IMPACT OF INSTRUCTIONAL SENSITIVITY ON HIGH-STAKES ACHIEVEMENT TEST ITEMS: A COMPARISON OF METHODS

By

Jie Chen

Submitted to the graduate degree program in Department of Psychology and Research in Education and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairperson: Dr. Neal Kingston

Dr. Bruce Frey

Dr. Vicki Peyton

Dr. William Skorupski

Dr. Reva Friedman

Date Defended: April 24, 2012

The Dissertation Committee for Jie Chen

certifies that this is the approved version of the following dissertation:

IMPACT OF INSTRUCTIONAL SENSITIVITY ON HIGH-STAKES ACHIEVEMENT TEST ITEMS: A COMPARISON OF METHODS

____

Chairperson: Dr. Neal Kingston

Date approved: April 24, 2012

ABSTRACT

Living in an era of test-based accountability systems, how do we hold accountability tests accountable? Many accountability decisions made today are based on the assumption that test scores successfully reflect the effect of instruction. However, only instructionally sensitive assessments, not the instructionally insensitive ones, reflect the impact of instruction. The purpose of this study is to explore the relationship between students' instructional experiences and their test scores on standardized achievement test items. The Mantel-Haenszel statistics, logistic regression and judgmental item-detection approaches were used to identify instructionally sensitive items in the Kansas Mathematics Interim Assessment for seventh graders. The two empirical methods performed very similarly. Many instructionally sensitive items were identified by the empirical methods. No strong agreement between the empirical and judgmental approaches was found. The implications of this study to educators and policymakers, the limitations of this study, and the directions for further studies are discussed.

ACKNOWLEDGEMENT

The completion of this dissertation would not be possible without many peoples' support and help. I would like to express my deep gratitude to my advisor, Dr. Neal Kingston. He gave me freedom to explore any topics that I was interested in, encouraged me to face and overcome every challenge that I met, and confirmed and was happy for every achievement I made. What he devoted was not only time advising a dissertation, but also meticulous care educating a graduate student who will benefit from her studies in graduate school and see the value of this experience in the future. My sincere thanks also go to other members of my dissertation committee: Dr. Bruce Frey, Dr. Vicki Peyton, Dr. William Skorupski, and Dr. Reva Friedman. Especially, I thank Dr. Frey for his help with the format and structure of my proposal, Dr. Skorupski for his help with the methodology part, and Dr. Peyton for giving me time working on my dissertation by kindly reducing my GTA responsibility. I also would like to thank Dr. Paul Johnson, who helped me to correctly interpret the results of logistic regression produced by SPSS. I am also grateful to Dr. Angela Broaddus for her generous help with data collection and item review and Gail Tiemann for her patient help with HSCL application.

My heartfelt gratitude is to my mother, Shumin Zhang, my father, Zhongyi Chen, my husband, Tao Guo, and my son, Chuxiao Guo. Their loving encouragement and enduring support are the warmest light in my life. They are the good reason for me to finish this journey and to start an even more exciting one.

I thank God for being my "ever-present help in trouble" and for making every impossible a possible in His hands.

Table of Contents

List of Tables

List of Figures

List of Appendices

Impact of Instructional Sensitivity on High-Stakes Achievement Test Items: A Comparison of Methods

CHAPTER 1

INTRODUCTION

This study examines instructional sensitivity of items on standardized achievement tests. Using a state interim assessment program, the present study seeks to determine the relationship between students' instructional experiences and their scores on test items. This chapter presents the background of the study, statement of the problem, purpose of the study, research questions, and significance of the study.

Background and Importance of the Study

It is important that accountability tests assess what is taught. Unfortunately, research suggests that many high-stakes achievement tests in the United States failed to effectively reflect whether students' teachers successfully covered and delivered the necessary content in their instruction (Popham, 2007a; Popham, 2007b; Pham, 2009). For example, Phillips and Mehrens (1988) examined the impact of different curricula on standardized achievement test scores both at item and at objective levels but failed to detect differential curricular impacts on students' test scores. In their study, Stanford Achievement Test scores for students in grades three and six in a school district from a middle-sized Midwestern city were used. Two indicators of curricular impact were (1) the degree to which the curricula matched the content of the standardized test and (2) the actual textbook series used within each building (classroom). The classical test theory (CTT) methods were used to examine the impact of curricular differences on standardized test item difficulties. This finding from their study parallels the results of the Pham (2009), the Mehrens and Phillips (1986), and the Phillips and Mehrens (1987) studies that reported at most there was minor impact of curricular differences detected by achievement tests. Further, the Mehrens and Phillips (1986) and the Phillips and Mehrens (1987 & 1988) studies emphasized that curricular differences at the objective and item levels were too small to have any practical significance. These studies pointed out the fact that the impact of curricular differences on the results of standardized achievement tests was peripheral. However, just as Goe (2007) cautioned, one of the reasons for the weak relationship between curricular differences and student performance could be due to the fact that the measurement tools (e.g., statewide standardized student achievement tests) are not sensitive enough to capture the effect of instruction or any other factors of interest.

Some other research accentuated the value of teaching strategies and content of instruction to performance assessments (Niemi et al., 2007). Niemi and colleagues investigated the instructional sensitivity[1] of a standards-based ninth-grade performance assessment (i.e., the California English-Language Arts Content) and found that assessment scores improved in response to instruction specifically targeting the assessed construct. However, in their study, the overall performance assessment score instead of the item score was utilized in the analyses. Therefore, no item information was available to capture effects of the instruction and to reveal which content may need more instruction. Still, more evidence is needed to corroborate the instructional sensitivity of standardized assessments.

Historically, standardized achievement tests are considered helpful because they rank students based on what students know and can do but not because they successfully measure how

[1] See pp. 3-4 for the definition of *instructional sensitivity*.

well those students have been taught. "When state standardized achievement tests are not aligned with standards, curriculum, and materials used in classrooms, student learning may not be reflected accurately in test scores" (Goe, 2007, p.15). As many studies were conducted purporting to explore the link between teacher quality[2] and student outcomes (Darling-Hammond, 2000; Fenwick, 2001; Jacob & Lefgren, 2004), some only found a weak relationship between factors that are supposed to be logically related to student achievement (Jacob & Lefgren, 2004); some even failed to detect this relationship.

However, in the era of accountability, schools are perceived as better or worse based on their proficiency, readiness, or growth; and teachers are believed to be more effective when their students perform better on high-stakes achievement assessments (Court , 2010 & Popham, 1995). Tests, as a tool of accountability, should be intended to measure how well students have been taught (Popham, 2010a). Therefore, the purpose of this study is to determine the relationship between students' instructional experiences and their scores on test items on standardized achievement tests. Students' instructional experiences were measured by their opportunities to learn (OTL). This research detected instructional sensitivity of standardized test items in a state testing program by looking at the linkage of item statistics and content covered in instruction.

Defining Instructional Sensitivity

According to Popham, *instructional sensitivity* is "the degree to which students' performances on a test accurately reflect the quality of instruction specifically provided to promote students' mastery of what is being assessed" (2006, p. 1). In Pham's (2009) dissertation, *instructional sensitivity* is defined as "responsiveness to the varying pedagogical practices of

[2] Goe (2007) grouped teacher quality into three categories: (1) teacher quality defined on **teacher inputs** which include teacher qualifications and teacher characteristics; (2) teacher quality defined on **teacher processes** (i.e., teacher practices); and (3) teacher quality defined on **outcomes** (i.e., teacher effectiveness).

teachers and [it] allows for standardized testing to be used as an accountability tool" (p. 117). Haladyna and Roid (1981) defined *instructional sensitivity* as "the tendency for an item to vary in difficulty as a function of instruction" (p. 40). In the Niemi and colleague (2007) study, instructionally sensitive assessments are "assessments that can measure the effects of previous teaching, and they can also be used as outcome measures to evaluate instruction, as well as to identify students who need additional instruction" (p. 216). All the above definitions emphasize a fact that instructional quality is an important part of the school environment and that instructional sensitivity is an important index of an effective or well-designed achievement test, which serves as a tool of accountability.

Statement of the Problem

Educational tests have been criticized by educators and researchers for their lack of accurately reflecting the quality of teachers' instructional efforts. This critique indicates that testing has lost its function of being a tool to evaluate how much students have benefited from schooling. Although recent attempts have been made to investigate the link between testing and instruction or teacher quality or effectiveness, studies have shown different foci and mixed findings in terms of the sensitivity of test items to instruction. From an educational measurement perspective, several issues in existing studies need to be addressed. First, when instructional sensitivity is discussed, the actual level of focus or interest should be at the item level, not at the test or objective level. A review of literature shows that many studies only explored the *instructional sensitivity* or the *instructional validity*[3], by using mean scores in standards-based tests (D'Agostino, Welsh & Corson 2007; Cohen & Hill, 1998; Niemi et al., 2007). However, it

[3] See p. 15 for the definition of *instructional validity*.

is very likely that the aggregated test scores mask the sensitivity of each test item (Airasian & Madaus, 1983). Thus, the objective-level score or overall test score can hardly be a good or accurate indicator because the total test score is greatly influenced by the dynamic of all the items in one test form, and is therefore not stable.

Second, although there are two ways to detect instructional sensitivity of test items, empirical strategies and judgmental strategies, most studies in instructional sensitivity so far have used empirical strategies. Because they are logistically simpler and less expensive, it may be desirable to make some efforts to implement judgmental item-detection strategies in the study of instructional sensitivity. One of the major strengths of judgmental methods is that they can estimate an accountability test’s instructional sensitivity as a continuum (Popham, 2003b). Popham (2007b) described using an 11-point Likert scale (Popham, 2003b, see Chart 1) to rate the tests on four evaluative dimensions, and the instructional sensitivity of items is just one of the dimensions.

0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Totally Unclear										Extremely Clear

Chart 1: The 10-Point Scales Rating Tests on Evaluative Dimensions

The four evaluative dimensions include (Popham, 2007b, p. 148):

1. The number of curricular aims assessed;
2. The clarity of assessment targets;
3. The number of items per assessed curricular aim;
4. The instructional sensitivity of items.

He further suggested that either equal weight or different weights needs to be assigned to each dimension when an instructional sensitivity review is practiced. Obviously, the instructional sensitivity review is imperatively important because, first, decisions of the panels of curriculum specialists and teachers are based on the evaluation of more than one category of a test; and, second, remedial actions can be taken if an accountability program's tests are indicated as instructionally insensitive.

Third, while there are some studies empirically determining instructional sensitivity at item level (Muthén, 1989; Muthén1988a; Mehrens & Phillips, 1986; Mehrens & Phillips, 1987), studies purporting to examine the congruity and/or incongruity between judgmental and empirical item-detection techniques have not been conducted thus far. Since different empirical techniques are based on different statistical assumptions, results obtained from using different techniques are slightly different in some aspects. Each of the instructional sensitivity indices has its strengths and weaknesses. Perkins (1984) compared seven[4] instructional sensitivity indices in regard to its accuracy of estimating instructional sensitivity of items in the grammar subtest of the Michigan Test of English Language proficiency. Results showed that the normalized difference between pre- and post-test logits of difficulty (item response theory) and the pre-to-post difference index (criterion-referencing) were the most robust. He further concluded that decisions about discarding or revising test items should not be based solely on instructional

[4] (1) the pre-to-post difference index (PPDI) introduced by Cox and Vargas (1966); (2) the percent of possible gain(PPG) introduced by Brennan and Stolurow (1971); (3) the combined-samples (instructed and uninstructed examinees) point-biserial correlation (COMPBI) introduced by Haladyna (1974); (4) four indices, 01, 11, 00, and 10, introduced by Popham (1971); (5) three test item discrimination indices, D1, D2, and D3, suggested by Helmstadter (1974); (6) item response indices: pretest logit of difficulty (PRELOGIT), posttest logit of difficulty (POSTLOGIT), pre-to-post difference in logit of difficulty (XLOGIT), and the normalized difference in item response difficulty estimates for the uninstructed and instructed samples (ZDIFF); and (7) three indices based on the Bayes' theorem which were proposed by Helmstadter (1974) and Haladyna and Roid (Perkins, 1984, pp. 4-5).

sensitivity indices. For this reason, a study that compares the two strategies of item-detection in terms of instructional sensitivity is requisite.

Purpose of the Study

Although the importance of instructional sensitivity of accountability tests has been recently advocated by researchers, there is little evidence to convince educators that their instruction has crucial impact on students' performance on high-stakes achievement tests. Neither do researchers know much about the extent to which the evidence of validity of teaching that is evaluated by test scores is reliable. This study explored the relationship between teaching and learning by relating achievement responses to instructional experiences. Students' instructional experiences (i.e., the opportunity to learn) were used as the grouping variable to classify students into two groups (i.e., the instructed group and the uninstructed group) for each of the test items. The Mantel-Haenszel (MH) tests (Mantel & Haenszel, 1959) and the logistic regression (LR) procedures were used to detect instructionally sensitive items on Kansas State Interim Assessment in Mathematics. Next, thirteen curriculum experts (i.e., middle school math teachers) were recruited to individually review and rate each of the test items in terms of its instructional sensitivity. The teachers were asked to report, in the format of written notes, to the researcher what items reflected the impact of instruction. The reported items were the ones that were instructionally sensitive according to the judgmental item-detection procedures. Finally, the pros and cons of judgmental and empirical item-detection strategies were compared and discussed for the purpose of determining the congruity of the two methods and the reliability of each of these methods.

Research Questions

In this study, the following research questions were addressed:

1. To what extent are the items in the state testing program sensitive to instruction?
2. Are item performance differences related to differences in curricular content covered in instruction?
3. How does the instruction in content, tested on the state testing program, influence students' performance?
4. To what extent do the instructional sensitivity judgments made by curriculum experts agree with the results of empirical methods?

Hypotheses

The researcher expected that students who have been presented the material tested as part of their instruction tend to answer correctly the items measuring the content covered in instruction. It was also hypothesized that students who were not instructed in the materials tested would tend to answer the corresponding questions correctly after receiving the instruction. The researcher believed that "uninstructed students perform at a low level prior to instruction and at a high level [*sic*] following instruction" (Haladyna & Roid, 1981, p. 40) if the intended effects of instruction on student learning could be successfully reflected by test results. It was further hypothesized that the items sensitive to instruction would reflect the impact of effective instruction on students' performance on standardized achievement test items. On the contrary, assessments containing items "that are not sensitive to well-designed instruction do not reveal what students have learned or should learn from instruction and are presumably measuring general intelligence, irrelevant constructs, or opportunities to learn outside of school" (Niemi et al., 2007, p. 216). The last hypothesis was that the items detected as instructionally sensitive by the curriculum experts and the items flagged as instructionally sensitive by using empirical methods should match systematically.

To sum up, the following three hypotheses were tested in this study:

1. The difference in difficulty between the group receiving instruction and the group not receiving instruction should be greater for items teachers judge to be instructionally sensitive than for items teachers judge to not be instructionally sensitive.
2. For items judged to be instructionally sensitive, at a fixed ability level, students who were taught the content tested should have a higher probability of responding correctly to the item than those who were not taught.
3. The items identified as functioning most differently for the instructed and not instructed groups should, to a great degree, match the items detected as sensitive by the curriculum experts.

Significance of the Study

The rationale of instructional sensitivity as an item characteristic is that students perform at a lower level prior to instruction and at a higher level after instruction. The overriding theme of the *No Child Left Behind Act* (NCLB) is accountability, including accountability through adequate yearly progress (AYP) (Simpson, LaCava & Graner, 2004). An appropriate NCLB test should be instructionally sensitive, meaning it has the capacity to detect genuine improvement in instruction (Popham, 2003a). In other words, students do relatively poorly on items before the content of the items is taught and relatively well after it is taught. Viewed from the item perspective, items should get significantly easier after instruction – they should be instructionally sensitive. Items for which this is not the case are inappropriate for accountability purposes.

Compared to the past studies in instructional sensitivity, this study is unique in its combination of judgmental item review and empirical item analysis and its effort made to explore what common characteristics the instructionally sensitive items have. Some researchers believed and argued that the accuracy of high-stakes achievement test scores heavily depends on the sensitivity of these tests to instruction (Popham, 2007a, 2010a, 2007b, 2006, 2003a; D'Agostino, Welsh & Corson, 2007). Then, the vital concern becomes how the educators and officials can be convinced that better instruction will lead to higher test scores.

Flagging items that are instructionally sensitive is important but not sufficient for improving instruction. Rather, it is more important to identify instructional approaches that influence item characteristics and also item characteristics that limit instructional sensitivity. D'Agostino and colleagues (2007) remarked that the previous studies did not pinpoint what instructional approaches influenced item characteristics because the empirical methods used did not have the actual measurement of instruction. The items or tests are "marked" as either sensitive or insensitive without successfully providing additional content-related evidence. On the other side, when the panels of trained judges (i.e., the curriculum experts or bias reviewers) are rating the tests or items, their judgment cannot be immediately confirmed until after the test is administered and scored. Therefore, it is promising to investigate the similarities between judgmental evidence and empirical evidence for the sake of credibility.

In her dissertation, Kao (1990) made an effort to explore a common characteristic of instructionally-sensitive items. However, her attempt only ended in finding out what topics the instructionally-sensitive items were measuring. Kao's findings include: 1) the eight detected sensitive items measured varied topics, such as *coordinates*, *measuring angles related to a triangle*, *congruence of plane figures*, *square roots*, and *powers and exponents*; 2) items measuring *square roots* tend to have high D values (the square root of the sum of squares of the distance between two item characteristic curves), indicating this topic may be easily impacted by learning experience rather than by ability; 3) items of topics related to definition or initial learning may be prone to be sensitive (Muthen, Kao & Burstein, 1991), but there is no evidence indicating that "items with these characteristics will be routinely sensitive to instructions" (Kao, 1990, p.92). To summarize, Kao's study uncovered topics that tend to be instructionally sensitive but did not further explore why items testing these topics were apt to be sensitive to instruction.

This study expands Kao's effort by especially investigating not only in topics, but also in content standard, what characteristics of the items were likely to contribute to an instructionally supportive test. In particular, Popham's and colleagues' criteria (2005) for an instructionally supportive accountability test were used for reference to make the judgment whether an item was well-developed to reflect and support instruction. According to these criteria, an instructionally supportive accountability test should (Popham et al., 2005, p. 125):

1. Measure a modest number of significant curricular aims;
2. Provide clear descriptions of what is to be assessed;
3. Supply instructionally informative results to teachers, students and students' parents.

This study employed judgmental approaches for item instructional sensitivity to answer the following questions:

1. Does the Kansas Interim Assessment in Mathematics for seventh graders measure a modest number of significant curricular aims?

2. Is each indicator of the benchmark stated with sufficient clarity that almost all of the state's teachers can identify what the indicator really means?
3. For purposes of a teacher's instructional planning, how clearly are the state test's assessment targets (the skills and knowledge it measures) described?

Summary

This chapter provides the background of the study, definition of the central concept, a statement of the problem, the purpose of the study, research questions and hypotheses, and the significance of the study. In Chapter 2, a review of relevant literature will be summarized. An overview of previous studies in instructional sensitivity, standards-based assessment, the opportunity to learn, two strategies of detecting instructionally sensitive items (with an emphasis on empirical item-detection approaches), and the importance of studying instructional sensitivity in the accountability tests will be presented. In Chapter 3, the research design will be described. It will include a description of the data source, model and variables selected, data analysis, and a summary of the research design.

CHAPTER 2

LITERATURE REVIEW

This chapter first provides an overview of the definitions or concepts of *instructional sensitivity, instructional validity, curricular validity* and the relationship among these three concepts. The second section reviews *standards-based assessment* and how it is related to instructional sensitivity. The third section presents studies on the *Opportunity to Learn* (OTL) and its relationship with instructional sensitivity. The fourth section focuses on the two strategies detecting instructional sensitivity*: judgmental item-detection techniques* and *empirical procedures*. In the last section, the relationship between instructional sensitivity and accountability tests is described.

