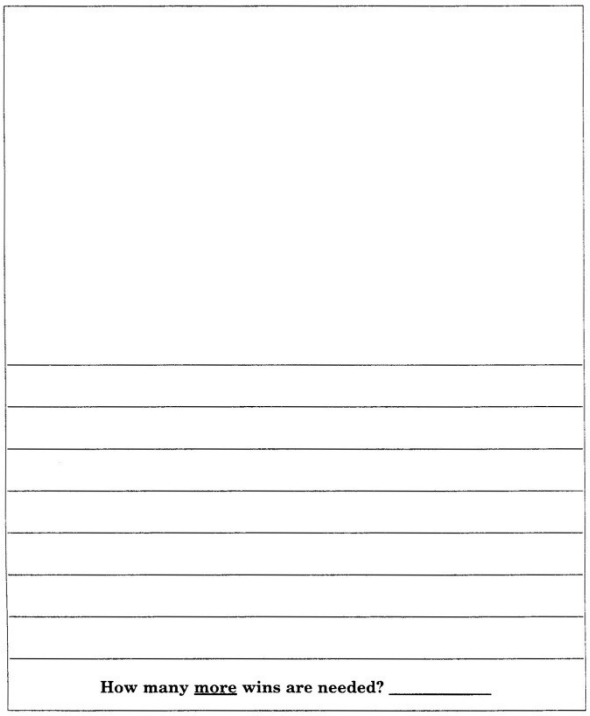
Item	Keyword
"Winning percentage for the season" in 2002 (44) and 2003 (44):	
44 A team has won 10 of the 15 games it has played. The team has 25 games left to play. The players figure they will make the playoffs if their winning percentage for the season is 60%. How many of the **remaining** games must the team win to have a winning percentage of 60% for the season? Clearly show how you arrived at your answer.  **How many <u>more</u> wins are needed? _____**	remaining; more

Table 14

Items for which Keywords were Bold-Faced or Underlined Both Years (Continued)

Item	Keyword
"Comparison of transportation costs" in 2003 (30) and 2004 (29):	
30 Earl is planning to travel from Seattle to Oklahoma City. His destination is 1,970 miles one-way. If he flies, he can get a one-way ticket for $400. If he drives, it will take him 3 days to get there, and the cost of renting a car would be $29 per day plus $0.19 per mile. Considering his transportation costs alone, would it cost more if he flew or drove? **Explain in detail** your answer using words, numbers, and/or diagrams. **Which way would be more expensive? _____**	Explain in detail

The next section reports the results of Robust Z-statistics and the signed area measure. In Table 15 and Table 16, the descriptive statistics of year to year equating as well as the Z-statistics and Robust Z-statistics are summarized. If the absolute value of the Z-score and Robust Z-score for an item was larger than 1.96, the item was flagged as drifting between test years. Seven items were flagged as drifting. The results of Robust Z-statistics flagged 4 more items.

Table 15

Difficulty parameters

Year Pairs	Item numbers	Mean		Linking Constant[a]	$Median_d$[b]	SD_d[c]	IQR_d[d]
		Bank Value	Initial Value Parameter				
1999-2000	15	0.057	-0.191	0.248	-0.008	0.099	0.095
2000-2001	12	0.352	0.033	0.318	0.037	0.127	0.150
2001-2002	12	0.136	0.226	-0.091	-0.002	0.125	0.120
2002-2003	12	-0.358	-0.167	-0.192	-0.007	0.154	0.059
2003-2004	14	-0.632	-0.098	-0.534	-0.006	0.064	0.087

Note.[a] Linking constant is the differences between average bank value and average initial parameter. [b]$Median_d$ is the median of the difference between bank value and adjust value. [c]SD_d is the standard deviation of the difference between bank value and adjust value. [d]IQR_d is interquartile range of the difference between bank value and adjust value.

Table 16

Descriptive Statistics and Z Statistics

	ID	Type	Bank Value	Initial Parameter	Adjusted Parameter	Difference	Z Statistic		Robust-Z Statistic	
	1	mc	-0.500	-0.660	-0.412	-0.088	-0.889		-1.138	
	2	mc	1.020	0.770	1.018	0.002	0.020		0.142	
	3	mc	-0.940	-1.160	-0.912	-0.028	-0.283		-0.284	
	4	mc	-0.370	-0.690	-0.442	0.072	0.728		1.138	
	5	mc	-0.080	-0.300	-0.052	-0.028	-0.283		-0.284	
	6	cr	0.150	-0.030	0.218	-0.068	-0.687		-0.853	
	7	mc	-0.540	-0.800	-0.552	0.012	0.121		0.284	
1999-2000	8	cr	0.540	0.510	0.758	-0.218	-2.203	**	-2.987	***
	9	mc	-0.040	-0.350	-0.102	0.062	0.627		0.996	
	10	cr	0.060	-0.160	0.088	-0.028	-0.283		-0.284	
	11	mc	-0.270	-0.600	-0.352	0.082	0.829		1.280	
	12	mc	0.210	-0.270	-0.022	0.232	2.345	**	3.414	***
	13	mc	0.430	0.230	0.478	-0.048	-0.485		-0.569	
	14	mc	0.910	0.610	0.858	0.052	0.526		0.853	
	15	mc	0.270	0.030	0.278	-0.008	-0.081		0.000	

*p<.10, **p<.05, ***p<.01

Table 16

Descriptive Statistics and Z Statistics (Continued)

	ID	Type	Bank Value	Initial Parameter	Adjusted Parameter	Difference	Z Statistic		Robust-Z Statistic	
2000-2001	16	mc	-0.180	-0.540	-0.222	0.042	0.328		0.045	
	17	mc	0.630	0.280	0.598	0.032	0.249		-0.045	
	18	mc	0.340	-0.050	0.268	0.072	0.564		0.315	
	19	mc	-1.040	-1.280	-0.962	-0.078	-0.617		-1.036	
	20	mc	0.060	0.000	0.318	-0.258	-2.034	**	-2.658	***
	21	mc	-0.260	-0.660	-0.342	0.082	0.643		0.405	
	22	mc	1.180	0.630	0.948	0.232	1.824	*	1.757	*
	23	cr	0.200	-0.100	0.218	-0.018	-0.144		-0.495	
	24	mc	-0.340	-0.730	-0.412	0.072	0.564		0.315	
	25	mc	1.150	0.980	1.298	-0.148	-1.168		-1.667	*
	26	mc	1.030	0.660	0.978	0.052	0.407		0.135	
	27	cr	1.450	1.210	1.528	-0.078	-0.617		-1.036	