Instructional Sensitivity, Instructional Validity & Curricular Validity

Instructional Sensitivity vs. Instructional Validity

Differences

Yoon and Resnick (1998) and Gordon (2008) described *instructional validity* as the extent to which an assessment is systematically sensitive to the nature of instruction offered. This differs from *instructional sensitivity,* which focuses on the item level and is not discussed on the test level.

Similarities

In some other studies, the terms *instructional sensitivity* and *instructional validity* were used interchangeably (D'Agostimo, Welsh & Corson, 2007). In order to explain the relationship between instructional validity and the OTL, D'Agostimo and colleagues (2007) stated that "instructional validity is a narrower concept that refers to the ability of a test to detect instructional differences that might arise due to OTL" (p. 4). This statement echoes, though differs from, Popham's (2006) definition of instructional sensitivity as "the degree to which students' performances on a test accurately reflect the quality of instruction specifically provided to promote students' mastery of what is being assessed" (p. 1).

The above comparison of *instructional sensitivity* and *instructional validity* indicates that they only differ on the level of analysis. Calfee (1983) questioned and clarified the match between the test and the curriculum by presenting a contrast: "should *instructional validity* mean that (1) the content of a test is adequately covered by the curriculum, or that (2) the curriculum (or some subset of the curriculum) is adequately covered by the test?" (pp. 106-107). In this current study, the term "instructional sensitivity" was used; the researcher believes both points remarked by Calfee are important. That is, the content of a test should be covered by the curriculum, and the test items should well represent the curriculum covered in instruction. Thus, the measurement process can help to model better instruction.

Curricular Validity vs. Instructional Validity

Unlike *instructional sensitivity* and *instructional validity*, which were regarded either as the same thing in some studies or as distinct concepts in some other studies, most studies distinguished *curricular validity* and *instructional validity* as two different concepts. According to Linn (1983), *curricular validity* refers to "the degree to which test items represent objectives of the curriculum," while *instructional validity* refers to "the degree to which the topics measured by the test were actually taught" (p. 115). Airasian and Madaus (1983) compared these two concepts and reached a similar conclusion: "Curricular validity refers to the link between test content and the content of the curriculum material used in schools. Instructional validity goes one step further in referring to the match between test content and the content of the instruction actually provided pupils in classrooms" (p. 109).

Although there are differences in definitions of the two concepts of *curricular validity* and *instructional validity*, educational specialists can easily find the connection of the concepts of *instructional validity* and *instructional sensitivity*. They both lay emphasis on the match between the knowledge and skills provided to students in instruction and the content assessed in the tests. Therefore, if a test is sensitive to instruction, the scores on this test should be differentially affected by high-quality instruction (Baker, 2008). To quote Haladyna and Roid (1981), *instructional sensitivity* is "the tendency for an item to vary in difficulty as a function of instruction" (p. 40). When the item is sensitive to instruction, students after instruction should have a higher probability of answering it correctly, compared to those who respond to the item prior to instruction.

To maintain consistency, this study uses the term *instructional sensitivity*, instead of *instructional validity*, to discuss the importance of the accountability tests accurately distinguishing students who have been well taught from those who have not. Figure 1 depicts the relationship among *instructional sensitivity*, *instructional validity* and *curricular validity* based on the literature review.

Figure 1. Relationship among Instructional Sensitivity, Instructional Validity and Curricular Validity

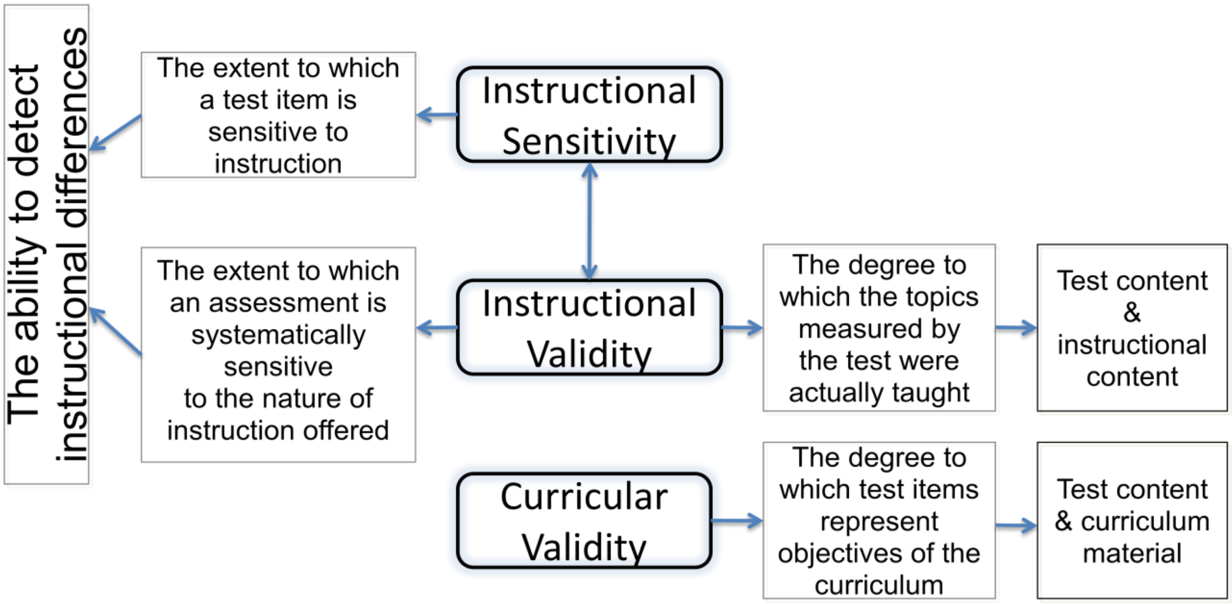

The ability to detect instructional differences
The extent to which a test item is sensitive to instruction
Instructional Sensitivity
The extent to which an assessment is systematically sensitive to the nature of instruction offered
Instructional Validity
The degree to which the topics measured by the test were actually taught
Test content & instructional content
Curricular Validity
The degree to which test items represent objectives of the curriculum
Test content & curriculum material

Figure 1. The double-headed arrow between Instructional Sensitivity and Instructional Validity indicates these two concepts were considered and used interchangeably in some studies and contexts. The lack of arrows or lines between Instructional Validity and Curricular Validity indicates that Curricular Validity is distinguished as a different concept from both Instructional Validity and Instructional Sensitivity.

Standards-Based Assessment and Instructional Sensitivity

Standards-Based Assessment

Standards-based assessment is a comparatively new concept that is a key component in standards-based reform. First, states set educational standards that define what students should know and be able to do. Then students are instructed to meet the expected standards. Finally, the students are assessed to determine if they meet these standards. Therefore, standards-based tests are designed to support improved student achievement, and the results of the tests should allow educators or other clients to determine whether a school has successfully promoted students' mastery of that state's content standards (Educational Commission of the States, 2002; Popham, 2001). In terms of assessments' function in increasing accountability and stimulating improvement in students' academic performance, standards-based assessments are characterized as follows (Educational Commission of the States, 2002, pp. 2-3):

- Closely links assessment to curriculum.
- Compares students to a standard of achievement, not to other students.
- Incorporates new forms of assessment (e.g., requiring students to write an essay or solve a real-life math problem).

The second feature listed above indicates how a standards-based test is connected to its traditional counterpart– a criterion-referenced test. Criterion-referenced tests are known for detecting an individual's status against some criteria or standards rather than other individuals (Popham & Husek, 1969). Another characteristic that standards-based tests and criterion-referenced tests share is that they both are used to evaluate instructional programs (Popham & Husek, 1969) and further to improve educational quality (Popham, 2001). To achieve this purpose, a pivotal factor that determines its potential for supporting improved student achievement is that standards-based assessment should be designed to align standards with assessment and instruction. However, most of today's standards-based measurements are criticized for their undetectable contribution to improved instruction (Popham, 2001). To ensure that standards-based assessment makes meaningful contributions to improved instructional quality, Popham (2001) proposed four rules to be followed (pp.4-6):

> Rule 1: Require curricular personnel to prioritize the most important outcomes they want children to achieve, and then develop tests to assess only the highest priority outcomes that can be both accurately assessed and instructionally accomplished.

Rule2: Construct all assessment tasks so an appropriate response will typically require the student to employ (1) key enabling knowledge and/or subskills, (2) the evaluative criteria to be used in judging a response's quality, or (3) both.

Rule3: Create a sufficiently clear description of the knowledge and/or skills represented by the test so that teachers will have an understanding of the cognitive demands required for students' successful performance.

Rule4: The items and description(s) of any high-stakes test should be reviewed at a level of rigor commensurate with the intended uses of the test.

Rules 3 and 4 reflect the concern about whether today's educational tests are instructionally functional.

Instructional Sensitivity and Standards-Based Tests

According to Popham (2001), large-scale educational testing is labeled "standards-based assessment" because of the national emphasis on promoting students' mastery of content standards. A standards-based test is used to see how well the test takers can do, and how much they master in terms of knowledge and skills, which they are expected to have mastered by a certain grade level.

Traditionally, instructional sensitivity indices (ISIs) were used with criterion-referenced mastery tests (Haladyna & Roid, 1981; Popham & Husek, 1969; Shannon & Cliver, 1987), which is congruent with most studies' concerns and arguments (i.e., Shannon & Cliver, 1987; van der Linden, 1981; Haladyna & Roid, 1981); it is primarily because close linkage between content taught and content tested is often found where criterion-referenced mastery tests are used (Hanna & Bennett, 1984). For this reason, criterion-referenced tests are expected to provide feedback to learners as well as teachers, so that instructional programs can be evaluated and improved (Haladyna & Roid, 1981). In the era of standards-based reform, a standards-based assessment is a crucial part of educational testing programs and plays a fundamental role in improving educational quality. A standards-based test must be able to detect substantial year-to-year improvements in students' scores. Otherwise, it is not an "instructionally helpful standards-based test" (Popham, 2001, p. 4). Then, what are the ingredients that characterize a helpful standards-based test? An effective standards-based test should be designed to align with the state's content standards and be composed with items sensitive to instruction.

Opportunity to Learn

The information on examinees' instructional experiences is essential in the investigation of instructional sensitivity. In previous studies, the opportunity to learn was used as the variable of students' instructional information (Kao, 1990; Kim, 1990; Lehman, 1986; Switzer, 1993; Yu, 2006). The OTL refers to whether the students are given equal opportunity to learn in classrooms. Yu, Lei and Suen (2006) classified the definition of OTL into two themes: OTL as *allocated time* for learning and OTL as *content coverage* in teaching (in other words, OTL as *content overlap* between what is taught and what is tested). In terms of content coverage in teaching, it can be either topic-related OTL or item-specific OTL (Kao, 1990). Walker (1983) summarized the indicators of a proper opportunity to learn as follows (p. 176):

1. The item appears in official (district/state) course of study repeatedly (at least three times), is flagged as important for teachers to teach, has adequate time allocations, and has provisions for activities facilitating retention and remediation.

2. The item appears in textbooks and other required curriculum materials repeatedly (at least three times), is emphasized there as important for students to learn, has sufficient page allotments, and is in a form students can comprehend.
3. The item repeatedly appears on classroom, school, and district tests used in grading (at least three times).
4. Teachers report providing their classes with an opportunity to learn it, spending sufficient time on it, and ensuring systematic attempts to identify and remediate learning failures.
5. Students report having been informed of test purpose and objectives before the learning opportunities had passed, having an opportunity to learn it, and spending sufficient time on it.

Similar to Walker's list above, Kao (1990) remarked, "[h]ow to measure OTL is an issue deserving more attention because the validity and reliability of OTL determines the success or failure of studying the instructional sensitivity issue" (p. 8). Yu and colleagues (2006), especially, pointed out the complication of assessing OTL for two reasons: (1) the multidimensional nature of OTL, and (2) the practical problem of how and from whom it is best to collect OTL data. The common methods used to measure OTL or to collect OTL data include the analysis of the instructional materials (Popham & Lindheim, 1981, cited from Kao, 1990), questionnaires for teachers' and/or students' self-report on instructional practices (Cohen & Hill, 1998; Yoon & Resnick, 1998; Wiley & Yoon, 1995; Kao, 1990; Kim, 1990; Yu, 2006) and teacher and/or student interviews (Goe, 2007; Gordon, 2008; Herman & Klein, 1997). Accordingly, a debate emerged about whether OTL information from teachers or from students was more reliable. Lehman (1986) argued that the OTL information from students was less reliable for several reasons. For example, students might not pay attention to what the teacher was trying to teach.

Students might not get the point of what the teacher was presenting. Students might be absent when the teacher was teaching a certain topic. Also, students might tend to believe they were not exposed to a certain topic if the item testing that topic is difficult. Although the shortcoming of the teacher's report of OTL information is that it may not reflect the fact that students in the same classroom may have very different instructional experiences (Kao, 1990; Switzer, 1993), most researchers believed that OTL information from teachers works better to represent the instructional coverage for test items (Lehman, 1986; Kao, 1990).

Obviously, there are important implications from the study of OTL to different stakeholders: policy makers, educational researchers, school teachers and administrators, and so on. This research will focus on its implications for better test development and construction. As stated by Yu and colleagues (2006), "[t]he core purpose of test fairness research is to find out if a test is biased against a particular group of examinees" (p. 6). Although research has been done on gender, race or socioeconomic status (SES) in test performance and test bias, it only puts emphasis on how children bring these differences to school, not on instruction that overlooks these differences. Popham (2007a) urged that students' scores should be capable of distinguishing between effective and ineffective instruction. Therefore, this study will continue this effort on detecting instructional sensitivities of items in accountability tests.

Empirical Detection of Instructional Sensitivity

To categorize broadly, there are two types of item-detection strategies employed in the study of instructional sensitivity (Popham & Kaase, 2009). The first is *judgmental* item-detection techniques. This strategy relies on the judgment of a group of bias reviewers (usually curriculum experts) who review a test's under-development items and suggest whether the items are potentially instructionally sensitive. The major advantage of the judgmental item-detection strategy is that the instructionally insensitive items can be filtered out before they are used in the test form. The second strategy involves the use of *empirical* procedures, which happen after the completion of a test. The empirical item-detection approaches are generally identified as a form of *differential item functioning* (DIF) tests, because they focus on detecting and isolating items that individuals with the same ability but from different groups do not have the same likelihood of success on. DIF is present when a test question is especially difficult (or especially easy) for a special group of test takers, after controlling for the overall ability of the group. In other words, when examinees from different groups have different probabilities of answering a test item correctly after they have been matched on the ability of interest, the test item can be flagged as a DIF item (Swaminathan & Rogers, 1990; Clauser & Mazor, 1998; Peyton, 2000; Woods & Grimm, 2011). This section focuses on the empirical procedures.

DIF is a necessary but not a sufficient condition for item bias (Clauser & Mazor, 1998). Clauser and Mazor (1998) further expounded that if differences exist after matching examinees on the ability of interest, then performance on that item depends on something else than that which has been taken into account. If the second thing is a nuisance, the item is biased. However, if the second thing is an additional interest of the research, there exists item impact, rather than item bias. In this study, students were matched on their general mathematics proficiency. If the group without instruction on the item was less proficient than the other with instruction, the between-group differences in performance were evidenced as DIF. However, this item may not be biased because instructional impact is considered a relevant factor with respect to the purpose of this study. On the contrary, the item was considered sensitive to instruction. Based on the examination of students' responses to test items, the purpose of empirically examining items is to improve items, instead of selecting them, before they are included in a domain (Haladyna and Roid, 1981).

Popham and Kaase (2009) believed that more interest and focus was needed on empirically identifying items that might be instructionally insensitive. Haladyna and Roid (1981) mentioned four theoretical contexts to consider instructional sensitivity empirically: a) *criterion-referenced*, b) *classical*, c) *item-response*, and d) *Bayesian*. The index introduced by Cox and Vargas (1966), for criterion-referenced tests was the *pre-to-post difference index* (PPDI), wherein an item which discriminates perfectly or nearly so, between pre- and post-training groups should fail pre- or poorly-trained test takers but be in favor of post- or well-trained test takers. PPDI simply refers to "the difference in difficulty (percent correct responses) observed when an item was given first to uninstructed students, and then to instructed students" (Haladyna and Roid, 1981, p. 40). This index was enhanced by Brennan and Stolurow (1971, cited from Haladyna & Roid, 1981) when they calculated the percent of possible gain (PPG) as PPG = PPDI/ (1.00 – Pretest Difficulty). The advantage of PPG over PPDI is that the former takes into account the potential for an item to demonstrate improvement.

The *classical item discrimination index* is known as the point-biserial correlation between an item and the test performance. It was not recommended for its suspected lack of variance in criterion-referenced test scores (Popham & Husek, 1969). One classical discrimination index introduced by Haladyna (1974) uses the combined-samples point-biserial correlation (COMPBI) to correlate an item and test performances when the sample is composed of a full range of instructed and uninstructed students. As shown in Formula 1, the point-biserial correlation (r_{pb}) discloses that "the mean differences in test scores for persons getting the item right and persons getting the item wrong are instrumental in determining the size of the coefficient" (Haladyna & Roid, 1981, p. 41)

$$r_{pb} = \frac{(M_p - M_q)\sqrt{pq}}{s_x} \tag{1}$$

Where

r_{pb} is the point-biserial correlation between item and test scores;

M_p is the mean test score of students who got the item correct;

M_q is the mean test score of students who got the item incorrect;

p is the proportion of examinees who got the item correct;

q is the proportion of examinees who got the item incorrect; and

s_x is the standard deviation of the test scores.

The magnitude of r_{pb} is a function of $M_p - M_q$. With combined samples, r_{pb} distinguishes better between instructed students with a tendency to perform successfully on items and uninstructed students with a tendency to perform unsuccessfully on items. However, it is not able to discriminate well when instructed students alone are examined (Haladyna & Roid, 1981).

Models based on *item response theory* (IRT) can provide sample-invariant estimates of item parameters. The Differential Item Functioning (DIF) procedure was most commonly used to detect item bias (Yu, Lei & Suen, 2006; Miller & Linn, 1988; Pham, 2009; Muthén, 1988a; Phillips & Mehrens, 1987; Muthén, 1989; Muthén, 1987; Muthén, Kao & Burstein, 1991; Shannon & Cliver, 1987). In an IRT model, if the probability for an examinee to give a correct answer to an item at a certain ability level, P(θ), is plotted as a function of ability, the result would be a smooth S-shaped curve, called item characteristic curve (ICC). The basic IRT models include the one-parameter logistic model, the two-parameter logistic model, and the three-parameter logistic model. In a three-parameter logistic IRT model (Birnbaum 1968; Hambleton, Swaminathan & Rogers, 1991):

$$P_i(\theta) = c_i + (1 - c_i)\{1 + \exp[-1.7a_i(\theta - b_i)]\}^{-1}$$

$$\text{or } P_i(\theta) = c_i + (1 - c_i)[1 + e^{-Da_i(\theta - b_i)}]^{-1}, \tag{2}$$

where

$P_i(\theta)$ is the probability that a randomly chosen examinee with ability θ answers item *i* correctly;

a_i is the discrimination power or discrimination parameter of item *i*;

b_i is the difficulty parameter of item *i*;

c_i is the lower asymptote of item *i;* it is called the *pseudo-chance-level* parameter;

D = 1.7 is a scaling factor introduced to make the logistic function as close as possible to the normal ogive function;

e is the base of the natural logarithm, the value of which is 2.718 (correct to three decimals); and

$P_i(\theta)$ is an S-shaped curve with values between 0 and 1 over the ability scale.

The *b* parameter for an item is the point on the ability scale (x-axis) where the probability of giving a correct answer to the item is 0.5. Therefore, the location of *b* indicates the difficulty of an item: the higher the *b*-point is on the ability (θ) scale, the more difficult the item is; and vice versa. The *a* parameter is proportional to the slope of the ICC at the point *b* on the ability scale. The guessing parameter, *c*, provides a possible nonzero lower asymptote for the ICC, representing the probability of examinees with very low ability “answering” (i.e., guessing) the item correctly. When guessing, the *c* parameter, is not allowed in the model, the lower asymptote of the ICC will be zero, and a two-parameter logistic model is formed:

$$P_i(\theta) = [1 + e^{-Da_i(\theta - b_i)}]^{-1} \tag{3}$$

Further, when the discrimination parameter, *a*, is constrained zero across all the items, it becomes a one-parameter logistic model:

$$P_i(\theta) = [1 + e^{(\theta - b_i)}]^{-1} \tag{4}$$

The IRT-based DIF methods use the estimate of latent ability (θ) rather than the observed score as matching variable. Item parameters are estimated separately for the reference and focal groups. After placing these estimates on the same scale, differences between the item parameters for the two groups can be compared (Clauser & Mazor, 1998). If the parameters are identical for the groups, the two ICCs overlap and there is no DIF present for the item (See Figure 2).

Figure 2. ICCs with No DIF

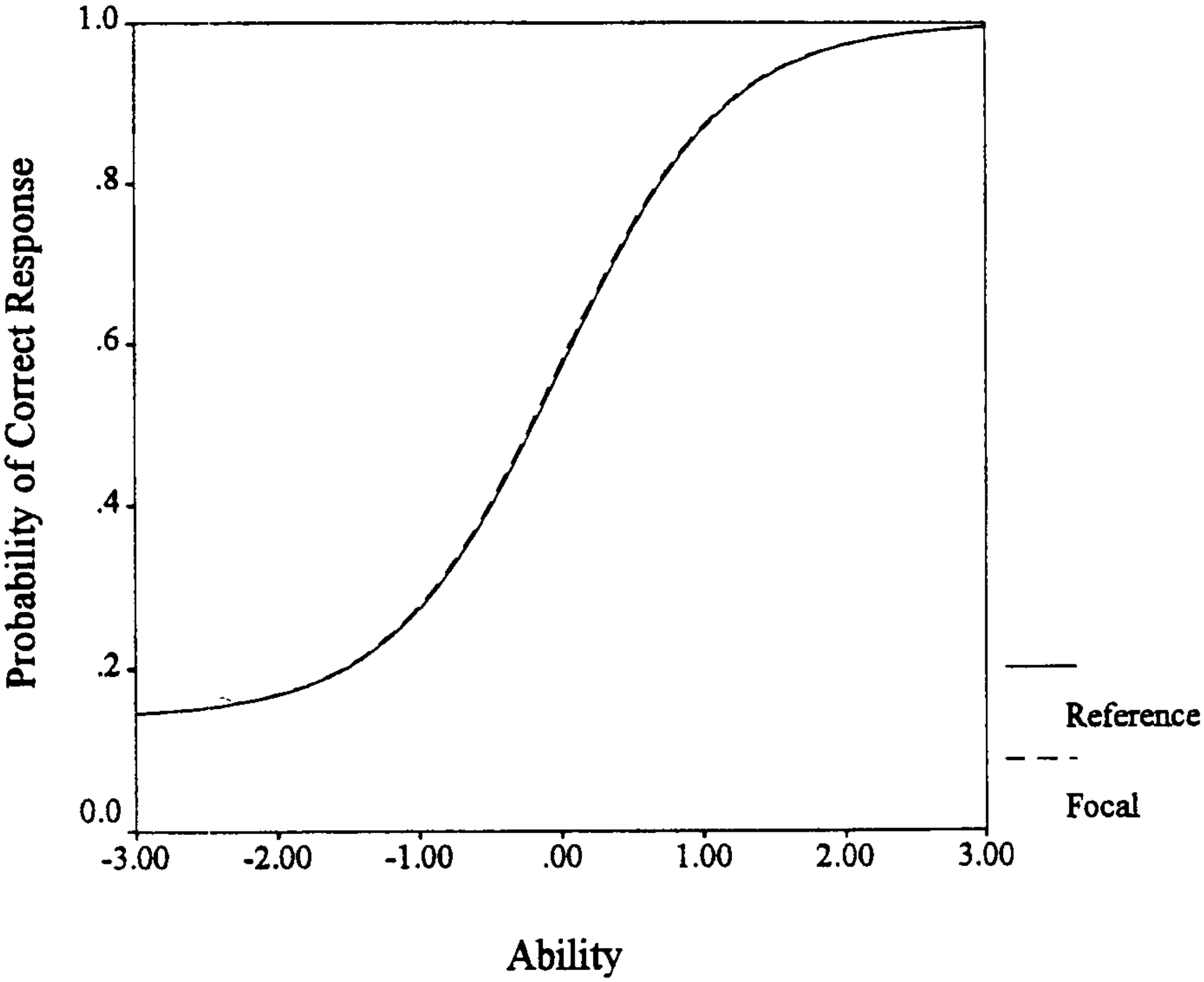


Figure 2. These are the ICCs of the focal and the reference groups for an item that displays no DIF. Taken from "Using statistical procedures to identify differentially functioning test items," by B. E. Clauser and K. M. Mazor, 1998, *Educational Measurement: Issue and Practice, 17,* p. 32.

If, instead of overlapping, there is an area between the ICCs for the two groups, DIF for that item is present. When items differ across groups only in terms of difficulty, uniform DIF is present and the two ICCs do not cross at any level of ability (See Figure 3).

Figure 3. ICCs with Uniform DIF

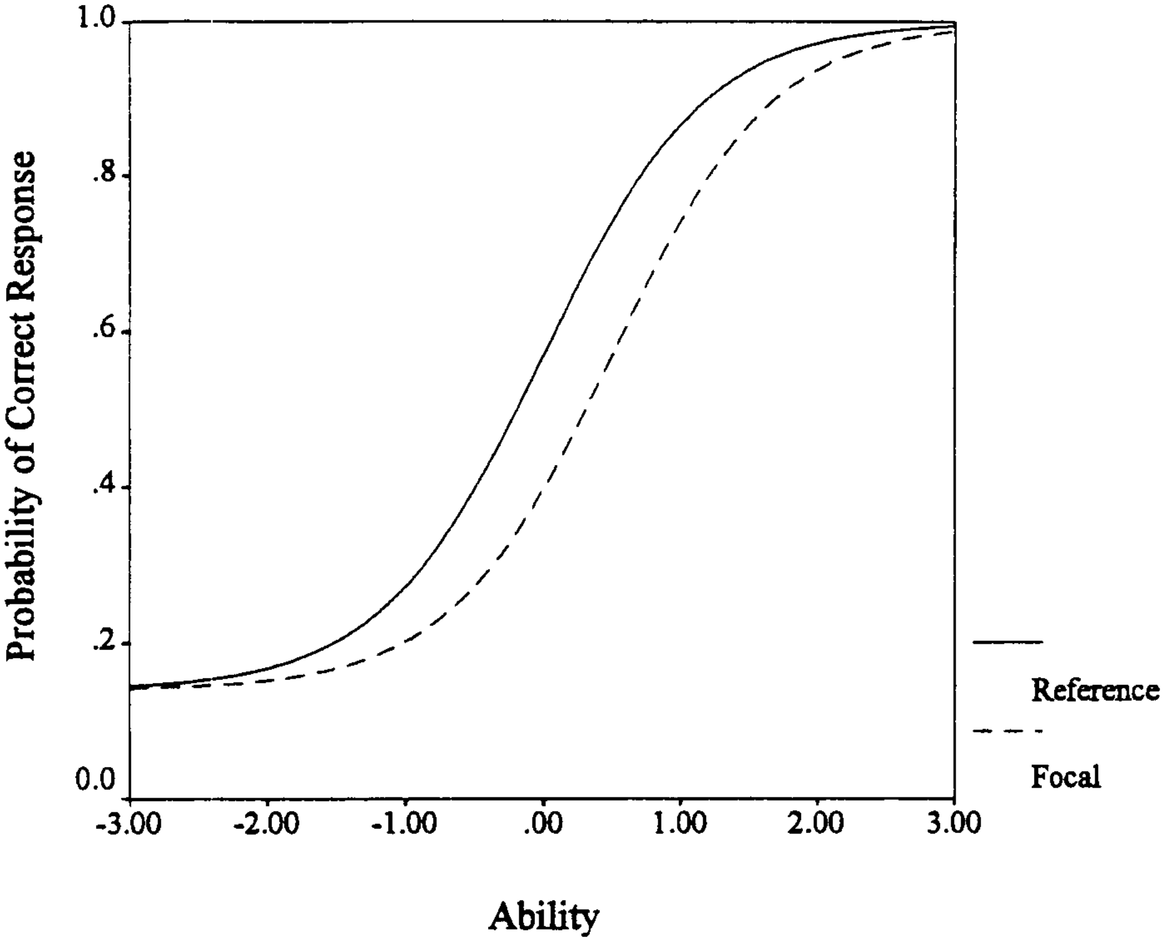


Figure 3. ICCs for the focal and reference groups for an item that displays uniform DIF. Taken from “Using statistical procedures to identify differentially functioning test items,” by B. E. Clauser and K. M. Mazor, 1998, *Educational Measurement: Issue and Practice, 17*, p. 33.

When items differ across groups in terms of not only difficulty but also discrimination (*a*), and/or sometimes pseudo-guessing (*c*), non-uniform DIF is present and the two ICCs cross at a certain point of ability level (See Figure 4).

Figure 4. ICCs with Non-uniform DIF

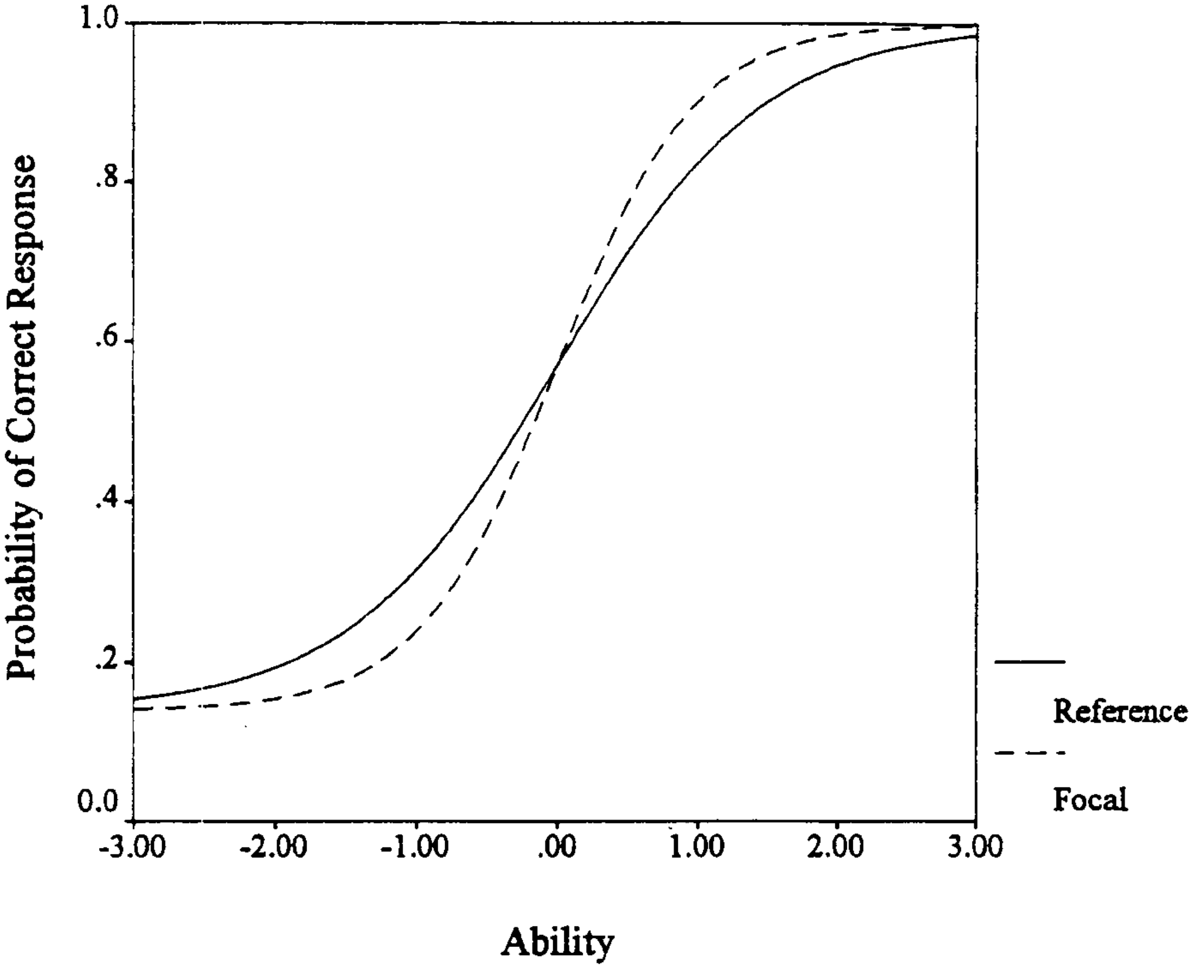


Figure 4. ICCs for the focal and reference groups for an item that displays non-uniform DIF. This figure shows that the item favors the reference group for examinees who have lower ability but favors the focal groups for the examinees who have higher ability. Taken from “Using statistical procedures to identify differentially functioning test items,” by B. E. Clauser and K. M. Mazor, 1998, Educational Measurement: Issue and Practice, 17, p. 34.

Muthén, Kao, and Burstein (1991) remarked that the standard IRT technique has an assumption that instruction increases the item performance through an increase in the latent trait (i.e., student’s ability: θ) level. Therefore, the item-trait relationship remains the same. They believed that this assumption was often too strong when the groups of students had very different content coverage. Muthén’s (1987, 1988b, 1989) contribution to the use of the IRT model is that he and his colleagues proposed an extended IRT-based detection technique of assessing item bias by generalizing the traditional IRT modeling to allow for item-specific variation in measurement relations across students with varying opportunity-to-learn. The traditional race or gender information, which is used to form the groups of students, puts a person in a group. The result of this grouping is constant over the items. However, the information of OTL is item specific (Muthén, 1988a, 1988b; Muthén, Kao & Burstein, 1991; Muthén, 1989). In order to take into account the instructional heterogeneity (i.e., item-specific variation in group membership), Muthén and colleagues created an OTL dummy variable for each item j, $z_j = 1$ represents OTL and $z_j = 0$ represents no OTL. Then the linear regression model is:

$$\eta = \gamma_X^{'} X + \gamma_Z^{'} Z + \zeta \ , \tag{5}$$

where x and z are vectors of variables and ζ is a normally distributed residual with zero mean, variance ψ, and where ζ is independent of x and z (Muthén, Kao & Burstein, 1991). To express the direct influence of the z variable on the item by using a latent response variable formulation, the following can be used:

$$y_j = 0, \text{if } y_j^* < \tau_j \text{or } y_j = 1, \text{ if otherwise,} \tag{6}$$

where τ_j is a threshold parameter defined on the continuous latent response variable y_j^*,

$$y_j^* = \lambda_j \eta + \beta_j z_j + \varepsilon_j \, . \tag{7}$$

The extended IRT model proposed by Muthén and his colleagues is also known as the multiple-indicator multiple-cause (MIMIC) model. The basic MIMIC model for instructional sensitivity detection or DIF testing is:

$$y_i * \ = a_i \theta + \beta_i G + \varepsilon_i \ , \tag{8}$$

where

$y_i *$ is the continuous response process that underlines a discrete y_i ;

a_i is the discrimination parameter;

β_i is the regression coefficient showing the group difference in the threshold; and

ε_i is the unique factor (error).

Figure 5 graphically depicts the basic MIMIC model.

Figure 5. A Basic MIMIC Model for DIF Testing

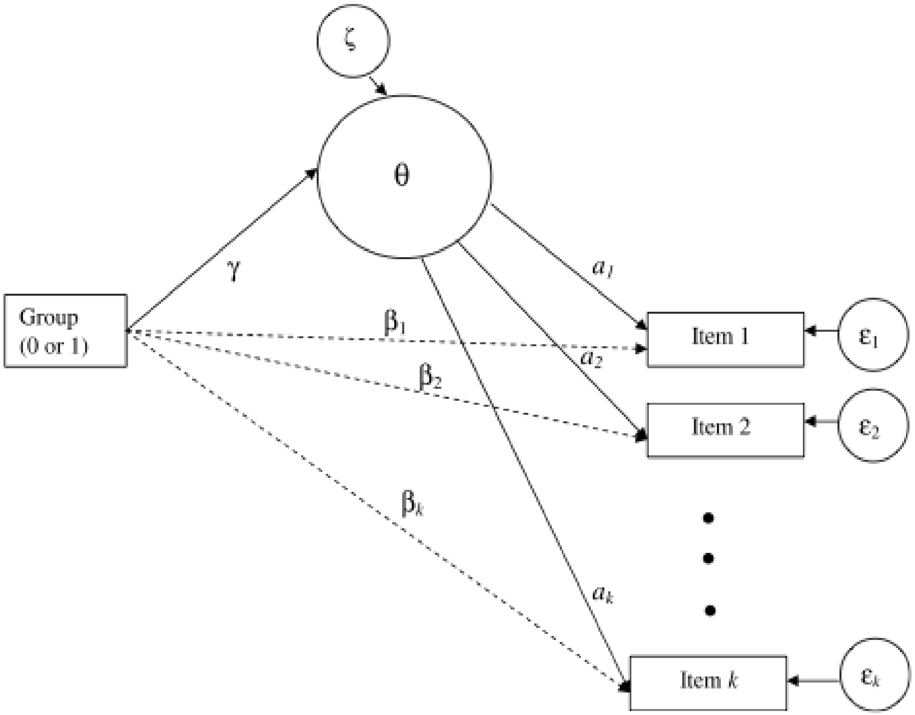


Figure 5. γ is the regression coefficient showing mean difference on the latent variable θ; β_i is the regression coefficient showing group difference in the threshold for item i; a_i is the loading of item *i* on the latent variable θ; ε_i is the measurement error for item *i*, and ξ= residual for θ. Taken from “Evaluation of MIMIC-Model methods for DIF testing with comparison to two-group analysis,” by C. M. Woods, 2009, *Multivariate Behavioral Research, 44*, p. 5.

The MIMIC model for testing non-uniform instructional sensitivity (or DIF, in general) is:

$$y_i * \ = a_i\theta + \beta_i G + \omega_i \theta G + \varepsilon_i, \qquad (9)$$

where θG is the interaction between the grouping variable (G) and the latent variable (θ), and ω_i is the non-uniform DIF effect. Figure 6 delineates the MIMIC model for testing non-uniform DIF.

Figure 6. A MIMIC Model for Testing Non-uniform DIF

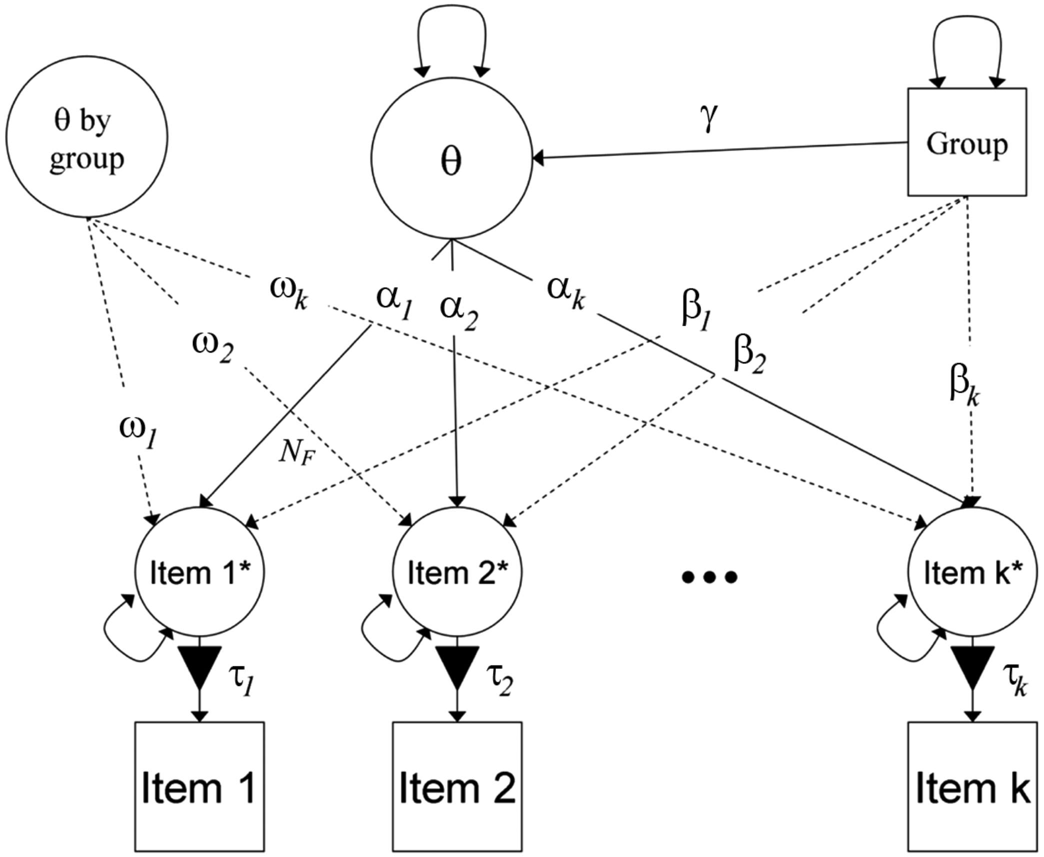


Figure 6. γ = mean difference on the latent variable, θ; item *i* = 1, 2, . . ., *k*; a_i = discrimination; ω_i = non-uniform DIF effect; τ_i = threshold; β_i = group difference in the threshold. Taken from "Testing for nonuniform differential item functioning with multiple indicator multiple cause models," by C. M. and Woods K. J. Grimm, 2011, *Applied Psychological Measurement*, *35*(5), p. 341.