*p<.10, **p<.05, ***p<.01

Table 16

Descriptive Statistics and Z Statistics (Continued)

	ID	Type	Bank Value	Initial Parameter	Adjusted Parameter	Difference	Z Statistic		Robust-Z Statistic	
2001-2002	28	mc	-0.283	-0.302	-0.393	0.110	0.879		1.253	
	29	mc	0.282	0.431	0.340	-0.058	-0.468		-0.635	
	30	cr	-0.450	-0.363	-0.454	0.004	0.029		0.062	
	31	mc	0.019	0.226	0.135	-0.116	-0.933		-1.287	
	32	cr	0.761	0.603	0.512	0.249	1.994	**	2.815	***
	33	mc	-0.043	0.140	0.049	-0.092	-0.740		-1.017	
	34	mc	0.219	0.331	0.240	-0.021	-0.171		-0.219	
	35	mc	-0.478	-0.412	-0.503	0.025	0.198		0.298	
	36	mc	1.518	1.496	1.405	0.113	0.903		1.287	
	37	mc	0.264	0.593	0.502	-0.238	-1.911	*	-2.658	***
	38	mc	-1.217	-1.119	-1.210	-0.007	-0.059		-0.062	
	39	cr	1.034	1.090	0.999	0.035	0.278		0.410	

*p<.10, **p<.05, ***p<.01

Table 16

Descriptive Statistics and Z Statistics (Continued)

	ID	Type	Bank Value	Initial Parameter	Adjusted Parameter	Difference	Z Statistic		Robust-Z Statistic	
2002-2003	40	mc	-1.024	-0.848	-1.040	0.016	0.103		0.513	
	41	mc	-0.631	-0.195	-0.387	-0.244	-1.581		-5.417	***
	42	mc	0.980	1.178	0.986	-0.006	-0.039		0.011	
	43	mc	-1.709	-1.905	-2.097	0.388	2.513	**	8.998	***
	44	cr	0.140	0.330	0.138	0.002	0.012		0.194	
	45	mc	-0.392	-0.193	-0.385	-0.007	-0.046		-0.011	
	46	cr	-0.720	-0.514	-0.706	-0.014	-0.091		-0.171	
	47	mc	0.612	0.663	0.471	0.141	0.913		3.364	***
	48	mc	0.045	0.284	0.092	-0.047	-0.305		-0.924	
	49	mc	-0.947	-0.681	-0.873	-0.074	-0.480		-1.540	
	50	cr	1.130	1.477	1.285	-0.155	-1.004		-3.387	***
	51	mc	-1.785	-1.594	-1.786	0.001	0.006		0.171	

*p<.10, **p<.05, ***p<.01

Table 16

Descriptive Statistics and Z Statistics (Continued)

	ID	Type	Bank Value	Initial Parameter	Adjusted Parameter	Difference	Z Statistic	Robust-Z Statistic
	52	mc	-1.242	-0.608	-1.142	-0.100	-1.562	-1.460
	53	mc	-0.462	0.111	-0.423	-0.039	-0.606	-0.513
	54	cr	-0.385	0.102	-0.432	0.047	0.742	0.823
	55	mc	0.090	0.684	0.150	-0.060	-0.935	-0.839
2003-2004	56	mc	-1.507	-0.967	-1.501	-0.006	-0.089	0.000
	57	mc	-1.658	-1.210	-1.744	0.086	1.353	1.429
	58	cr	-0.180	0.311	-0.223	0.043	0.679	0.761
	59	mc	-0.501	-0.033	-0.567	0.066	1.040	1.118
	60	mc	0.157	0.731	0.197	-0.040	-0.622	-0.528

*p<.10, **p<.05, ***p<.01

Table 17 summarizes the results of area measure for detecting item parameter drift. The curves were compared by evaluating the signed area between a pair of common ICCs which was labeled to investigate the difference involving item difficulty, discrimination and guessing parameters. The criteria of cut-off score were determined first. According to the Table 15, in 1999 and 2000, SD_d was 0.099, so the criterion of the item difficulty difference between two years was the product of 1.96 and 0.099 which was 0.194. The values of the probabilities were then simulated using the item difficulty as 0.00 and 0.194. After taking the absolute value of the differences between the probabilities and multiplying the difference by the interval width (the θ increment used in the calculations was 0.01 in this study) and summing, the criterion of cut-off score equaled to 0.191, which was generated. Based on the same logics, in 2000 and 2001, SD_d is 0.127, the item difficulty difference is 0.249, and the determined cut-off score was 0.246. In 2001 and 2002, SD_d is 0.125, the item difficulty difference is 0.245, and the determined cut-off score was 0.242. In 2002 and 2003, SD_d is 0.154, the item difficulty difference is 0.302, and the determined cut-off score was 0.298. In 2003 and 2004, SD_d is 0.064, the item difficulty difference is 0.125, and the determined cut-off score was 0.123. Twelve items were found for which the area between two ICCs was larger than the determined cut-off score, and then these items were seen as flagged items. Figure 1 shows the visual representation of DIF, in which the bank values were drawn as a solid curve and the adjusted values were drawn as a dashed curve.

Table 17

Area between Curves

			First year (bank value)			Second year (adjusted value)			Area	
	ID	Type	b	a	c	b	a	c		
1999-2000[a]	1	mc	-0.50	0.97	0.01	-0.412	1.05	0.00	0.144	
	2	mc	1.02	0.89	0.01	1.018	0.9	0.02	0.062	
	3	mc	-0.94	1.06	0	-0.912	1.13	0	0.053	
	4	mc	-0.37	1.29	0	-0.442	1.36	0	0.074	
	5	mc	-0.08	0.99	0	-0.052	1.02	0	0.034	
	6	cr	0.15	1.17	0	0.218	1.13	0	0.069	
	7	mc	-0.54	1.29	0	-0.552	1.41	0	0.055	
	8	cr	0.54	1.13	0	0.758	1.33	0	0.227	*
	9	mc	-0.04	1.08	0	-0.102	1.07	0	0.062	
	10	cr	0.06	0.92	0	0.088	0.99	0	0.067	
	11	mc	-0.27	0.57	0.04	-0.352	0.79	0.03	0.408	*
	12	mc	0.21	0.7	0.03	-0.022	0.8	0.02	0.263	*
	13	mc	0.43	0.63	0.04	0.478	0.76	0.03	0.256	*
	14	mc	0.91	1.17	0	0.858	1.2	0	0.052	
	15	mc	0.27	1.16	0	0.278	1.06	0	0.067	

Note. [a] Significant area with z > 1.96 and difference >.194; critical value=0.191. *p<.05.

Table 17

Area between Curves (Continued)

			First year (bank value)			Second year (adjusted value)			Area	
	ID	Type	b	a	c	b	a	c		
	16	mc	-0.18	1.45	0	-0.222	1.41	0.00	0.042	
	17	mc	0.63	0.83	0.02	0.598	0.67	0.04	0.318	*
	18	mc	0.34	0.9	0.02	0.268	0.86	0.02	0.075	
	19	mc	-1.04	1.27	0	-0.962	1.27	0	0.078	
	20	mc	0.06	0.85	0.02	0.318	0.86	0.02	0.253	*
2000-2001[b]	21	mc	-0.26	1.41	0	-0.342	1.38	0	0.082	
	22	mc	1.18	0.77	0.03	0.948	0.74	0.03	0.224	
	23	cr	0.2	0.91	0	0.218	0.88	0	0.034	
	24	mc	-0.34	1.54	0	-0.412	1.49	0	0.072	
	25	mc	1.15	1.09	0	1.298	1.03	0.01	0.175	
	26	mc	1.03	0.81	0.03	0.978	0.74	0.03	0.099	
	27	cr	1.45	1.03	0	1.528	1	0	0.078	

Note. [b] Significant area with $z > 1.96$ and difference >.249; critical value=0.246. *p<.05.

Table 17

Area between Curves (Continued)

			First year (bank value)			Second year (adjusted value)			Area	
	ID	Type	b	a	c	b	a	c		
	28	mc	-0.283	0.88	0.02	-0.407	1.04	0.00	0.236	
	29	mc	0.282	0.47	0.05	0.326	0.70	0.04	0.546	*
	30	cr	-0.45	1	0	-0.468	1.14	0	0.101	
	31	mc	0.019	1.19	0	0.121	1.29	0	0.107	
	32	cr	0.761	1	0	0.498	0.98	0	0.263	*
2001-2002[c]	33	mc	-0.043	0.89	0.03	0.035	0.94	0.03	0.083	
	34	mc	0.219	0.76	0.03	0.226	0.79	0.03	0.039	
	35	mc	-0.478	1.04	0.01	-0.517	1.13	0	0.103	
	36	mc	1.518	0.79	0.02	1.391	0.83	0.02	0.124	
	37	mc	0.264	0.65	0.03	0.488	0.86	0.01	0.446	*
	38	mc	-1.217	1.01	0	-1.224	1.06	0	0.038	
	39	cr	1.034	1.16	0	0.985	1.21	0	0.053	

Note. [c] Significant area with z > 1.96 and difference >.245; critical value=0.242. *p<.05.

Table 17

Area between Curves (Continued)

			First year (bank value)			Second year (adjusted value)			Area	
	ID	Type	b	a	c	b	a	c		
	40	mc	-1.024	1.01	0	-1.040	0.97	0.00	0.036	
	41	mc	-0.631	0.60	0.05	-0.387	0.50	0.06	0.320	*
	42	mc	0.98	1.22	0	0.986	1.16	0	0.035	
	43	mc	-1.709	1.04	0	-2.097	1.02	0	0.386	*
	44	cr	0.143	1.08	0	0.138	1.06	0	0.015	
2002-2003[d]	45	mc	-0.392	1.3	0	-0.385	1.19	0	0.058	
	46	cr	-0.719	1.12	0	-0.706	1.13	0	0.014	
	47	mc	0.612	0.69	0.03	0.471	0.66	0.02	0.116	
	48	mc	0.045	1	0.02	0.092	0.91	0.04	0.164	
	49	mc	0.045	1	0.02	0.092	0.91	0.04	0.122	
	50	cr	1.13	1.09	0	1.285	1.08	0	0.155	
	51	mc	-1.785	1.14	0	-1.786	1.11	0	0.019	

Note. [d] Significant area with $z > 1.96$ and difference >.302; critical value=0.298. *$p<.05$.

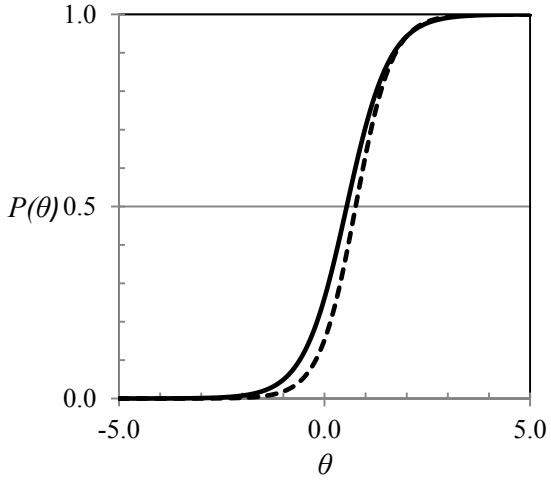
Table 17

Area between Curves (Continued)