The *Bayesian* technique (Helmstadter, 1974), which was rarely mentioned in the scholarship of instructional sensitivity, has three indices: "(a) B1, the probability that the student has knowledge, given that the student gets the item right, (b) B2, the probability that the student does *not* have knowledge, given that the student gets the item wrong, and (c), B3, the probability of making a right decision, due either to mastery or non-mastery" (Haladyna & Roid, 1981, p. 42):

$$B1 = \frac{(POSTDIFF)(COMDIFF)}{(POSTDIFF)(COMDIFF) + (PREDIFF)(1 - COMDIFF)} \tag{10}$$

$$B2 = \frac{(1 - PREDIFF)(1 - COMDIFF)}{(1 - PREDIFF)(1 - COMDIFF) + (1 - POSTDIFF)(COMDIFF)} \tag{11}$$

$$B3 = (POSTDIFF - COMDIFF) + (1 - PREDIFF) + (COMDIFF - POSTDIFF) \tag{12}$$

where

B1, B2, and B3 are the three Bayesian indices, and PREDIFF is the pre-instruction sample difficulty, POSTDIFF is the post-instruction sample difficulty, and COMDIFF is the mean of PREDIFF + POSTDIFF (i.e., the combined-samples' difficulty). The Bayesian indices were not recommended by previous research because these indices are not only unstable but also require complex computation. For example, high pretest and posttest difficulties can easily influence B1, while low pretest and posttest difficulties influence B2 easily; and B3 is influenced by both ceiling and floor effects (Haladyna & Roid, 1981).

This study used Mantel-Haenszel statistic and logistic regression to identify instructionally sensitive items (i.e., items that function differently across instructed and uninstructed groups at the same proficiency level). As Clauser and Mazor (1998) summarized, the MH method compares the likelihood of success on the item for members of the two groups that are matched on ability or proficiency. The ratio of these likelihoods is used as the index to identify instructional sensitivity. This method has been very popular since it was first recommended by Holland and Thayer (1988, cited from Clauser & Mazor 1998) because of its statistical power and easy computational procedures. To implement the MH method, the examinees are first divided into levels based on proficiency. Usually the total score is used as the matching criterion. Then a 2 X 2 contingency table is used to arrange data for DIF detection, one item at a time. For a binary item scored 0, 1, data from the two groups of respondents can be arranged in a 2 X 2 table as below (Table 1):

Table 1. Mantel-Haenszel 2 X 2 Contingency Table for Binary Item *i*

	Performance on Item *i*		
Group	1	0	Total
Reference (Uninstructed)	a_i	b_i	$N_{ri} = a_i + b_i$
Focal (Instructed)	c_i	d_i	$N_{fi} = c_i + d_i$
Total	$N_1 = a_i + c_i$	$N_0 = b_i + d_i$	$N_i = a_i + b_i + c_i + d_i$

For each score interval *i*, α_i is calculated as (Angoff, 1993; See Peyton, 2000):

$$\alpha_i = \frac{p_{ri}}{q_{ri}} \bigg/ \frac{p_{fi}}{q_{fi}} = \frac{a_i}{b_i} \bigg/ \frac{c_i}{d_i} = \frac{a_i d_i}{b_i c_i}, \quad i = 1, 2, \ldots, k \qquad (13)$$

where

i is the number of levels of the total score (i.e., the matching criterion). For example, a scale contains 20 binary items (scored 0, 1), there would be 21 test score levels, ranging from 0 to 20;

p_{ri} is the proportion of the uninstructed group in score interval *i* who answered the item correctly;

$q_{ri} = 1 – p_{ri}$ is the proportion of the uninstructed group in score interval *i* who failed to answer the item correctly;

P_{fi} is the proportion of the instructed group who answered the item correctly,

$q_{fi} = 1 – p_{fi}$ is the proportion of the instructed group who failed to answer the item correctly; and

α_i is the ratio of the odds (p/q) that the uninstructed group succeeded on the item to the odds that the instructed group succeeded on the item. α_i ranges from 0 to ∞, with the value of 1.0 indicating no DIF, values less than 1.0 indicating that the item is in favor of the instructed group and values greater than 1.0 indicating that the item is in favor of the uninstructed group, after students from two groups have been matched on their proficiency.

An item presents uniform instructional sensitivity if the odds of correctly answering the item at a given score level *i* is different for the two groups at a certain matching level of proficiency or total test scores.

Logistic regression is another popular DIF procedure that is regarded as a link between the contingency table methods (e.g., Mantel-Haenszel) and the IRT methods (Clauser & Mazor, 1998; Peyton, 2000). According to Swaminathan and Rogers (1990), the Mantel-Haenszel procedure can be considered as a special case of the logistic regression model where the ability variable is discrete and no interaction term between the grouping variable and ability is estimated. On the other hand, in the logistic regression model, the ability variable is assumed to be continuous and an interaction term between the grouping variable and ability is included.

Therefore, logistic regression procedure is capable of identifying both uniform and non-uniform instructional sensitivity. The basic model of logistic regression is

$$P(U=1)=\frac{e^{z}}{1+e^{z}} \text{ or } P(U=1|x)=\frac{e^{f(x)}}{1+e^{f(x)}}, \tag{14}$$

where $P(U=1|x)$ is the conditional probability of responding to the item correctly given x (vector of independent variable); and *f*(x) is the function defining the linear combination of the independent variables. In DIF analysis, the dependent variable is the item score; the independent variables are the grouping variable (G), the examinee's ability (θ) or the observed total test score that is used as the matching criterion, and the interaction between the examinee's membership and his/her ability (θ*G).

When

$$\text{Z (or } f(x)) = \beta_0 + \beta_1\theta + \beta_2 G, \tag{15}$$

β_2 is the measure of uniform DIF, where θ is the matching ability and G is the grouping variable.

When

$$Z = \beta_0 + \beta_1\theta + \beta_2 G + \beta_3(\theta * G), \tag{16}$$

the added term represents the interaction between ability and group and, therefore, both the uniform DIF and non-uniform DIF can be tested simultaneously by comparing the fit of the final model (including β_0, $\beta_1\theta$, β_2G and $\beta_3\theta*G$) to that of the simple model (including only β_0 and $\beta_1\theta$) (Clauser & Mazor, 1998).

Studies using both simulated and real data showed that logistic regression procedure produced results similar to the Mantel-Haenszel statistic when testing for uniform DIF but was superior to the Mantel-Haenszel statistic when identifying non-uniform DIF (Rogers & Swaminathan, 1993). However, a recent study showed that only the standard MH procedure was less powerful in DIF detection than the LR procedure; the modified MH procedure showed similar power to the LR procedure (Hidalgo & López-Pina, 2004). The modified MH procedure improves detection rates of non-uniform DIF over the standard MH procedure without increasing the Type I error rate (Mazor et al., 1994). Both the MH and the LR procedures were used in this study so that the results produced from both methods could be compared.

Instructional Sensitivity and Accountability Tests

Accountability tests have become increasingly important (Popham, 2007b), and "highly qualified teachers" are an important accountability component of NCLB (Simpson, 2004). One category of defining teacher quality, according to Goe (2007), is based on the outcome – teacher effectiveness. To better understand it, teacher effectiveness can be reflected by effective instruction. Thus, sensitivity to effective instruction becomes an imperative index that helps to determine whether an achievement test is accountable or not in measuring how well students have been taught. Since a fundamental function of educational tests is to make inferences from test results, Popham (2010a) discerned the essential difference between two types of test-based inference: when the test scores only allow people to ascertain what knowledge and skills the students possess, they have *achievement test inference*; when the test scores allow people to tell how well the students have been taught the tested content, they have *accountability test inference*.

Most of today's accountability tests fail to hit the target of providing an accurate estimate of how well a group of students has been taught (Popham, 2010a). These tests measure what students bring to school, but not what they learn from school (Popham, 2010a). With the inaccurate or even wrong test-based evidence, the presence of instructional improvement (or the opposite) cannot be determined. Without a doubt, an accountability test must be instructionally sensitive to do an adequate job of measuring instructional sensitivity. Popham (2003a, Popham et al., 2005) summarized crucial attributes of an instructionally sensitive accountability test:

(1) *clear descriptions of what is assessed*, meaning sufficiently clear descriptions of the content standards that can help educators and teachers easily and quickly figure out the test's true assessment targets;

(2) *a modest number of curricular aims assessed*, meaning the number of significant curricular aims must be reasonably manageable for teachers to focus on in their instruction;

(3) *instructionally informative results*, meaning an instructionally sensitive accountability test must report students' performances for teachers and parents to identify which content standards the student has or has not mastered;

(4) *gauging sensitivity*, meaning school administrators need to appraise the state's NCLB tests based on the previous three requirements, and distinguish immediately whether their state's NCLB tests are instructionally sensitive or not.

Summary

A review of literature with a focus on instructional sensitivity provides a set of key elements discussed in relevant studies. They are (a) comparison and contrast of instructional sensitivity, instructional validity, and curricular validity; (b) relation of instructional sensitivity to standards-based assessments; (c) definition and measure of opportunity to learn (OTL); (d) empirical and judgmental strategies of detecting instructional sensitivity; and (e) relationship between instructional sensitivity and accountability tests. Based on the importance of instructional sensitivity in high-stakes achievement tests reviewed in the literature, this study tested the following hypotheses:

1. The difference in difficulty between the group receiving instruction and the group not receiving instruction should be greater for items teachers judge to be instructionally sensitive than for items teachers judge to not be instructionally sensitive.
2. For items judged to be instructionally sensitive, at a fixed ability level, students who were taught the content tested should have higher probability of responding correctly to the item than those who were not taught.
3. The items identified as functioning most differently for the instructed and not instructed groups should, to a great degree, match the items detected as sensitive by the curriculum experts.

CHAPTER 3

METHODS

This chapter contains three major sections: (1) a review of purpose and research questions; (2) a description of the data source used in the present study; and (3) a description of analytical procedures employed to detect instructional sensitivity of test items.

Purpose Overview and Research Questions

The purpose of this study is to examine the relationship between students' performance on high-stakes achievement tests and their instructional experiences. The Mantel-Haenszel tests and logistic regression procedures were used to detect instructional sensitivity of the test items. Thirteen mathematics teachers, who had experience teaching seventh grade math, were recruited to review 35 multiple-choice mathematics items on the Kansas Interim Assessment and rate the items as to their sensitivity to good instruction. Twelve of the thirteen curriculum experts were currently teaching in Kansas middle schools, and one of them was a graduate student majoring in Curriculum and Teaching with a specialization in Mathematics Education at the University of Kansas. Results from empirical detection and judgmental detection of item sensitivity were compared to see how much they agreed or disagreed with each other for the sake of credibility and reliability.

Data from 2010-2011 Kansas Interim Assessment in seventh grade mathematics (Test Window Two) were used in this study. Responses from 3,446 students to 35 multiple-choice items were analyzed. The following questions were addressed:

1. To what extent are the items in the state testing program sensitive to instruction?

2. Are item performance differences related to differences in curricular content covered in instruction?
3. How does the instruction in content tested on the state testing program influence students' performance?
4. To what extent do the instructional sensitivity judgments made by curriculum experts agree with the results of empirical methods?

Data Source and Participants for Empirical Methods

The data for this study were collected from the Kansas Interim Assessment in Mathematics in 2010-2011. According to the manual for the assessment, its purpose is to provide estimates of student achievement in regard to tested indicators for mathematics at three time points prior to the summative assessment. The interim assessment program is designed to measure the same specific indicators within the Kansas Curricular Standards that are measured by the summative state assessments, which are administered every spring.

The interim assessment has three testing windows. Window One (i.e., the Fall One Window) was open from September 15 to October 29 in 2010; Window Two (i.e., the Fall Two Window) was open from October 30 to December 31 in 2010; Window Three (i.e., the Winter Window) was open from January 1 to February 28 in 2011. The assessment was given in third through eighth grades. However, for this particular study, only test results from the seventh graders were examined. As for the sample size, 3,503 seventh graders took the test in Window One; 5,678 seventh graders took the test in Window Two; and 4,577 seventh graders took the test in Window Three. Only data from Window Two were used in this study for two reasons. First, Window Two has the biggest sample size. Second, only in Window Two was part of the tested content taught while the other part was not. Therefore, it is meaningful to examine items' sensitivity to instruction by using Window Two data. In Window One, nearly no content tested was taught, and in Window Three, nearly all the content tested was taught. Thus, no comparison in Windows One and Three could be made between the examinees in terms of the impact of their instructional experiences on performance.

The interim assessment is a multi-stage adaptive design. Each test is composed of three parts that were administered in one testing session. The three sections in each assessment session are called parts or testlets. Students received more or less challenging testlets based on their responses to previous testlets during the same test window. The testlet was selected from the testlet pool. Altogether there were about 90 items available to form one testlet for each grade. All the test items were multiple choice questions. The suggested session length was 45 to 60 minutes (see Appendix A for more information).

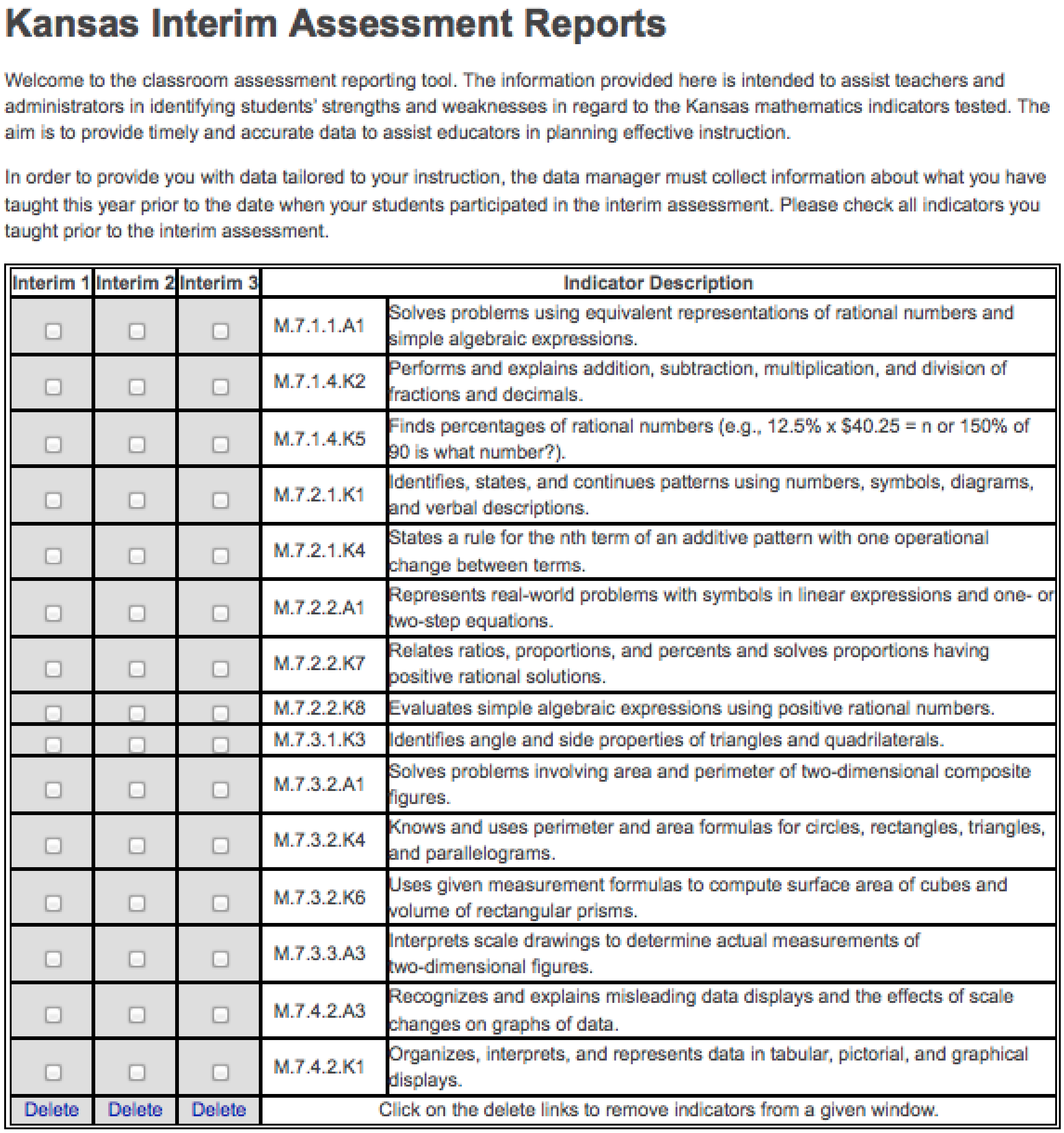
Four mathematical standards were measured in each test session. They are *Numbers and Computation*, *Algebra*, *Geometry*, and *Data* (see Appendix B for more information). For the tests given to the seventh graders, there were 15 indicators tested in each test session. Under each indicator, there were two to four test questions. Appendix C presents all the indicators within the state's curricular standards for seventh grade, and the items, used in Test Window Two, under the above 15 tested indicators are listed. The 15 tested indicators are in bold in Appendix C.

Opportunity to Learn Measures

To get instructional information for the opportunity to learn (OTL) variable, teachers were asked to login online and enter instructed indicators after administering the interim assessment. A list of tested indicators for the teacher's grade and subject were presented after his/her login. The teacher was expected to click the check box beside any indicators that were taught prior to the interim assessment. The teacher was assumed to have considered whether the instruction provided prior to an assessment for each indicator was adequate or not for students to be able to successfully answer all potential items assessing that indicator. Figure 7 shows how the screen actually appeared to the teachers after their login. When at least five indicators had been taught prior to an interim assessment, then the teacher's report contained scores reflecting student performance on all tested indicators, as well as scores reflecting student performance on only those indicators taught prior to the test. If a teacher chose not to submit information about what was taught prior to the interim assessment, then that teacher's report contained only scores reflecting student performance on all tested indicators; no specific score information was available to reflect student performance on the indicators taught prior to the test.

The grouping variable (i.e., the OTL variable) was created based on this information provided by teachers. Therefore, each student was assigned either to the instructed group or to the uninstructed based on his or her teacher's "feedback" on each specific item. In this study for instructional sensitivity analysis using empirical item-detection techniques, the uninstructed group was treated as the reference group (coded as 0), and the instructed group was treated as the focal group (coded as 1).

Figure 7. A Screen Capture of Kansas Mathematics Interim Assessment Reports



Kansas Interim Assessment Reports

Welcome to the classroom assessment reporting tool. The information provided here is intended to assist teachers and administrators in identifying students' strengths and weaknesses in regard to the Kansas mathematics indicators tested. The aim is to provide timely and accurate data to assist educators in planning effective instruction.

In order to provide you with data tailored to your instruction, the data manager must collect information about what you have taught this year prior to the date when your students participated in the interim assessment. Please check all indicators you taught prior to the interim assessment.

Interim 1	Interim 2	Interim 3	Indicator Description	
□	□	□	M.7.1.1.A1	Solves problems using equivalent representations of rational numbers and simple algebraic expressions.
□	□	□	M.7.1.4.K2	Performs and explains addition, subtraction, multiplication, and division of fractions and decimals.
□	□	□	M.7.1.4.K5	Finds percentages of rational numbers (e.g., 12.5% x $40.25 = n or 150% of 90 is what number?).
□	□	□	M.7.2.1.K1	Identifies, states, and continues patterns using numbers, symbols, diagrams, and verbal descriptions.
□	□	□	M.7.2.1.K4	States a rule for the nth term of an additive pattern with one operational change between terms.
□	□	□	M.7.2.2.A1	Represents real-world problems with symbols in linear expressions and one- or two-step equations.
□	□	□	M.7.2.2.K7	Relates ratios, proportions, and percents and solves proportions having positive rational solutions.
□	□	□	M.7.2.2.K8	Evaluates simple algebraic expressions using positive rational numbers.
□	□	□	M.7.3.1.K3	Identifies angle and side properties of triangles and quadrilaterals.
□	□	□	M.7.3.2.A1	Solves problems involving area and perimeter of two-dimensional composite figures.
□	□	□	M.7.3.2.K4	Knows and uses perimeter and area formulas for circles, rectangles, triangles, and parallelograms.
□	□	□	M.7.3.2.K6	Uses given measurement formulas to compute surface area of cubes and volume of rectangular prisms.
□	□	□	M.7.3.3.A3	Interprets scale drawings to determine actual measurements of two-dimensional figures.
□	□	□	M.7.4.2.A3	Recognizes and explains misleading data displays and the effects of scale changes on graphs of data.
□	□	□	M.7.4.2.K1	Organizes, interprets, and represents data in tabular, pictorial, and graphical displays.
Delete	Delete	Delete	Click on the delete links to remove indicators from a given window.	

In the dataset used for this study, the 3,446 students were nested within 142 teachers. The minimum number of students nested within each teacher was one; the maximum number of students nested within each teacher was 83; and the most frequent number of students associated with each teacher was one (i.e., fourteen out of the 142 teachers only had one student nested within each of them). Only one teacher had 83 nested students. Table 2 below presents the frequency of the number of students nested within each teacher.

Table 2. Frequencies of Number of Students Nested per Teacher

No. of Students per Teacher	83	76	75	67	66	65	64	62	60	59	56	50	48	47	44	43	42
Frequency	1	1	1	2	1	1	1	1	3	2	3	2	2	1	2	3	4

(Table 2 continued)

No. of Students per Teacher	39	37	36	35	34	33	29	28	27	26	25	24	23	22	21	20	19
Frequency	1	4	3	2	1	4	2	3	3	2	3	4	3	3	2	4	4

(Table 2 continued)

No. of Students per Teacher	18	17	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
Frequency	1	2	1	4	1	4	6	1	5	1	2	3	4	2	6	6	14

The same information in Table 2 is also presented by a line graph with students per teacher on the x-axis and frequency on the y-axis (See Figure 8 below).

Figure 8. Number of Student Nested within per Teacher

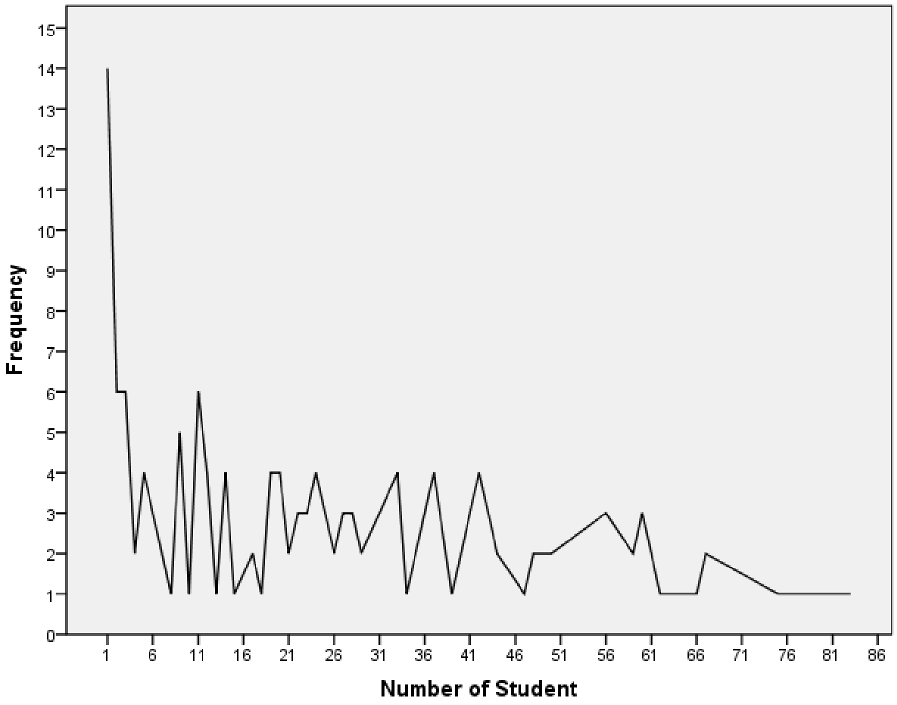


Figure 8. The maximum number of students per teacher was 83; the minimum number of students per teacher was 1; teachers who had one student nested were counted as most frequent (n = 14).

Data Analysis for Empirical Methods

The Mantel-Haenszel tests and logistic regression techniques in testing for instructional sensitivity were used in this study to detect items that functioned differently when responded by students from different instructional groups. Because the grouping variable in this study is the students' instructional experience, the membership of each student may change across the items. For example, Student # 1 may belong to the instructed group for Item #1 but to the uninstructed group for Item # 2. This characteristic of the grouping variable made it impossible to analyze all the test items at one time, so each test item was analyzed one at a time. Therefore, the traditional item response theory (IRT) methods (IRT-based methods) for detecting DIF were not feasible. Neither would the multiple-indicator multiple-cause models work because they require all the test items to be included in one model to construct the baseline model (Muthén, Kao & Burstein, 1991). Further, “the drawback of IRT-based procedures is that they are sensitive to sample size . . . and the indices [*sic*] such as the area between item characteristic curves have no associated tests of significance” (Swaminathan & Rogers, 1990, p. 362). Also, the IRT methods are usually complex and expensive (Mazor et al., 1994). On the contrary, the major advantage of MH procedure is that it is easy to implement and has an associated test of significance. As for the LR model, it is attractive because it takes into account the continuous nature of the ability scale and is powerful in identifying both uniform and non-uniform DIF (Swaminathan & Rogers, 1990; Zumbo, 1999). It also has associated tests of significance. Therefore, Mantel-Haenszel tests and logistic regression are more appropriate approaches in this study.

The MH statistics test the H_0 against the alternative (Clauser & Mazor, 1998; Hidalgo & López-Pina, 2004):

$$H_1 : \frac{P_{ri}}{Q_{ri}} = \alpha \frac{P_{fi}}{Q_{fi}} \quad i = 1, 2, . . . ,k \, , \tag{17}$$

Where $\alpha \neq 1$ and i is the number of levels of total score or matching criterion (refer to Equation 13); the α is estimated by:

$$\hat{\alpha}_{MH} = \frac{\sum_{i=1}^{k} a_i d_i / N_i}{\sum_{i=1}^{k} b_i c_i / N_i} \text{ (refer to Table 1 for denotes) .} \tag{18}$$

The statistics for detecting DIF in an item takes the form:

$$MH\chi^2 = \frac{\left(\left|\sum_i a_i - \sum_i E(A_i)\right| - \frac{1}{2}\right)^2}{\sum_i \mathrm{var}(a_i)}, \tag{19}$$

where

$$E(A_i) = (N_{Ri} N_{1.i}) / N_{..i} \text{ , and} \tag{20}$$

$$\mathrm{var}(a_i) = \frac{N_{ri} N_{fi} N_1 N_0}{N_i^2 (N_i - 1)}. \tag{21}$$

The test statistics is compared to a chi-square distribution with one degree of freedom.

The Mantel-Haenszel tests are one of the most popular methods evaluating DIF but have been criticized for two limitations: a) they do not have latent variables in their models to adjust for measurement errors (Woods & Grimm, 2011); and b) they are not powerful in detecting non-uniform DIF (Woods & Grimm, 2011; Peyton, 2000; Swaminathan & Rogers, 1990). Often, the items that the MH procedure is likely to miss are non-uniform DIF items of medium difficulty (Mazor et al., 1994). However, the first limitation of not having latent variables included in the models was solved in this study because students' proficiency measures (i.e., the latent θ), instead of their total test scores, were used for the data analyses. Further, Mazor, Clauser, and Hambleton (1994) proposed a modification of the MH statistics that improves non-uniform DIF detection. In their simulation study, a standard MH procedure was used first. Then, the examinees were split into two samples approximately at the middle of the test score distribution. Data of the low-performing sample and data from the high-performing sample were analyzed separately by using MH procedure. Compared to the total sample procedure, this variation improved detection rates of non-uniform DIF substantially without increasing the Type I error rate. Therefore, the second limitation can be taken care of as well.

The logistic regression procedure, which is powerful in detecting both uniform and non-uniform DIF, was also used in this study. While the usual approach to logistic regression would be to use raw scores as a continuous variable, due to their greater measurement precision, this study used the latent variable, θ, estimated by using the one-parameter item response model. Thus, θ, G and θ*G (refer to Equations 15 &16) were entered into the regression model successively (Hidalgo & López-Pina, 2004). To evaluate instructional sensitivity, the unique contribution of each successive model term (θ, G, θ*G) is statistically calculated. A statistically significant regression coefficient for G indicates uniform instructional sensitivity, and a statistically significant regression coefficient for θ*G indicates non-uniform instructional sensitivity; both statistics follow a χ^2 distribution with *df* = 1. The LR analysis also allows simultaneously testing for both uniform and non-uniform instructional sensitivity. To test this joint hypothesis, the χ^2 value of the null model (θ included only) is compared with the χ^2 value of the final model (θ, G, and θ*G included). This statistic follows a χ^2 distribution with *df* = 2.

The results obtained from the two techniques were compared and discussed. Further, the results were examined in the context of some measure of effect size for two reasons. First, with small sample sizes, the results may not be statistically significant but may be significant in effect size terms; and with large sample sizes, the results may be statistically significant, but the effect size may be small. Second, results from the Jodoin and Gierl (2001) study indicate that the inclusion of the effect size measure can substantially reduce Type I error rates when large sample sizes are used, although there is also a reduction in power. In the Mantel-Haenszel models, both α and Δ_{MH} are measures of effect size. The α is the ratio of the odds that a reference (or uninstructed) group gets an item correct to those for a matched focal (or instructed) group (Clauser & Mazor, 1998). If an item favors the uninstructed group, α falls between one and infinity; if an item favors the instructed group, α ranges from zero to one. When $\alpha = 1$, the item is insensitive. For the convenience of interpretation, logistic transformation was made by multiplying the α by -2.35 to produce the Δ_{MH} (Holland & Thayer, 1988; cited by Clauser & Mazor, 1998; Hidalgo & López-Pina, 2004):

$$\Delta_{\alpha(MH)} = -2.35\ln\left(\alpha_{MH}\right). \tag{22}$$

Thus, the distribution of the resulting values is symmetric around zero, and a negative value indicates that the item favors the instructed group while a positive value indicates the opposite; and a zero value indicates an absence of instructional sensitivity. The popular three-level DIF classification system used by Educational Testing Service (ETS) is as follows (proposed by Zwick and Ercikan, 1989; cited by Hidalgo & López-Pina, 2004):

- Type A items – negligible DIF: items with $\Delta_{\alpha(MH)} < |1|$;
- Type B items – moderate DIF: items with $|1| \leq \Delta_{\alpha(MH)} \leq |1.5|$, and MH test ($\mathrm{MH}\chi^2$) is statistically significant ($p < .05$);
- Type C items – large DIF: items with $\Delta_{\alpha(MH)} > |1.5|$, and MH test is statistically significant.

The combination of statistical significance and effect size helps to avoid identifying items that present practically unimportant but statistically significant DIF (Clauser and Mazor, 1998).

The logistic regression also provides a measure of effect size. The ΔR^2, a weighted lease squares effect size measure, is usually used to quantify the magnitude of uniform or non-uniform DIF when LR is applied to DIF detection (Hidalgo & López-Pina, 2004). The ΔR^2 used for non-uniform instructional sensitivity is the difference between the Nagelkerke R^2 value of the model in Step 3 and that of the model in Step 1:

$$\Delta R^2 = R^2(M3) - R^2(M1). \tag{23}$$

Guidelines suggested by Zumbo and Thomas (1977, cited by Hidalgo & López-Pina, 2004) are as follows:

- Type A items – negligible DIF: $\Delta R^2 < 0.13$;
- Type B items – moderate DIF: $0.13 \leq \Delta R^2 \leq 0.26$, and the two-*df* χ^2 test is statistically significant ($p < .05$);
- Type C items – large DIF: $\Delta R^2 > 0.26$, and the two-*df* χ^2 test is statistically significant.

Jodoin and Gierl (2001) proposed other DIF classification criteria based on the SIB (Simultaneous Item Bias Test)[5] effect size measure (Roussos & Stout, 1996):

- Type A items – negligible DIF: $\Delta R^2 < 0.035$;
- Type B items – moderate DIF: $0.035 \leq \Delta R^2 \leq 0.070$, and the null hypothesis is rejected;
- Type C items – large DIF: $\Delta R^2 > 0.070$, and the null hypothesis is rejected.

Jodoin and Gierl (2001) reported that only 6.8% of the items with DIF were identified as Type B items when the classification criteria proposed by Zumbo and Thomas (1997) was used; but 68.2% of the items were identified as Type B items when their own criteria were used (Hidalgo & López-Pina, 2004).

The two techniques in testing for DIF mentioned above were used to test the first two hypotheses:

1. The difference in difficulty between the group receiving instruction and the group not receiving instruction should be greater for items teachers judge to be instructionally sensitive than for items teachers judge to not be instructionally sensitive.

[5] SIB is an alternative statistical method for detecting DIF proposed by Shealy and Stout (1993). SIBTEST is intended to model multidimensional data; it can be used for unidimensional data as well (Gierl et al., 1999).

2. For items judged to be instructionally sensitive, at a fixed ability level, students who were taught the content tested should have higher probability of responding correctly to the item than those who were not taught.

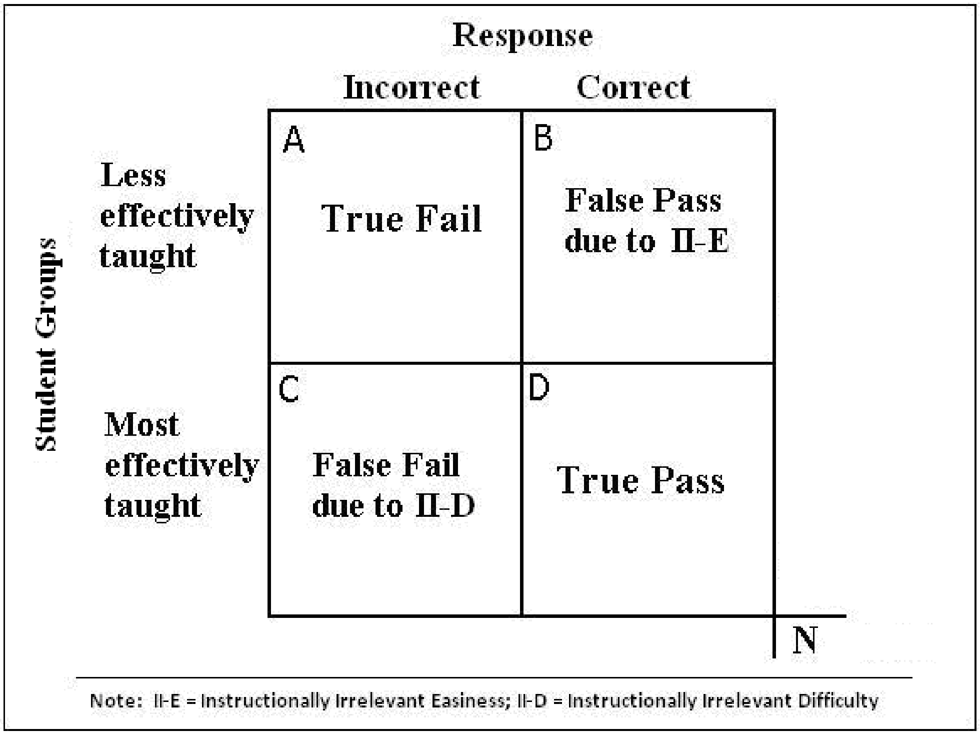
Judgmental Approach

In order to determine how the judgmental and empirical item-detection techniques agree or disagree with each other (i.e., H3: The items identified as functioning most differently for the instructed and not instructed groups should, to a great degree, match the items detected as sensitive by the curriculum experts.), some curriculum experts were recruited to review the interim assessment items and provide their judgment. Popham (2003b) suggested a small group of teachers, perhaps a half-dozen or so, should be sufficient. Snowball, chain, or network sampling was used to recruit thirteen mathematics teachers, who were currently teaching or recently taught seventh grade mathematics, from the Kansas middle schools and the University of Kansas. To be specific, the researcher located one or two key participants who easily met the criteria she had established for participation in the study. As she worked with early key participants, the researcher asked each participant to refer her to other participants.

The participants were instructed to review and rate 35 multiple-choice items from the 2010-2011 Kansas Interim Assessment in Mathematics for seventh graders. They provided their comments on the test items by filling out a form with predetermined questions (See Appendix D). It took about 1 hour to 90 minutes for the participant to review all the items and to report, in the form of written notes, his/her ratings on item sensitivity. The 60- to 90-minute item review took place on the Lawrence Campus of the University of Kansas. The entire review was composed of three parts. In the first part, the participants were presented the test items and instructed to rate the items based on an 11-point Likert scale of instructional sensitivity (i.e., the first form for item review, see p.2 of Appendix D). In the second part, the researcher withdrew the test items and presented the descriptions of the indicators within the state’s curricular standards for seventh grade mathematics. The participants were instructed to read the descriptions and provide their feedback on the clarity of the indicators (i.e., the second form for item review, see pp. 3-6 of Appendix D). In the third part, the researcher withdrew the second form, presented the test items again and the third form that repeated the description of the indicators (see pp. 7-8 of Appendix D), and instructed the participants to assign items under different indicators.

During the review, the researcher was present but did not interfere with the participant’s independent review and judgment. First, the researcher briefly introduced herself, gave some information about what instructional sensitivity is, presented an example of how a sensitive item or an insensitive item functions (see Figure 9), made it clear that the target test takers for the items to be reviewed were typical seventh graders, and then gave instructions for the steps of the review. Second, the participants consented to their own participation and signed the letter of confidentiality (See Appendix E). Then, the researcher monitored the three parts of the item review. The participants' notes were given identifying code numbers and stored securely while not in use. The notes will be preserved indefinitely; however, any labels which contain code numbers will be removed from them after two years. When reporting the judgments from the participants, all information that could be used to identify individuals will be omitted. The participants’ names were and will not be associated in any way with the information collected about them or with the research findings from this study. The researcher only collected and used their name as part of the informed consent form and the letter of confidentiality.