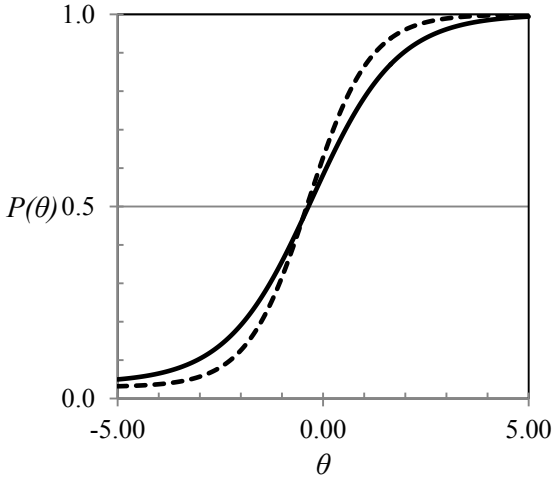
			First year (bank value)			Second year (adjusted value)			Area	
	ID	Type	b	a	c	b	a	c		
2003-2004[e]	52	mc	-1.242	1.22	0	-1.142	1.18	0.00	0.100	
	53	mc	-0.462	1.04	0.00	-0.423	1.05	0.01	0.047	
	54	cr	-0.385	1.12	0	-0.432	1.1	0	0.047	
	55	mc	0.09	1.42	0	0.150	1.37	0	0.060	
	56	mc	-1.507	0.92	0.03	-1.501	0.9	0	0.111	
	57	mc	-1.658	1.01	0	-1.744	1.02	0	0.086	
	58	cr	-0.18	0.93	0	-0.223	0.82	0	0.122	
	59	mc	-0.501	1.31	0	-0.567	1.31	0	0.066	
	60	mc	0.157	0.69	0.04	0.197	0.75	0	0.281	*

Note. [e] Significant area with z > 1.96 and difference >.125; critical value=0.123. *p<.05.

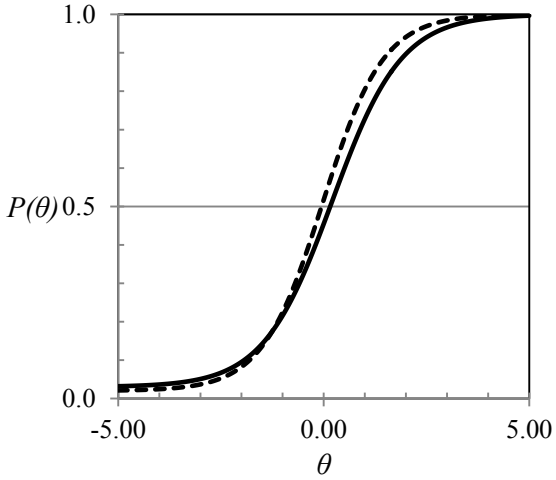
Item 8: “Equation representing apartment rent” in 1999(21) and 2000(21)



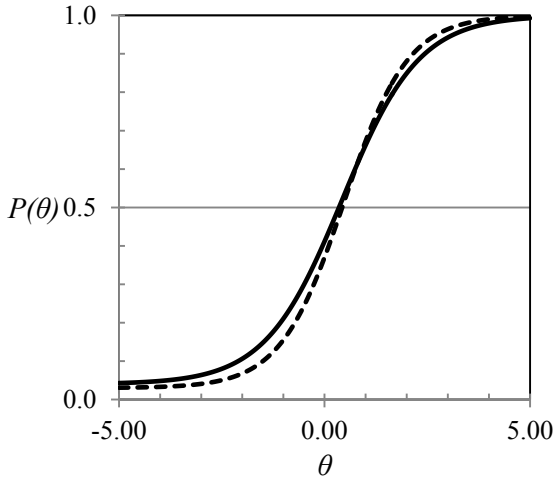
Item 11: “Graph of employee salary” in 1999(27) and 2000(27)



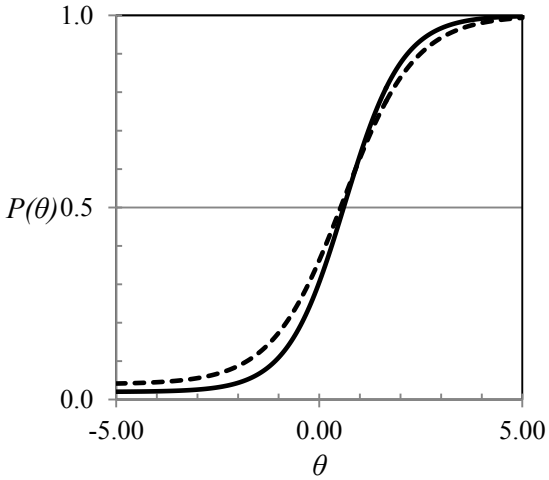
Item 12: “Expression be divisible by 6” in 1999(31) and 2000(31)



Item 13: “Objects balanced on scale” in 1999(32) and 2000(32)



Item 17: “Probability of winning tickets” in 2000(5) and 2001(5)



Item 20: “Position of new coordinates of point” in 2000(22) and 2001(20)

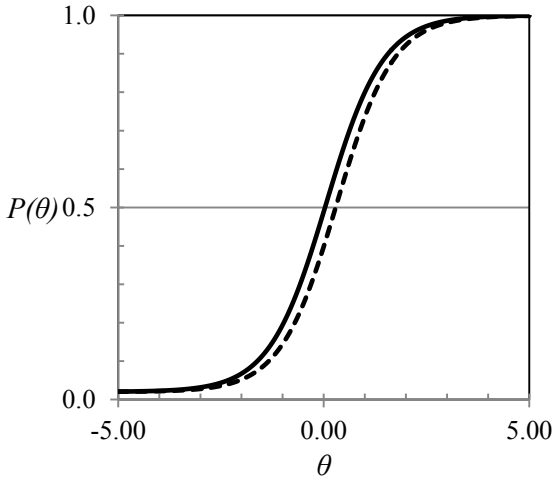
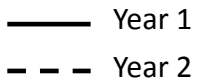
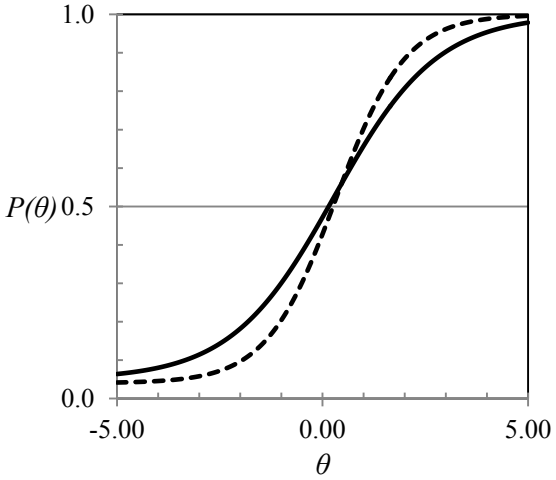