Figure 9. Logic of the Appraisal Applied to a Perfectly Sensitive Test Item



		Response	
		Incorrect	Correct
Student Groups	Less effectively taught	A True Fail	B False Pass due to II-E
	Most effectively taught	C False Fail due to II-D	D True Pass

N

Note: II-E = Instructionally Irrelevant Easiness; II-D = Instructionally Irrelevant Difficulty

Figure 9. The degree to which actual performance corresponds with expectation reflects the degree to which the item is sensitive to instruction. (A+D)/N serves as an adequate appraiser of instructional sensitivity. Taken from “Instructional sensitivity of accountability tests: Recent refinements in detecting insensitive items,” by S. C. Court, 2010, *CCSO’s National Conference on Student Assessment*, p. 5.

Teachers’ ratings on item sensitivity and indicator clarity and their assignments of items to indicators were analyzed. The intra-class correlation was used to evaluate raters’ agreement. Further, items flagged by the empirical techniques and items suggested by the curriculum experts as instructionally sensitive were compared. Possible reasons for the disagreement of the two methods were discussed. Then, implications helpful to improve the achievement measurement process were summarized. Finally, the results of this comparison can help the researcher to determine: a) Do instructionally sensitive items have common characteristics? b) Do items that test any specific mathematic content tend to be instructionally sensitive?

Summary

This chapter reviews the research methods used in this study. For the empirical methods, the data source is described and the research design is specified. For the judgmental approach, the procedures of data collection and item review are presented. This study has been approved by the Human Subjects Committee University of Kansas, Lawrence Campus (HSCL). It was also approved by the Lawrence school district office to conduct this research with school teachers.

CHAPTER 4

RESULTS

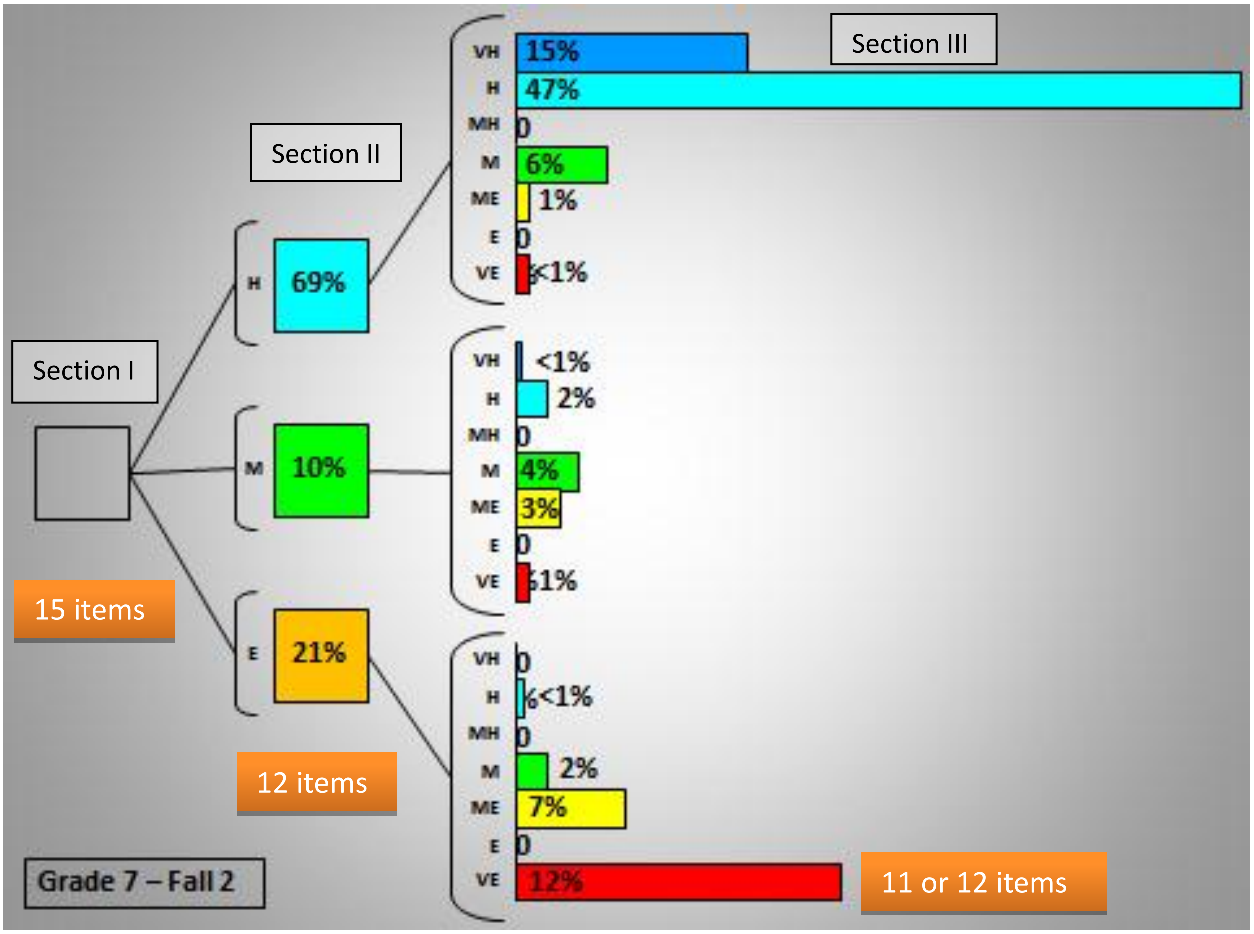
The focus of this chapter is to present the results of analyses related to the four research questions of the study using Mantel-Haenszel tests, Logistic Regression and judgmental strategy for instructional sensitivity. This chapter consists of three sections: (1) empirical methods, (2) judgmental approaches, and (3) the comparison of results of empirical methods and those of judgmental approaches. The section on empirical methods includes: (a) data description, (b) results of Logistic regression analyses, (c) results of Mantel-Haenszel tests, and (d) a comparison of the two methods. The section on judgmental approaches includes: (a) sample description, (b) data analyses, and (c) results. The third section compares items detected by different approaches and examines how results of empirical methods correlate to the ones of judgmental method.

Empirical Methods

Data Description

The Kansas State Interim Assessments are multi-stage adaptive design computer tests. In total, ninety items were given to 5,510 seventh graders in the second testing window of the assessment in mathematics. The second testing window was open from October 30 to December 31, 2010. It was composed of three sections: (1) in Section I, 15 items were given to all the test takers; (2) in Section II, three sets of 12 items of varying difficulties were given based on students’ performance on the first 15 items; (3) in Section III, students were further divided into 21 groups based on their performance in Section II. In Section II, test takers were divided into three groups – high difficulty group, medium difficulty group, and the easy group. As Figure 10 shows, 12 items of high difficulty were given to 69% of the total test takers; 12 items of medium difficulty were given to 10% of the total test takers, and 12 easy items were given to the remaining 21% of test takers. In Section III, students in some groups were given 11 items, while some were given 12 items. Thus, in this window, some students were given 38 items and some were given 39 items in total.

Figure 10. Multi-Stage Adaptive Design and Student Sample Distribution in Pathways



In order to ensure a large enough sample size, the common items used for the first two pathways were selected for analysis. To be specific, fifteen items were given in Section I; twelve items were given in Section II; and eight items used for both pathways of "very high difficulty"

("VH" in Figure 10) and of "high difficulty" ("H" in Figure 10) were in common in Section III. Thus, in total, thirty-five items were used for analysis in this study.

Demographic Information

Out of the 5,510 test takers, 3,446 students were given these 35 items. Approximate demographics of these 3,446 students were as follows: 1.1% Native American, 1.9% Asian, 4.1% Black, 13.7% Hispanic, 74.9% white, 4% multiracial, and .2% Pacific Islander. About 11% of these students received reduced cost lunch and about 27% received free lunch. About 62% of these students did not report what kind of lunch program they received. Table 3 presents the details of demographic information.

Table 3. The Demographics of the Student Sample

	Gender			**Lunch**			**Race**	
	n	percent		*n*	percent		*n*	percent
Girl	1734	50.3	*Reduced*	380	11.0	*Native American*	38	1.1
Boy	1706	49.5	*Free*	932	27.0	*Asian*	67	1.9
						Black	42	4.1
						Hispanic	71	13.7
						White	2577	74.8
						multiracial	138	4.0
						Pacific Islander	7	.2
Missing		.2		2134	61.9		6	.2
Total	3446	100.0		3446	100.0		3446	100.0

Descriptive Statistics of Items

After the test was administered to the students, the teachers were asked to login online and enter the indicators they instructed. A list of tested indicators for the teacher's grade and subject were presented after the teacher's login (refer to Figure 7). Thus, the membership (i.e., the instructed group or uninstructed group) of each student changed across the items. The number of students in each group was also different for different items. Table 4 summarizes the sample size of each group for each item, the mean value of θ (proficiency) of each group for each item, and the *p*-value (item difficulty in classical test theory) of each item for each group.

Table 4. Descriptive *Statistics* of Test Items

Item	**Group**	**n** **(total = 3446)**	$\bar{\theta}$ **(SD)** **(Proficiency)**	***p*-value**
1	Uninstructed	2306	.257(.76)	.45
	Instructed	1140	.207 (.73)	.52
2	Uninstructed	2420	.224(.75)	.73
	Instructed	1026	.281(.76)	.78
3	Uninstructed	1624	.185(.70)	.66
	Instructed	1822	.291(.79)	.81
4	Uninstructed	195	.221(.78)	.77
	Instructed	3251	.242(.75)	.82
5	Uninstructed	195	.221(.78)	.87
	Instructed	3251	.242(.75)	.91
6	Uninstructed	1748	.185(.72)	.62
	Instructed	1698	.298(.78)	.77
7	Uninstructed	2467	.196(.73)	.29
	Instructed	979	.353(.80)	.52
8	Uninstructed	825	.144(.68)	*.42*
	Instructed	2621	.271(.77)	*.41*
9	Uninstructed	2664	.234(.75)	.43
	Instructed	782	.264(.77)	.52
10	Uninstructed	490	.043(.63)	.60
	Instructed	2956	.274(.77)	.68
11	Uninstructed	825	.144(.68)	*.76*
	Instructed	2621	.271(.77)	*.75*
12	Uninstructed	825	.144(.68)	.32
	Instructed	2621	.271(.77)	.40
13	Uninstructed	195	.221(.78)	.38
	Instructed	3251	.242(.75)	.55
14	Uninstructed	2664	.234(.75)	.57
	Instructed	782	.264(.77)	.67
15	Uninstructed	2467	.196(.73)	.48
	Instructed	979	.353(.80)	.68
16	Uninstructed	2420	.224(.75)	.65
	Instructed	1026	.281(.76)	.81
17	Uninstructed	2092	.157(.69)	*.82*
	Instructed	1354	.371(.82)	*.81*
18	uninstructed	1738	.102(.67)	.46
	Instructed	1708	.382(.81)	.54
19	Uninstructed	1738	.102(.67)	.18
	Instructed	1708	.382(.81)	.26

(Table 4 continued)

		n	**$\bar{\theta}$ (SD)**	
Item	**Group**	**(total = 3446)**	**(Proficiency)**	***p*-value**
20	Uninstructed	1702	.162(.69)	.79
	Instructed	1744	.318(.80)	.94
21	Uninstructed	490	.043(.62)	.83
	Instructed	2956	.274(.77)	.86
22	Uninstructed	2092	.157(.69)	.53
	Instructed	1354	.371(.82)	.61
23	Uninstructed	1624	.185(.70)	.55
	Instructed	1822	.291(.79)	.65
24	Uninstructed	1738	.102(.67)	.70
	Instructed	1708	.382(.81)	.77
25	Uninstructed	2694	.232(.75)	.63
	Instructed	752	.272(.75)	.67
26	Uninstructed	2467	.196(.73)	.90
	Instructed	979	.353(.80)	.91
27	Uninstructed	2397	.255(.76)	.68
	Instructed	1049	.209(.74)	.70
28	Uninstructed	1222	.217(.73)	.*94*
	Instructed	2224	.254(.77)	.*94*
29	Uninstructed	1222	.217(.73)	.59
	Instructed	2224	.254(.77)	.62
30	Uninstructed	1748	.185(.72)	.29
	Instructed	1698	.298(.78)	.41
31	Uninstructed	1702	.162(.69)	.51
	Instructed	1744	.318(.80)	.72
32	Uninstructed	2694	.232(.75)	.21
	Instructed	752	.272(.75)	.41
33	Uninstructed	195	.221(.78)	.57
	Instructed	3251	.242(.75)	.62
34	Uninstructed	2306	.257(.76)	.12
	Instructed	1140	.207(.73)	.36
35	Uninstructed	2397	.255(.76)	.30
	Instructed	1049	.209(.74)	.34

Note: The θ ranges from -1.234 to 4, with a mean of .241.

Data from Table 4 show that 1) on average, students in the instructed group had higher performance on the test items; 2) generally speaking, the mean value of θ (proficiency) is higher for the instructed group than that for the uninstructed group; and 3) the *p*-values of most items are higher for the instructed group than those for the uninstructed group, indicating that most items appear easier to the instructed group but harder to the uninstructed group. The exception is that students in the uninstructed group were of a higher proficiency level than those in the instructed group for Item 1 ($\bar{\theta}_{uninstructed}$ = .257; $\bar{\theta}_{instructed}$ = .207), Item 27 ($\bar{\theta}_{uninstructed}$ = .255; $\bar{\theta}_{instructed}$ = .209), Item 34 ($\bar{\theta}_{uninstructed}$ = .257; $\bar{\theta}_{instructed}$ = .207) and Item 35($\bar{\theta}_{uninstructed}$ = .255; $\bar{\theta}_{instructed}$ = .209).

In addition, for Items 8, 11, 17 and 28, students in the uninstructed group performed slightly higher than or equally well (i.e., Item 28) as those in the instructed group. Further, students from both instructed and uninstructed groups had nearly identical high performance on these items, indicating that these items did not distinguish students well based on the instruction they received.

The comparison of the *p*-values shows that Items 5, 20, 26 and 28 were the easiest items among the 35 items given to this sample of students. The considerably large difference between the *p*-values of Item 20 for the uninstructed and instructed groups ($p_{\text{uninstructed}}$ = .79, $p_{\text{instructed}}$ = .94) indicates that this item discriminated students well in terms of their performance. The other three items were not discriminant. On the contrary, Items 19, 32, 34 and 35 were the hardest items. Students from both groups did not perform well on these items. However, all four items, though very hard, discriminated students well, especially Items 19, 32 and 34.

<u>Research Questions and Results</u>

This study addressed four questions:

1. To what extent are the items in the state testing program sensitive to instruction?
2. Are item performance differences related to differences in curricular content covered in instruction?

3. How does the instruction in content tested on the state testing program influence students' performance?
4. To what extent do the instructional sensitivity judgments made by curriculum experts agree with the results of empirical methods?

In order to answer the first three research questions, each item was analyzed using the DIFAS 4.0 (differential item functioning analysis system) (Penfield, 2007) for the standard MH procedures and with both the SPSS program (Version 18) and SAS program (Version 9.3) for LR analysis. In both cases, the θ estimated by using the 1-PL IRT model was used as the matching criterion.

Logistic Regression Procedures

The basic logistic regression equation is denoted as $P(U=1)=\frac{e^{z}}{1+e^{z}}$ (refer to Equation 14). In this study, three models were used to detect instructionally sensitive items. In Model 1, Z = $\beta_0 + \beta_1\theta$, where the student's proficiency (θ) was the only independent variable that predicted his/her probability of answering the item correctly. In Model 2, the categorical grouping variable (G) was added to the model: Z = $\beta_0 + \beta_1\theta + \beta_2 G$ (refer to Equation 15); thus, the difference in probability of responding correctly to the item due to membership could be measured after matching students on the same proficiency levels. In Model 3, the interaction between the student's proficiency and his/her membership was taken into account: Z = $\beta_0 + \beta_1\theta + \beta_2 G + \beta_3(\theta * G)$ (refer to Equation 16). By comparing the fit of Model 3 (i.e., the χ^2statistics) to that of Model 1, the uniform and the non-uniform DIF can by tested simultaneously. The uniform DIF can be tested by comparing the fit of Model 2 to that of Model 1. For this study, a significant $\Delta\chi^2$ from Model 1 to Model 3 indicates instructional sensitivity of an item due to the interaction between a student's proficiency and his/her membership in instruction. By the same token, a mere significant $\Delta\chi^2$from Model 2 to Model 3 indicates instructional sensitivity of an item due to the interaction only.

The $R^2$of each model represents the practical importance of the statistical differences. The comparison of the Nagelkerke R^2 (i.e., the ΔR^2) between Model 1 and Model 2 provides the information about how important the uniform DIF is if there is a uniform DIF detected. Similarly, the comparison of the Nagelkerke R^2 between Model 1 and Model 3 shows the practical importance of the non-uniform DIF if a non-uniform DIF is detected. For the purpose of this study, the comparison of Nagelkerke R^2 shows the degree to which a test item accurately reflects the impact of instruction on the content tested by the item. In other words, the ΔR^2 shows how sensitive an item is to instruction.

Table 5 presents the results of the logistic regression procedures. Items were reordered based on the effect size (i.e., the value of ΔR^2 between Models 1 and 3). The larger the ΔR^2 is, the more sensitive the test item is.

Table 5. Report of Logistic Regression Model Comparisons: χ^2, $\Delta\chi^2$, R^2, and ΔR^2

	$\chi^2(df=1)$	R^2	$\chi^2(df=1)$	R^2	$\chi^2(df=1)$	R^2	$\Delta\chi^2(df=2)$		$\Delta\chi^2(df=1)$		$\Delta\chi^2(df=1)$		ΔR^2
Item	**Model 1**	**M1**	**Model 2**	**M2**	**Model 3**	**M3**	**M3: M1**	***p***	**M2: M1**	***p***	**M3: M2**	***p***	**M3: M1**
34	207.395	.093	512.273	.220	512.838	.220	305.443	**<.001**	304.878	<.001	.565	.452	.127
20	147.514	.076	296.841	.149	298.579	.150	151.065	**<.001**	149.327	<.001	1.738	.187	.074
31	302.990	.114	438.110	.162	445.596	.165	142.606	**<.001**	135.120	<.001	7.486	.006	.051
32	307.054	.126	434.984	.175	440.039	.177	132.985	**<.001**	127.930	<.001	5.055	.025	.051
7	236.321	.091	356.944	.135	357.932	.135	121.611	**<.001**	120.623	<.001	.988	.320	.044
16	285.711	.113	378.029	.147	379.021	.148	93.310	**<.001**	92.318	<.001	.992	.319	.035
3	380.805	.154	472.591	.189	473.496	.189	92.691	**<.001**	91.786	<.001	.905	.341	.035
15	210.473	.079	303.074	.112	306.915	.114	96.442	**<.001**	92.601	<.001	3.841	.050	.035
6	485.551	.185	566.721	.214	567.666	.214	82.115	**<.001**	81.170	<.001	.945	.331	.029
30	159.185	.062	202.597	.079	205.664	.080	46.479	**<.001**	43.412	<.001	3.067	.080	.018
1	640.742	.226	663.851	.234	666.086	.235	25.344	**<.001**	23.109	<.001	2.235	.135	.009
14	549.541	.199	572.591	.207	573.404	.207	23.863	**<.001**	23.050	<.001	.813	.367	.008
13	462.775	.168	485.397	.176	485.501	.176	22.726	**<.001**	22.622	<.001	.104	.747	.008
23	494.312	.181	519.461	.189	519.682	.189	25.370	**<.001**	25.149	<.001	.221	.638	.008
9	418.082	.153	437.841	.160	439.243	.160	21.161	**<.001**	19.759	<.001	1.402	.236	.007
19	260.097	.112	265.901	.114	275.737	.118	15.640	**<.001**	5.804	.016	9.836	.002	.006
35	445.278	.171	454.339	.174	455.602	.174	10.324	**<.001**	9.061	.003	1.263	.261	.003
12	381.273	.143	390.122	.146	392.172	.146	10.899	**<.001**	8.849	.003	2.050	.152	.003
2	356.817	.145	363.410	.147	363.602	.148	6.785	**<.001**	6.593	.010	.192	.661	.003
5	81.514	.051	84.215	.053	84.240	.053	2.726	.256	2.701	.100	.025	.874	.002
10	493.588	.185	495.521	.186	498.511	.187	4.923	.085	1.933	.164	2.990	.084	.002
11	477.427	.192	482.307	.194	482.352	.194	4.925	.085	4.880	**.027**	.045	.832	.002
24	463.653	.183	465.344	.184	468.361	.185	4.708	.095	1.691	.193	3.017	.082	.002
4	361.146	.162	363.863	.163	364.566	.163	3.420	.181	2.717	.099	.703	.401	.001
22	711.314	.250	713.037	.250	715.189	.251	3.875	.144	1.723	.189	2.152	.142	.001
25	374.759	.141	376.534	.142	377.653	.142	2.894	.235	1.775	.183	1.119	.290	.001
27	114.588	.046	116.663	.047	118.528	.047	3.940	.139	2.075	.150	1.865	.172	.001
28	79.109	.063	79.129	.063	79.545	.064	.436	.804	.020	.888	.416	.519	.001
29	428.771	.159	431.509	.160	433.222	.160	4.451	.108	2.738	.098	1.713	.191	.001
33	723.972	.257	725.122	.258	726.516	.258	2.544	.280	1.150	.284	1.394	.238	.001
17	25.855	.012	27.857	.013	28.255	.013	2.400	.301	2.002	.157	.398	.528	.001
8	267.942	.101	271.297	.102	272.200	.102	4.258	.119	3.355	.067	.903	.342	.001
18	666.120	.234	666.832	.235	666.922	.235	.802	.670	.712	.399	.090	.764	.001
21	119.775	.061	120.125	.061	120.378	.061	.603	.740	.350	.554	.253	.615	0
26	165.593	.100	165.770	.100	165.820	.100	.227	.893	.177	.674	.050	.823	0

Note: "M1" means "Model 1", the same to "M2" and "M3"; "M3: M1" means "M3 vs. M1", the same to "M2: M1" and "M3: M2". Sensitive items are in bold.