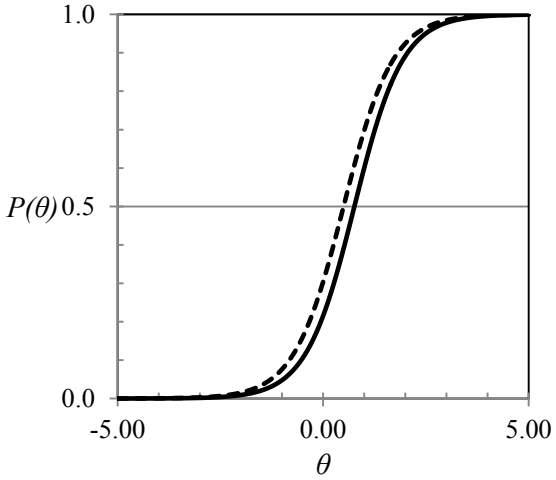
Fig1 Item response functions for selected items among first year and second year



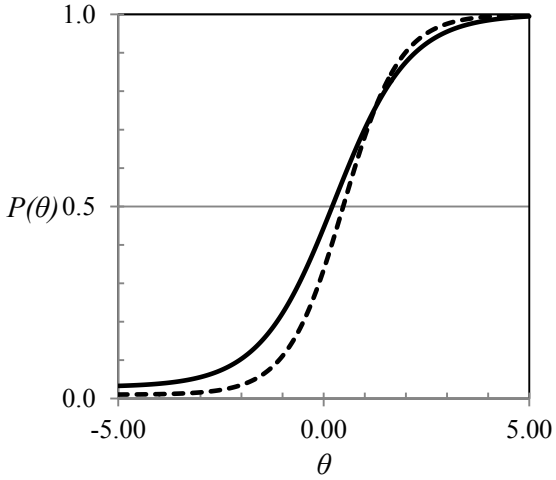
Item 29: “Probability of sum on spinners” in 2001(3) and 2002(3)



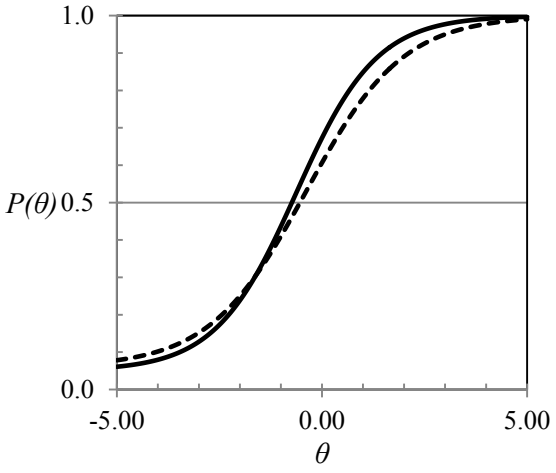
Item 32: “Line graph of two temperatures scales” in 2001(9) and 2002(9)



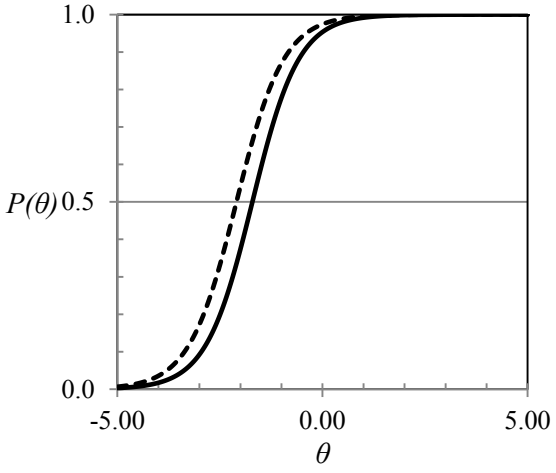
Item 37: “Boxes ordered least to greatest volume” in 2001(24) and 2002(25)



Item 41: “Containers ordered greatest to least volume” in 2002(10) and 2003(10)



Item 43: “Expression representing dollars left” in 2002(13) and 2003(13)



Item 60: “Sand cost needed to fill sandbox” in 2003(43) and 2004(44)

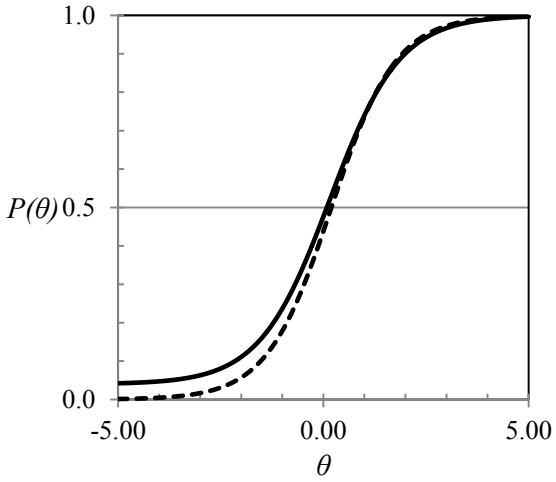


Fig1 Item response functions for selected items among first year and second year (Continuous)

Year 1

Year 2

The following section summarizes the patterns related to IPD with item characteristics given. Flagged items based on Z-Statistics and the area measure, as well as their characteristics are summarized in Table 18.

The results show that the parameters of Process Strand items are less stable than Content Strand items. Among Content Strand items, items belonging to Probability and Statistics were less stable than others. Although there were fewer construct-response items used in WASL and used as equating items, the results did not confirm that constructed-response items were less stable than multiple-choice items.