In Table 5, the χ^2statistics (refer to columns labeled as "**χ^2 (*df* = 1) Model 1**", "**χ^2 (*df* = 1) Model 2**"and "**χ^2 (*df* = 1) Model 3**") and the NagelkerkeR^2 effect size (refer to columns labeled as "**R^2M1**", "**R^2M2**" and "**R^2M3**") for each model were reported first. Next, the pair-wise comparison of χ^2 statistics between each set of two models (refer to two columns labeled as "**$\Delta\chi^2$**") and its corresponding significance test (***p***) were reported. Finally, the ΔR^2 between Model 3 and Model 1 was reported to indicate the practical importance of instructional sensitivity due to the interaction between students' proficiency and their instructional experience (i.e., their membership).

The measure of ΔR^2 represents the degree of instructional sensitivity of the item and was used to classify the degree of instructional sensitivity of the items. The magnitude of the effect size (ΔR^2) for the items shows that Items 34, 20, 31, 32, 7, 3, 15 and 16 were considerably more sensitive to instruction than others. Item 34 was the most sensitive item ($\Delta R^2 = .127$). Items 31 and 32 were equally sensitive ($\Delta R^2 = .051$), and Items 3, 15 and 16 were equally sensitive ($\Delta R^2 = .035$). Results show that twenty out of thirty-five items were instructionally sensitive.

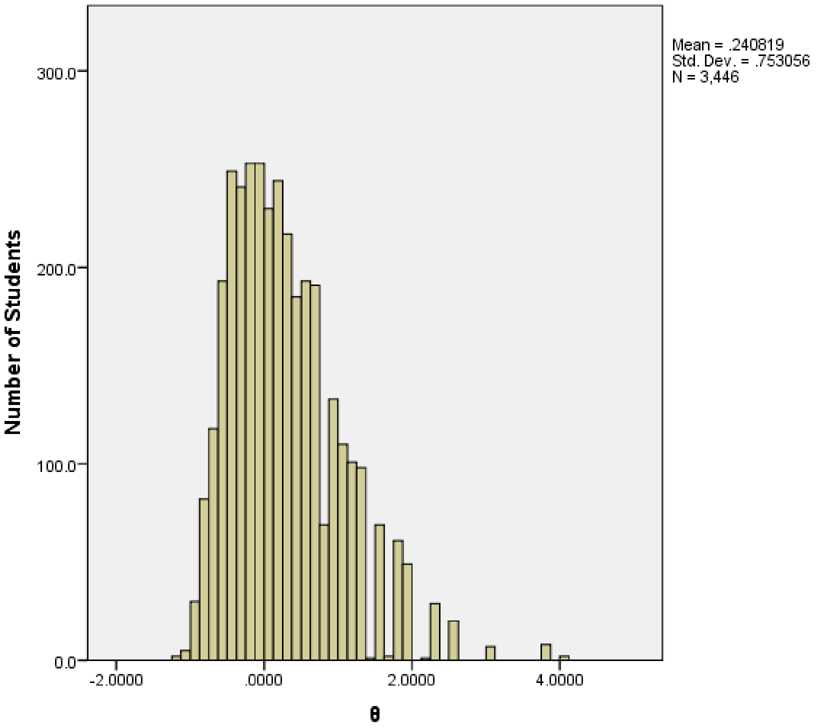
Mantel-Haenszel Tests

The DIFAS 4.0 (Penfield, 2007) was used to conduct Mantel-Haenszel tests to detect instructionally sensitive items. Students' proficiency levels (θ) were used as the matching criteria. As shown in Figure 10, students who were given the items presented in the top two pathways (i.e., the "Very High" difficulty and the "High" difficulty pathways) accounted for about 60% of the entire sample size. Using items from these two pathways for this study yielded a sub-sample of 3,446 students, which was about 63% of the original total sample size (i.e., 5,510). Because these 3,446 students were given the most difficult and/or the difficult items in Section III based on their performance on the previous two sections, they were competitive children in the cohort. Therefore, the θ for this group was not symmetrically distributed based on the original scale with a mean of 0, a minimum score of -4 and a maximum score of 4. Instead, the minimum θ score for this group was -1.234, indicating that these students were of comparatively high proficiency in mathematics. Table 6 is a summary of descriptive statistics of θ, and Figure 11 graphically shows how θ for the sample used in this study was distributed.

Table 6. Descriptive Statistics of Math Proficiency (θ)

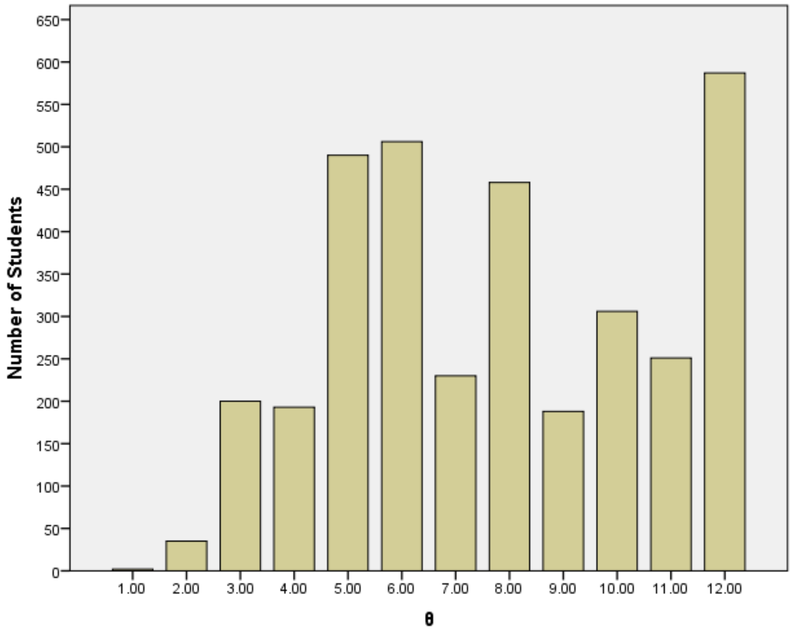
	N	**Min.**	**Max.**	**Mean**	**SD**	**25th Percentile**	**50th Percentile**	**75th Percentile**
θ	3,446	-1.234	4.000	.241	.75	-.352	.128	.640

Figure 11. Distribution of Math Proficiency (θ)



To conduct MH tests, the continuous variable θ was converted into a categorical variable with 12 categories. A histogram showing θ distribution in categories is presented in Figure 12.

Figure 12. A Histogram of Proficiency (θ) Distribution in Categories



The Mantel-Haenszel method is a contingency table method that compares the likelihood of success on the item for students of the two groups after they were matched on proficiency. The ratio of these likelihoods was used as the index to identify instructional sensitivity. Students in the sample were matched based on their proficiency (θ). Referring to Equation 13, α_i is the ratio of the odds (p/q) that the uninstructed group of students succeeded on the item to the odds that the instructed group of students succeeded on the item. The α_i ranges from 0 to ∞, with the value of 1.0 indicating the item is not sensitive, with values less than 1.0 indicating the item is in favor of the instructed group, and with values greater than 1.0 indicating that the item is in favor of the uninstructed group, after students from both groups have been matched on their proficiency. For the purpose of convenient interpretation, logistic transformation was made by multiplying the α by -2.35 to produce the Δ_{MH} (refer to Equation 22). The values are then asymptotically normally distributed around zero, and a negative value indicates that the item favors the instructed group while a positive value indicates the opposite (Camilli & Shepard, 1994). A zero value indicates no instructional sensitivity. The DIFAS 4.0 reports the Mantel-Haenszel common log-odds ratio (MH LOR), which was used as the measure of effect size. Table 7 presents the results from the MH tests on detecting instructionally sensitive items. The χ^2 statistics, MH LOR, and the standard error of LOR are reported. Items were reordered based on the values of effect size (i.e., the MH LOR).

Table 7. Results of Mantel-Haenszel Tests

	MH χ^2	**MH**	
Item	**(*df* = 1)**	**LOR**	**SE of LOR**
34	301.419	-1.550	.09
20	141.556	-1.318	.12
32	135.781	-1.046	.09
7	125.014	-.889	.08
16	87.393	-.869	.09
31	132.652	-.853	.07
3	91.095	-.799	.08
13	21.514	-.766	.16
15	89.807	-.761	.08
6	79.484	-.722	.08
30	43.426	-.487	.07
14	22.257	-.434	.09
9	19.324	-.386	.09
23	25.159	-.377	.07
5	2.329	-.373	.23
1	19.637	-.359	.08
4	2.822	-.340	.19
12	8.950	-.268	.09
19	7.567	-.244	.09
35	7.882	-.242	.08
2	6.163	-.235	.09
11	4.367	.212	.10
33	.947	-.181	.17
10	1.945	-.158	.11
8	2.751	.144	.08
25	2.021	-.133	.09
29	2.837	-.133	.08
22	2.322	-.124	.08
24	2.012	-.122	.08
17	1.602	.121	.09
27	1.528	-.104	.08
18	1.195	.086	.08
21	.202	-.070	.13
26	.110	.053	.13
28	.004	-.021	.15

Note: Items were reordered by the magnitude of the value of MH LOR. A negative value indicates that the item is in favor of the instructed group. Items 4 and 5 are in bold, indicating that although they have larger effect size, they were not detected as instructionally sensitive based on the MH χ^2 statistics.

As shown in Table 7, nineteen out of the thirty-five items were detected as instructionally sensitive by using MH tests. Among the 19 instructionally sensitive items, ten were sensitive to a large degree, two were sensitive to a moderate degree, and seven were sensitive to a negligible degree according to the ETS classification. When the effect size was combined with the statistical significance to classify the sensitive items, only one out of nineteen sensitive items was sensitive to a large degree, two out of nineteen sensitive items were sensitive to a moderate degree, and all the others were only sensitive to a negligible degree. Items 34, 20 and 32 were detected as the most sensitive items. Additionally, all the sensitive items were in favor of the instructed group.

Comparison of Logistic Regression and Mantel-Haenszel Tests

The results obtained from the logistic regression procedure and the Mantel-Haenszel tests were compared (see Table 8) to address the following questions:

1) Did both methods detect the same items as instructionally sensitive?
2) Based on the measure of effect size, to what degree do the two methods agree with each other?
3) Were both methods equally powerful in detecting items that were sensitive due to the interaction of students' instructional experience and their proficiency? If not, which method was more powerful?

Items in Table 8 were ordered based on the effective sizes of the two methods, respectively.

Table 8. A Comparison of Results Obtained from Two Empirical Methods

Logistic Regression			Mantel-Haenszel Tests		
Item Order	$\Delta\chi^2$	Interaction	Item Order	LOR	Interaction
34	.127	Y	34	-1.550	Y
20	.074	Y	20	-1.318	N
31	.051	Y	32	-1.046	N
32	.051	Y	7	-.889	N
7	.044	Y	16	-.869	N
16	.035	Y	31	-.853	Y
3	.035	Y	3	-.799	N
15	.035	Y	13	-.766	N
6	.029	Y	15	-.761	N
30	.018	Y	6	-.722	N
1	.009	Y	30	-.487	N
14	.008	Y	14	-.434	N
13	.008	Y	9	-.386	N
23	.008	Y	23	-.377	N
9	.007	Y	5	-.373	-
19	.006	Y	1	-.359	N
35	.003	Y	4	-.340	-
12	.003	Y	12	-.268	N
2	.003	Y	19	-.244	Y
5	.002	-	35	-.242	N
10	.002	-	2	-.235	N
11	.002	N	11	.212	-
24	.002	-	33	-.181	-
4	.001	-	10	-.158	-
22	.001	-	8	.144	-
25	.001	-	25	-.133	-
27	.001	-	29	-.133	-
28	.001	-	22	-.124	-
29	.001	-	24	-.122	-
33	.001	-	17	.121	-
17	.001	-	27	-.104	-
8	.001	-	18	.086	-
18	.001	-	21	-.070	-
21	0	-	26	.053	-
26	0	-	28	-.021	-

Note: Results produced by the MH methods were sorted descending by the magnitude of the value of LOR. Only three (Items 34, 16, and 2) out of nineteen items detected as instructionally sensitive by the MH tests are sensitive due to the interaction between proficiency and membership.

Table 8 compares results produced from both methods. The first three columns contain information from the LR procedure, and the second three columns contain information from the MH tests. For both methods, the items were reordered based on their degrees of sensitivity (see Columns 1 and 4). Columns 2 and 5 report the measures of effect size for both methods. Based on the measure of effect size per method, the items were reordered, respectively. Columns 3 and 6 report information about whether the sensitivity of the detected items was due to the interaction of students' proficiency and their instructional experiences. A "Y" represents sensitivity due to interaction; an "N" represents sensitivity due to membership only; and a "-" means the item was not sensitive.

The comparison of these two methods indicates the following findings:

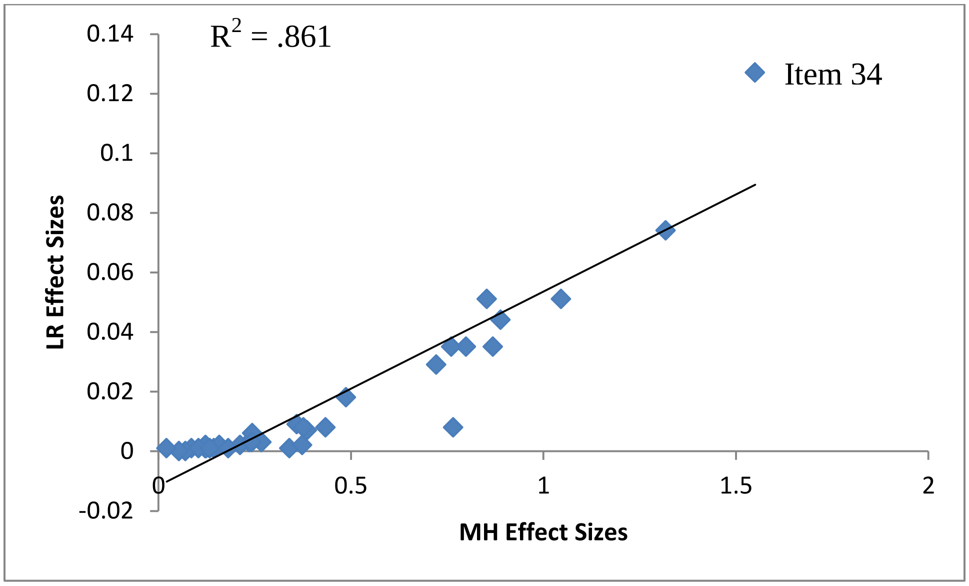
First, both methods detected 19 items in common which were instructionally sensitive. Logistic Regression procedures detected one more sensitive item (Item 11) in addition to the 19 items. However, the degrees of sensitivity for items detected were not exactly the same when ranked by the two methods, although the rankings were similar. In Table 9, items with the same rankings were bolded; items with adjacent rankings were underlined. Both methods detected Item 34 as the most sensitive item, Item 20 as the second most sensitive item, and Item 2 as the least sensitive item.

Table 9. Ranking of Sensitivity for LR and MH Methods

		More Sensitive														**Less Sensitive**				
LR	**34**	**20**	**32**	31	7	16	**3**	15	6	30	1	**14**	13	**23**	9	19	35	12	**2**	11
MH	**34**	**20**	**32**	7	16	31	**3**	13	15	6	30	**14**	9	**23**	1	12	19	35	**2**	X

Second, the effect sizes from both methods were highly correlated (*r* = .93), which means the two methods agreed with each other to a high degree in detecting instructionally sensitive items. Figure 13 is the scatter plot showing the relationship between the effect sizes from both methods.

Figure 13. A Scatterplot of the Relationship between the MH and LR Effect Sizes



Third, in terms of the degree of sensitivity, the LR procedure detected eight items with a large to moderate degree of sensitivity, while the MH method only detected three items that were of a large to moderate degree of sensitivity. Items 34, 20 and 32 were the common items meeting the criteria of being moderate or large in sensitivity for both methods, and Item 34 was the only item that was of a large degree of sensitivity “ranked” by both methods.

Fourth, among the 20 sensitive items detected by the LR procedure, 19 of them were sensitive due to the interaction of students’ instructional experience (i.e., membership or grouping variable) and their proficiency. However, among the 19 sensitive items detected by the MH tests, only three (Items 34, 16, and 2) were sensitive due to the interaction of students’ instructional experience and their proficiency. Therefore, the LR procedure is more effective in detecting items' instructional sensitivity due to the interaction of students' instructional experience and their proficiency.

Answers to Research Questions

Question 1: To what extent are the items in the state testing program sensitive to instruction? Results show that more than half of the items in the state interim assessment on mathematics for seventh graders were sensitive to instruction. Among the 19 items detected by both MH tests and LR procedures as instructionally sensitive, one item was classified as sensitive to a large degree, about five items (the numbers were different when different statistical methods were used) were classified as sensitive to a moderate degree, and all the others were classified to a small degree.

Question 2: Are item performance differences related to differences in curricular content covered in instruction? Results show that all the detected items were in favor of the instructed group. In other words, students from the instructed group had a higher probability of succeeding on each of the detected items than those who were from the uninstructed group, after they were matched on proficiency. This finding indicates that the item performance differences are due to differences in curricular content covered in instruction. Students from different groups had significant differences in their performance on the detected items.

Question 3: How does the instruction in content tested on the state testing program influence students' performance? The finding is that students who received adequate instruction to be able to successfully answer the potential item prior to the assessment had a higher probability of responding to the item correctly than those who did not receive adequate instruction, after they were matched on proficiency. This finding indicates that instruction positively influenced students' performance.

Judgmental Approach

Sample Description

Thirteen[6] middle school teachers were recruited to review the same 35 multiple-choice items used for empirical methods and to rate which items reflected the impact of instruction on students' performance based on their knowledge and experience. Twelve of them were currently teaching or had recently taught seventh grade mathematics in Kansas. One of them taught seventh grade math in Ecuador, using the connected mathematics project (CMP)[7], but was familiar with the Kansas mathematics indicators for seventh grade. Nine out of the thirteen teachers were female; ten of them were currently teaching seventh grade math; and seven of them had been teaching math for more than ten years. Table 10 is a summary of the participants' background information.