The parameters of three different levels of Cognitive Complexity did not affect the stability of item parameters. In terms of Item Concreteness, the parameters of Contextual and Concrete/Story items were less stable than the Abstract items. The item structure, i.e., the relationship of stimulus to stem, did not affect the stability of item parameters; however, specific types of stimuli were associated with less stable item parameters. The parameters of the items that were coded with Expression and Text as stimuli did not show drift. However, the parameters of items that were coded with Picture of Phenomena as well as Graph or Chart as stimulus were less stable.

Furthermore, various changes in the items were identified. In order to eliminate item position effects, equating items were administered in the same or almost the same position in

each year's test. The results confirmed that once item was either moved forward two positions or backward two positions, the parameters were less stable. For example, the "Position of new coordinates of point" in 2000 (Item 22) and 2001 (Item 20) was moved forward two positions and the "Measure of central angle of section" item in 2000 (Item 35) and 2001 (Item 37) was moved backward two positions. Changes in numbers of page and item positions affected the stability of item parameters. However, the item format of the adjacent items didn't negatively impact the stability of item parameters.

As expected, making text bold-faced, italic, or underlined to emphasize key information affected the stability of item parameters. For example, the phrase "at least" in the item entitled "Sum of two rolled number cubes" was bold-faced in 2001 (Item 29), and the phrase "least to greatest" was bold-faced in the item entitled "Boxes ordered least to greatest volume" in 2002 (Item 25). Both items were not stable. However, when "percentage" was bold-faced in the item entitled "Bar graph representing percentage" in 2002 (Item 4), the item parameters did not show instability. Considering other items emphasizing keywords in both years, the terms that referenced sorting or that were related to statistics (i.e., "all", "at least", "least to greatest", "greatest to least", "median", and "remaining") had more item parameter drift.

Finally, the parameters of the item entitled "Line graph of two temperatures scales" were not stable. Observing the different item characteristics between two years, "title, scale labels,

axis labels" were listed vertically in the first second, however, they were listed horizontally in the second year. The x-axis and y-axis were added and labeled on the coordinate plane in the second year, too.

Table 18

Flagged items and their item characteristics

Item	Years	Strand	Format	Cognitive Complexity	Item Concreteness	Stimulus	Sig[a]	Sig[b]	Sig[c]
Equation representing apartment rent	1999(21), 2000(21)	AS	SA	Routine	Concrete	N/A	X	X	X
Graph of employee salary	1999(27), 2000(27)	AS	MC	Routine	Contextual	Graph & Chart			X
Expression be divisible by 6	1999(31), 2000(31)	NS	MC	Comprehension	Abstract	N/A	X	X	X
Objects balanced on scale	1999(32), 2000(32)	SR	MC	Nonroutine/ Insightful	Contextual	Picture of Phenomena			X
Probability of winning tickets	2000(5), 2001(5)	PS	MC	Comprehension	Concrete	N/A			X
Position of new coordinates of point	2000(22), 2001(20)	GS	MC	Routine	Abstract	Grid	X	X	X
Sum of two rolled number cubes	2000(29), 2001(29)	PS	MC	Comprehension	Concrete	Object Cue	X	X	
Measure of central angle of section	2000(35), 2001(37)	MC	MC	Comprehension	Contextual	Graph & Chart		X	
Probability of sum on spinners	2001(3), 2002(3)	PS	MC	Comprehension	Contextual	Picture of Phenomena			X
Line graph of two temperatures scales	2001(9), 2002(9)	CU	SA	Comprehension	Contextual	Graph & Chart	X	X	X
Boxes ordered least to greatest volume	2001(24), 2002(25)	ME	MC	Routine	Contextual	Geometric Figure	X	X	X
Containers ordered greatest to least	2002(10), 2003(10)	ME	MC	Comprehension	Abstract	Geometric Figure		X	X
Expression representing dollars left	2002(13), 2003(13)	AS	MC	Routine	Concrete	N/A	X	X	X
Median amount of precipitation	2002(32), 2003(32)	PS	MC	Routine	Contextual	Graph & Chart		X	
Winning percentage for the season	2002(44), 2003(44)	SR	SA	Comprehension	Concrete	N/A		X	
Sand cost needed to fill sandbox	2003(43), 2004(44)	ME	MC	Comprehension	Concrete	N/A			X

Note. [a] Z-Statistics. [b] Robust Z-Statistics. [c] Area measure.

CHAPTER V: Discussion and Conclusion

Standardized tests are increasingly used in large scale educational achievement testing programs. The primary concern of standardized tests is the accuracy of test score interpretation and the appropriateness of test score use across multiple tests. Equating is essential for any testing program that continually produces new parallel forms of a test and for which the expectation is that scores from these forms have the same meaning over time. For some state assessment programs, the assessments across test administrations, are calibrated using IRT, and are equated using common-item nonequivalent groups design with equating constants for year-to-year equating (Kolen & Brennan, 2004).

An important property of IRT is that item parameters calibrated on samples from the same population are invariant, even over different testing administrations. Using common-item nonequivalent groups design for equating tests, the statistical properties of the equating items should be stable across forms. When two groups respond to two alternate forms, the equating items must function similarly in both forms. Nevertheless, if two groups of examinees respond differently to the same item, then that item might not be appropriate to be included in the equating process. However, due to random or systematic errors and item parameter drift (IPD) (Goldstein, 1983), the parameters are not always the same as the ones in the item bank when used in multiple test forms.

In this study, using real data, the threats to item parameter invariance existing at the item level were explored. Specifically, content and context effects on item parameter estimates occurred and affected item parameter instability.

In this paper, the focus was to explore the issues that need to be addressed to better understand how anchor test design can maintain common items' statistical properties across multiple forms and make test equating effective. The data came from the WASL tenth grade mathematics exams administered from 1999 to 2004. Sixty items were identified as equating items across these six years. The present study answered three sets of questions.

The first question sets related to item characteristics itself. Are the parameters of constructed-response items less stable than the parameters of multiple-choice items? Do *Content* and *Process Strands* affect the stability of item parameters? Are the parameters of *Routine* items less stable than the parameters of *Comprehension* and *Nonroutine/Insightful items*? Are the parameters of *Contextual* and *Story/Concrete* items less stable than the parameters of *Abstract* items?

Although Taylor and Lee (2010) found that the parameters of constructed-response items, used as anchor items in an equating process, were less stable than the parameters of dichotomous items, a similar result was not found in this study. Next, the results showed that *Process Strand* items were less stable than *Content Strand* items. *Process Strand* generally included more

complex procedures than *Content Strand* items. Items that are contextual or have a story may have been easier to remember than abstract items. If teachers teach to the remembered context, the parameters for a story problem (as an anchor item) could drift. Although fewer Nonroutine/Insightful items were used as equating items, the parameters of three different levels of Cognitive Complexity didn't affect the stability of item parameters. In terms of Item Concreteness, the parameters of Contextual and Concrete/Story items were less stable than the Abstract items. Items requiring problem-solving or reasoning were likely to be placed in a contextual story context. Items measuring problem-solving and reasoning may be more easily recalled by teachers – leading to an emphasis on activities that prepare students for these items.

The second question related to item structure which consists of stem and stimulus? Do any of their characteristics affect the stability of item parameters? The results did not show any effect of item structure; however, the types of stimuli did seem to relate to item parameters instability. The parameters of items that were coded as Picture of Phenomena as well as Graph or Chart as stimuli were less stable; but items that were coded as Text and Equation were stable. Solving items presented with a visual representation, for example, table, graph, or diagram, examinees need to understand, use, think and explain in terms of images. The ability of reading images may cause item parameters to be unstable. Therefore, complex contexts and translation for visual representation may be related to stability.

Finally, the third question considered the changes in common items over years. Do changes in wording affect the stability of item parameters? Do changes in context affect the stability of item parameters? Previous studies have examined the effects of changes in item order on test performance (Whitely & Dawis, 1976; Haertel, 2004; Meyers et al., 2009). The results shown in this study confirmed that item position effects varied depending on factors such as how much the item position changes and the direction of change. In this study, change in item location and appearance were also shown to affect equating results. This study found that when common items were not in the same operational location, the item parameters were unstable. However, the direction of item position movement was not clear. Further, this study agreed with previous studies that any differences in wording, typeface, and other formatting usages were likely to affect the performance for one group of examinees.

One of the drift sources found in the literature came from "security breaches". Researchers pointed out the fact that "teaching tested content" was widespread when tests were used for high stakes decisions. As a result, item release made it difficult for test developers to examine test score validity and interpret test scores. Different local curricular and instructional characteristics may influence parameter estimates as well. Not enough evidence could support these assumptions in this study. However, it is possible that items in memorable contexts may have affected teachers' instructional focuses. For example, Statistics and Probability items with stimuli

coded as Graph or Chart were less stable. Items with stimuli that were coded as Pictures of Phenomena were also less stable. The focuses of these items would be fairly easy to recall.

Equating is a statistical process to establish comparability between alternate forms of a test built to the same content and statistical specifications by placing scores on a common scale. Using common-item non-equivalent groups design, by equating the current year's state test to state tests given in previous years, the performance standard can be maintained over time. Test equating is necessary to be fair to examinees taking different test forms and to provide score-users with scores that mean the same thing, regardless of the items appearing in different test years and taken by different examinees. In order to minimize the threats to item parameter invariance existing at the item level, testing programs measure the same constructs and are usually built to the same test specifications or test blueprint. However, different editions or forms of a test almost always differ somewhat in their statistical properties. Few studies have examined the stability of anchor items after controlling for the context effects between two forms and compared the characteristics of stable items and unstable items. The present study focused on identifying content and context effects that influenced item performance, considering a range of plausible explanations to decide whether a misbehaving item should be discarded from the common-item pool or not, and providing suggestions of choosing common items as equating items in the future. This study reinforced previous studies conducted with multiple-choice items

and, therefore, expands our understanding of what factors to consider when selecting and managing linking items in common-item, nonequivalent-groups equating designs.

REFERENCES

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the Causes of DIF in Translated Verbal Items. *Journal of Educational Measurement, 36,* 3, 185-98.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bock, R. D., Murakl, E., & Pfeiffenberger, W. (1988). Item Pool Maintenance in the Presence of Item Parameter Drift. *Journal of Educational Measurement, 25*(4), 275-285.

Cizek, G. J. (1994). The Effect of Altering the Position of Options in a Multiple-Choice Examination. *Educational and Psychological Measurement. 54*(1), 8-20.

Donoghue, J. R., & Isham, S. P. (1998). A Comparison of Procedures to Detect Item Parameter Drift. *Applied Psychological Measurement. 22*(1), 33-51.

Eignor, D.R. (1985). *An investigation of the feasibility and practical outcomes of preequating the SAT verbal and mathematical sections* (Research report 85-10). Princeton, NJ: Educational Testing Service.

Eignor, D. R., Council of Chief State School Officers, & Association of Test Publishers. (2010). *Operational best practices for statewide large-scale assessment programs*. Washington, D.C:

Council of Chief State School Officers and the Association of Test Publishers.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying Sources of Differential Item and Bundle Functioning on Translated Achievement Tests: A Confirmatory Analysis. *Journal of Educational Measurement, 38,* 2, 164-187.

Goldstein, H. (1983). Measuring Changes in Educational Attainment over Time: Problems and Possibilities. *Journal of Educational Measurement, 20,* 4, 369-377.

Haertel, E. H. (2004). *The behavior of linking items in test equating* (CSE Tech. Rep. 630). Los Angeles, CA: Center for the Study of Evaluation (CSE), National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising Standardized Achievement Test Scores and the Origins of Test Score Pollution. *Educational Researcher, 20*(5), 2-7.

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

Holland, P. W. (2007). A Framework and History for Score Linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5-30). Statistics for social and behavioral sciences. New York: Springer.

Kingston, N. M., & Dorans, N. J. (1984). Item Location Effects and Their Implications for IRT Equating and Adaptive Testing. *Applied Psychological Measurement, 8,* 2, 147-154.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. Statistics for social science and public policy. New York: Springer.

Koretz, D. (2003). *Using multiple measures to address perverse incentives and score inflation.* Educational Measurement: Issues and Practice *22*(2), 18-26.

Linn, R. L. (1993). Educational Assessment: Expanded Expectations and Challenges. *Educational Evaluation and Policy Analysis, 15,* 1, 1-16.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, N.J: L. Erlbaum Associates.

Madaus, G. F. (1988). The Distortion of Teaching and Testing: High-Stakes Testing and Instruction. *Peabody Journal of Education, 65*(3), 29-46.

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

National Research Council (U.S.)., Koretz, D. M., Bertenthal, M. W., & Green, B. F. (1999). *Embedding questions: The pursuit of a common measure in uncommon tests*. Washington, DC: National Academy Press.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased Item Detection Techniques. *Journal of Educational Statistics, 5,* 3, 213-233.

Skaggs, G., & Lissitz, R. W. (1986). IRT Test Equating: Relevant Issues and a Review of Recent Research. *Review of Educational Research. 56*(4), 495-529.

Smith, M. L. (1991). Meanings of Test Preparation. *American Educational Research Journal, 28*(3), 521-542.

Stone, C. A., & Lane, S. (January 01, 1991). Use of Restricted Item Response Theory Models for Examining the Stability of Item Parameter Estimates over Time. *Applied Measurement in Education, 4*(2), 125-41.

Taylor, C. S. (1999a). Washington Assessment of Student Learning, Technical report for 1998, grade 4. Olympia, WA: Office of the Superintendent of Public Instruction.

Taylor, C. S. (1999b). Washington Assessment of Student Learning: Technical report for 1998, grade 7. Olympia, WA: Office of the Superintendent of Public Instruction.

Taylor, C. S. (1999c). Washington Assessment of Student Learning: Technical report for 1998, grade 10. Olympia, WA: Office of the Superintendent of Public Instruction.

Taylor, C. S., & Lee, Y. (2010). Stability of Rasch Scales over Time. *Applied Measurement in Education*, *23*(1), 87-113.

Tenenbaum, I., Lindsay, S., Siskind, T., Wall-Mitchell, M. E., & Saunders, J. (2001). Technical documentation for the 2000 Palmetto achievement challenge tests of English language arts and mathematics. Columbia, SC: South Carolina Department of Education.

Traub, R.E. (1983). Apriori considerations in choosing an item response. In R.K. Hambleton (Ed.), *Applications of item response theory* (pp. 57–70). Vancouver, BC: Educational Research Institute of British Columbia.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The Effect of Item Parameter Drift on Examinee Ability Estimates. *Applied Psychological Measurement. 26*(1), 77-87.

Whitely, S. E., & Dawis, R. V. (1976). The Influence of Test Context on Item Difficulty. *Educational and Psychological Measurement, 36,* 2, 329-337

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.

Yen, W. M., Green, D. R., & Burket, G. R. (1987). Valid Normative Information from Customized Achievement Tests. *Educational Measurement: Issues and Practice, 6,* 1, 7-13.

APPENDIX

WASL equating items from 1999 to 2004

Year	ID	Item No. in Year 1	Item No. in Year 2	Item
1999-2000	1	1	1	Units of measure in the metric system
	2	4	4	The greatest rectangular garden area
	3	12	12	Mean test score in algebra class
	4	13	13	Points on Enrique's grid
	5	17	17	Chirping affected by air temperature
	6	18	18	Graph of population distribution
	7	20	20	Pizza check split by four
	8	21	21	Equation representing apartment rent
	9	25	25	Length of model car from ratio
	10	26	26	Moth trace sketched symmetrically
	11	27	27	Graph of employee salary
	12	31	31	Expression be divisible by 6
	13	32	32	Objects balanced on scale
	14	44	44	Number of triangles continuing pattern
	15	45	45	Total cost of the fencing
2000-2001	16	2	2	Shadow length of hoop from ratio
	17	5	5	Probability of winning tickets
	18	10	10	Ratio of surface area to volume
	19	19	19	Equation based on x/y table
	20	22	20	Position of new coordinates of point
	21	24	26	Measure of angle in parallelogram
	22	29	29	Sum of two rolled number cubes
	23	33	35	Numbers in increasing pattern
	24	34	36	Volume of cylindrical storage tank
	25	35	37	Measure of central angle of section
	26	41	43	Number of apples bought at store
	27	43	44	Average weight of a bobsled team
2001-2002	28	1	2	Properties of a and b
	29	3	3	Probability of sum on spinners
	30	4	4	Bar graph representing percentage
	31	7	7	Factor of a expression (3a^2+12a)
	32	9	9	Line graph of two temperatures scales
	33	12	12	Equation of car rental cost
	34	14	14	Explanation of "Drama" in diagram
	35	16	16	Angle relationships in triangle
	36	17	17	Total amount of concrete in cubic feet
	37	24	25	Boxes ordered least to greatest volume
	38	33	34	Triangle having most lines of symmetry
	39	34	35	Width of strip on perpendicular sides

Year	ID	Item No. in Year 1	Item No. in Year 2	Item
2002-2003	40	8	8	Graph of tape length and recording times
	41	10	10	Containers ordered greatest to least volume
	42	11	11	Original cost price before discount
	43	13	13	Expression of number of dollars left
	44	20	20	Angle in isosceles triangle
	45	27	27	Patterns of series of numbers
	46	29	29	Money earned based on salary options
	47	32	32	Median amount of monthly precipitation
	48	37	37	Area of carpet in triangular showroom
	49	39	40	Probability of tied chess game
	50	44	44	Winning percentage for the season
	51	45	45	Likely result of flipping pennies
2003-2004	52	1	1	Total number of cans displayed
	53	3	3	Point facing on diagram after rotations
	54	9	9	Possible outcomes of cube and coin
	55	14	14	Probability of hitting in a dart game
	56	16	16	Equation of four symbols
	57	21	21	Shaded figure reflected horizontally
	58	30	29	Comparison of transportation costs
	59	33	33	Total salary based on TV sold
	60	43	44	Sand cost needed to fill sandbox