[6] Popham (2003b) suggested a sample size could be as small as a half-dozen or so.

[7] The CMP website: http://connectedmath.msu.edu/.

Table 10. Background Information of Teachers

Participant	Gender	Currently Teaching 7th Grade Math?	When Taught 7th Grade Math?	Years Teaching Math	Years Teaching Math in Kansas
1	F	Y	-	27.0	27.0
2	M	N	2008-2011	10.0	10.0
3	M	Y	-	10.0	10.0
4	F	N	2010-2011	4.0	4.0
5	M	Y	-	1.0	1.0
6	F	Y	-	1.5	1.5
7	F	Y	-	21.0	21.0
8	F	Y	-	11.0	11.0
9	M	Y	-	3.0	3.0
10	F	Y	-	6.0	6.0
11	F	Y	-	14.0	14.0
12	F	N	2010-2011	1.0	0.0
13	F	Y	-	10.0	10.0

Data Analysis

Ratings on Items

In the first part of the test review, the participants reviewed the 35 multiple-choice items and rated them on an 11-point Likert scale “measuring” the degree of instructional sensitivity, where “0” represents “totally insensitive” and “10” represents “totally sensitive.” Results show that the teachers who participated in this research had very different opinions in regards to instructional sensitivity of the examined items, and their ratings were very dispersed on these items. Comparatively speaking, only ratings on Items 1, 7, 11, 19 and 34 were less dispersed (see Tables 12 & 13). Table 11 is a summary of teachers’ judgment on the items regarding instructional sensitivity. The items were reordered based on their means. The higher the mean value of an item, the more sensitive the item was rated by the teachers.

Table 11. Teachers' Ratings on Item Sensitivity

Item	Mean	SD	Min.	Max.
1	9.08	1.12	7	10
7	**8.92**	1.32	6	10
15	**8.92**	1.75	4	10
31	8.54	1.56	5	10
34	8.38	1.45	6	10
11	8.31	0.95	6	10
4	8.23	2.01	3	10
2	8.15	1.52	5	10
3	**8.00**	2.12	3	10
10	**8.00**	2.38	2	10
20	**8.00**	2.04	4	10
5	7.92	1.71	5	10
35	7.85	1.57	5	10
8	7.77	2.39	2	10
19	7.69	1.25	5	9
17	7.62	2.57	3	10
6	7.54	2.47	2	10
21	7.46	1.98	3	10
13	**7.38**	1.56	4	10
30	**7.38**	2.66	0	10
16	7.31	2.36	3	10
18	7.15	1.63	4	9
22	**6.92**	3.17	1	10
24	**6.92**	1.85	3	9
25	6.77	3.30	1	10
14	**6.69**	2.32	2	10
29	**6.69**	2.14	3	10
33	**6.69**	1.75	4	10
27	6.62	2.18	3	10
23	6.54	2.30	1	10
9	6.38	2.63	2	10
12	6.23	2.92	0	10
32	5.77	2.17	2	9
26	4.77	2.62	0	10
28	4.46	2.15	2	9

Note: The mean values in bold indicate these items shared the same ranking.

The analysis of intra-class correlations shows that the agreement or consensus among teachers on their ratings on item sensitivity was poor: *r* = .155 (when measures of absolute agreement were applied) and *r* = .182 (when measures of consistency were applied). The two-way random effect model was used to calculate intra-class correlations. For the measure of absolute agreement among raters, the single measure intra-class correlation is .155, with a 95% confidence interval between .085 and .270; for the measure of consistency, the single measure intra-class correlation is .182, with a 95% confidence interval between .102 and .308. For both measures, *F* (34, 408) = 3.889, *p* < .001; the Cronbach's Alpha (reliability) is .743 (N of raters = 13; N of items = 35).

Further, teachers who had ten or more years of teaching experience were selected and intra-class correlations were calculated again with this sub-sample. The agreement among teachers on item sensitivity increased slightly but was still poor: *r* = .224 (when measures of absolute agreement were applied) and *r* = .252 (when measures of consistency were applied). For the measure of absolute agreement among raters, the single measure intra-class correlation is .224, with a 95% confidence interval between .093 and .600; for the measure of consistency, the single measure intra-class correlation is .252, with a 95% confidence interval between .107 and .638. For both measures, *F* (6, 204) = 12.814, *p* < .001; the Cronbach's Alpha (reliability) is .922 (N of raters = 7; N of items = 33).

Based on the information presented in Table 11, most items were rated as considerably sensitive although the teachers' ratings obviously varied for each single item. Among thirty-five items, twenty-two of them were rated above Point 7 based on the mean value of ranking, and only two of them were rated below Point 5 based on the mean value of ranking. The most sensitive items were Items 1, 7 and 15; the least sensitive items were Items 26 and 28. Every item was rated as "totally sensitive" (Point 10) or close (Point 9) by at least one teacher. When reordered based on the range of rating for each item, Items 1, 7, 11, 19 and 34 shared the most agreement among teachers (range = 3 or range = 4), and Items 12, 26 and 30 were of the least agreement (range = 10). Teachers also greatly disagreed on Items 22, 23 and 25 (range = 9). Table 12 shows the degree to which the teachers agree or disagree on the items based on the range of rating for each item. Because the value of range is very easily influenced by outliers, the conclusions made above might not be reliable. Therefore, the items were than reordered based on their standard deviations of teachers' ratings that are also an indicator of dispersion. Results show that Items 11, 1 and 19 were the top three that shared most agreement. Items 12, 22 and 25 shared least agreement on rating. Table 13 shows that order of the items from the least dispersed to the most dispersed in rating based on standard deviation.

As far as the relationship between the teachers' teaching experience and rating, a simple bivariate correlation shows that there was no relationship between these two variables: r = -.078, p = .80.

Table 12. Item Order on Rating Ranges

Item	Range	Min.	Max.	Mean	SD
1	3	7	10	9.08	1.12
7	4	6	10	8.92	1.32
34	4	6	10	8.38	1.45
11	4	6	10	8.31	0.95
19	4	5	9	7.69	1.25
31	5	5	10	8.54	1.56
2	5	5	10	8.15	1.52
5	5	5	10	7.92	1.71
35	5	5	10	7.85	1.57
18	5	4	9	7.15	1.63
15	6	4	10	8.92	1.75
20	6	4	10	8.00	2.04
13	6	4	10	7.38	1.56
24	6	3	9	6.92	1.85
33	6	4	10	6.69	1.75
4	7	3	10	8.23	2.01
3	7	3	10	8.00	2.12
17	7	3	10	7.62	2.57
21	7	3	10	7.46	1.98
16	7	3	10	7.31	2.36
29	7	3	10	6.69	2.14
27	7	3	10	6.62	2.18
32	7	2	9	5.77	2.17
28	7	2	9	4.46	2.15
10	8	2	10	8.00	2.38
8	8	2	10	7.77	2.39
6	8	2	10	7.54	2.47
14	8	2	10	6.69	2.32
9	8	2	10	6.38	2.63
22	9	1	10	6.92	3.17
25	9	1	10	6.77	3.30
23	9	1	10	6.54	2.30
30	10	0	10	7.38	2.66
12	10	0	10	6.23	2.92
26	10	0	10	4.77	2.62

Table 13. Item Order on Standard Deviation of Ratings

Item	SD	Mean	Min.	Max.
11	0.95	8.31	6	10
1	1.12	9.08	7	10
19	1.25	7.69	5	9
7	1.32	8.92	6	10
34	1.45	8.38	6	10
2	1.52	8.15	5	10
13	1.56	7.38	4	10
31	1.56	8.54	5	10
35	1.57	7.85	5	10
18	1.63	7.15	4	9
5	1.71	7.92	5	10
33	1.75	6.69	4	10
15	1.75	8.92	4	10
24	1.85	6.92	3	9
21	1.98	7.46	3	10
4	2.01	8.23	3	10
20	2.04	8.00	4	10
3	2.12	8.00	3	10
29	2.14	6.69	3	10
28	2.15	4.46	2	9
32	2.17	5.77	2	9
27	2.18	6.62	3	10
23	2.30	6.54	1	10
14	2.32	6.69	2	10
16	2.36	7.31	3	10
10	2.38	8.00	2	10
8	2.39	7.77	2	10
6	2.47	7.54	2	10
17	2.57	7.62	3	10
26	2.62	4.77	0	10
9	2.63	6.38	2	10
30	2.66	7.38	0	10
12	2.92	6.23	0	10
22	3.17	6.92	1	10
25	3.30	6.77	1	10

Rating on Indicators

In the second part of the test review, the teachers reviewed the descriptions of the Standards, Benchmarks and Indicators for seventh grade mathematics to rate the clarity of each indicator based on the following predetermined questions:

1) For purposes of a teacher's instructional planning, how clearly are this state test's assessment targets (the skills and knowledge it measures) described?
2) Are they stated with sufficient clarity that almost all of the state's teachers can identify what the benchmark and/or indicator really means?

They were instructed to respond to these questions by circling one number from the number line measuring clarity of the descriptions for each indicator and providing their rationale or comments. The number line is an 11-point Likert scale (from 0 to 10) with "0" representing "totally unclear" and "10" representing "extremely clear."

Table 14 below is a brief summary of the indicators. There were four standards to be measured for a seventh grader: I) Numbers and Computation, II) Algebra, III) Geometry, and IV) Data. Under Standard I, there were two benchmarks (Benchmarks 1 and 4); further, there was one indicator under Standard I – Benchmark 1 and five indicators under Standard I – Benchmark 4. Under Standard II, there were two benchmarks (Benchmarks 1 and 2); further, there were three indicators under Standard II – Benchmark 1 and three indicators under Standard II – Benchmark 2. Under Standard III, there were three benchmarks (Benchmarks 1, 2 and 3); further, there were seven indicators under Standard III – Benchmark 1; there were four indicators under Standard III – Benchmark 2; and there was one indicator under Standard III – Benchmark 3. Under Standard IV, there was only one Benchmark, and there were nine indicators under this construct. Appendix D (pp. 121-124) provides more details on the content tested in this assessment.

Table 14. A Summary of Standards, Benchmarks, and Indicators

Standards	Benchmarks	Indicators
Standard I *Number and Computation*	**Benchmark 1** Number Sense	Ind. A1a
	Benchmark 4 Computation	Ind. K2a Ind. K2b Ind. K2c Ind. K2d Ind. K5
Standard 2 *Algebra*	**Benchmark 1** Patterns	Ind. K1a Ind. K1b Ind. K4
	Benchmark 2 Variable, Equations, and Inequalities	Ind. K7 Ind. K8 Ind. A1
Standard 3 *Geometry*	**Benchmark 1** Geometric Figures and Their Properties	Ind. K3a Ind. K3b Ind. K3c Ind. K3d Ind. K3e Ind. K3f Ind. K3g
	Benchmark 2 Measurement and Estimation	Ind. K4 Ind. K6a Ind. K6b Ind. A1c
	Benchmark 3 Transformational Geometry	Ind. A3
Standard 4 *Data*	**Benchmark 2** Statistics	Ind. K1a Ind. K1b Ind. K1c Ind. K1d Ind. K1e Ind. K1f Ind. K1g Ind. A3a Ind. A3b

Results show that the teachers had more agreement on indicators than on items. They had 100% agreement on five indicators. The descriptions of these ten indicators were rated as extremely clear. These ten indicators were under Standard III – Benchmark 1. In addition to the five clearest indicators, the teachers shared fairly high agreement on most of the rest of the indicators. The indicators that they had most disagreement on were “s1b1a1a”, “s2b1k1a”, “s4b2a3a”, and “s2b1k4.” Table 15 below shows the reordered indicators starting from the clearest one based on the measure of standard deviation.

Table 15. Indictors Ordered by Teachers' Agreement on Clarity Ratings

Indicator	SD	Mean	Min.	Max.
s3b1k3b	0.00	10.00	10	10
s3b1k3c	0.00	10.00	10	10
s3b1k3d	0.00	10.00	10	10
s3b1k3e	0.00	10.00	10	10
s3b1k3f	0.00	10.00	10	10
s1b4k2a	0.28	9.92	9	10
s1b4k2b	0.28	9.92	9	10
s1b4k2c	0.28	9.92	9	10
s3b1k3a	0.28	9.92	9	10
s3b2k6a	0.28	9.92	9	10
s3b2k6b	0.28	9.92	9	10
s1b4k2d	0.44	9.77	9	10
s3b1k3g	0.55	9.85	8	10
s4b2k1b	0.63	9.69	8	10
s4b2k1g	0.63	9.69	8	10
s4b2k1d	0.65	9.62	8	10
s4b2k1e	0.65	9.62	8	10
s4b2k1f	0.65	9.62	8	10
s4b2k1a	0.77	9.62	8	10
s4b2k1c	0.78	9.54	8	10
s3b2k4	0.87	9.62	7	10
s2b2k8	1.04	9.38	7	10
s3b2a1c	1.22	9.00	6	10
s2b1k1b	1.27	8.54	6	10
s2b2a1	1.30	8.77	7	10
s1b4k5	1.44	8.92	6	10
s4b2a3b	1.45	8.54	6	10
s3b3a3	1.55	9.08	5	10
s2b2k7	1.60	8.69	5	10
s1b1a1a	1.98	6.62	3	10
s2b1k1a	2.02	7.92	4	10
s4b2a3a	2.07	7.54	3	10
s2b1k4	2.30	8.46	3	10

Note: "s1b4k2a" represents "Standard I, Benchmark 4, Indicator K2a".

Results also show that most of the indicators were rated as very clear. To be specific, five indictors had a mean score of 10; nineteen indicators had a mean score between 9 and 10; six indicators had a mean score between 8 and 9; two indicators had a mean score between 7 and 8;

only one indicator had a mean score of 6.62, the lowest mean score of the rating. This indicator was “s1b1a1a”. Thus, most indicators were clear enough for the state’s teachers to identify what the indicators really mean, based on these teachers’ knowledge and experience. Table 16 below orders the indicators based on their clarity (i.e., the mean scores of the rating).

Table 16. Indicators Ordered by Teachers’ Ratings on Clarity

Indicator	Mean	SD	Min.	Max.
s3b1k3b	10.00	0.00	10	10
s3b1k3c	10.00	0.00	10	10
s3b1k3d	10.00	0.00	10	10
s3b1k3e	10.00	0.00	10	10
s3b1k3f	10.00	0.00	10	10
s1b4k2a	9.92	0.28	9	10
s1b4k2b	9.92	0.28	9	10
s1b4k2c	9.92	0.28	9	10
s3b1k3a	9.92	0.28	9	10
s3b2k6a	9.92	0.28	9	10
s3b2k6b	9.92	0.28	9	10
s3b1k3g	9.85	0.55	8	10
s1b4k2d	9.77	0.44	9	10
s4b2k1b	9.69	0.63	8	10
s4b2k1g	9.69	0.63	8	10
s3b2k4	9.62	0.87	7	10
s4b2k1a	9.62	0.77	8	10
s4b2k1d	9.62	0.65	8	10
s4b2k1e	9.62	0.65	8	10
s4b2k1f	9.62	0.65	8	10
s4b2k1c	9.54	0.78	8	10
s2b2k8	9.38	1.04	7	10
s3b3a3	9.08	1.55	5	10
s3b2a1c	9.00	1.22	6	10
s1b4k5	8.92	1.44	6	10
s2b2a1	8.77	1.30	7	10
s2b2k7	8.69	1.60	5	10
s2b1k1b	8.54	1.27	6	10
s4b2a3b	8.54	1.45	6	10
s2b1k4	8.46	2.30	3	10
s2b1k1a	7.92	2.02	4	10
s4b2a3a	7.54	2.07	3	10
s1b1a1a	6.62	1.98	3	10

The two-way random effect model was used to calculate intra-class correlations. For the measure of absolute agreement among raters, the single measure intra-class correlation is .339, with a 95% confidence interval between .229 and .492; for the measure of consistency, the single measure intra-class correlation is .371, with a 95% confidence interval between .256 and .525. For both measures, *F* (32, 384) = 8.662, *p* < .001; the Cronbach's Alpha (reliability) is .885 (N of raters = 13; N of items = 33). Thus, the agreement or consensus among teachers on their rating on indicator clarity was fair.

Further, teachers who had ten or more years of teaching experience were selected and intra-class correlations were calculated again with this sub-sample. The agreement among teachers on item sensitivity decreased from fair to poor: *r* = .121 (when measures of absolute agreement were applied) and *r* = .200 (when measures of consistency were applied). For the measure of absolute agreement among raters, the single measure intra-class correlation is .121, with a 95% confidence interval between .043 and .425; for the measure of consistency, the single measure intra-class correlation is .200, with a 95% confidence interval between .077 and .572. For both measures, *F* (6, 192) = 9.246, *p* < .001; the Cronbach's Alpha (reliability) is .892 (N of raters = 7; N of items = 33).

Again, there was no relationship between the teacher's teaching experience and his or her rating on indicators: *r* = -.011, *p* = .97. Both the relationship between teachers' teaching experience and their ratings on item sensitivity and the relationship between teachers' teaching experience and their ratings on indictor clarity were negative. However, the relationship between the ratings on item sensitivity and those on indicator clarity is positive and statistically significant: *r* = .571, *p* = .04.

Item Assignment to Indicators

In the third part of the item review, both the test items and the descriptions of indicators were presented to the teachers. Teachers were instructed to assign the test items under the appropriate indicators based on the descriptions of the indicators. It was made clear to the teachers that, under one indicator, there could be multiple items, but one item should not be assigned to more than one indicator. For four out of 35 items, all 13 teachers assigned them to the right indicators. These items are Items 7, 15, 17 and 34. For another nine items, 12 out of 13 teachers assigned them to the right indicators. For another eight items, 11 out of 13 teachers assigned them to the right indicators. Out of 13 teachers, only five of them assigned Item 1 to the right indicator. Table 17 lists the frequencies of correct assignment for each item. It should be noted that both Items 1 and 34 belong to the same indicator; all teachers assigned Item 34 correctly, but only five teachers (the fewest) assigned Item 1 correctly. Despite this, Item 1 was rated as the most sensitive item by the teachers.

Table 17. Frequencies of Correct Assignment for Each Item

Items	34	17	7	15	3	23	31	21	19	24	22	26	32	4	28	18	2
Frequency	13	13	13	13	12	12	12	12	12	12	12	12	12	11	11	11	11

(Table 17 continued)

Items	16	9	14	25	33	13	20	12	30	10	27	35	5	29	8	11	6	1
Frequency	11	11	11	11	10	10	10	9	9	9	9	9	8	8	7	7	7	5

When comparing the order of teachers' consensus on item assignment to the orders of degrees of item sensitivity both from the empirical methods and the judgmental approach, there was no relationship between the degree of item sensitivity and the likelihood of assigning an item to the right indicator. However, Item 34 was not only rated as the most sensitive item by the empirical methods but also correctly assigned to the right indicator by all the teachers. In addition, Item 7 was also both ranked very high in sensitivity and shared consensus from all the teachers on item assignment.

When looking at the relationship between teaching experience and correct item assignments, although there was no strong relationship between more experienced teachers and more correct item assignments (r = .493, *p* = .087), descriptive statistics showed 1) the more experienced the teachers were, the more likely they assigned the items to the right indicators and, 2) the more experienced the teachers were, the more likely they reached a consensus on the item assignments. Table 18 presents the number of items (out of 35) assigned correctly and standard deviation of correct assignments after data were divided into three groups based on teachers' teaching experience in mathematics.

Table 18. Years of Teaching and Correct Item Assignments

Years Teaching Math	N	Min.	Max.	Mean	SD
Year >20	2	31	34	32.50	2.121
10 ≤ Year <20	5	26	34	30.20	2.863
Year < 10	6	15	33	26.83	6.462

As shown in Table 18, two teachers had more than 20 years of teaching experience in mathematics. One of them assigned 31 out of 35 items correctly, and the other had 34 items assigned correctly. Five teachers had 10 or more but less than 20 years of experience in mathematics. The one who had the least correct item assignments had 26 items assigned correctly; and the one who had the most correct assignments had 34 items assigned correctly. There were no big differences in correct assignments between these two groups in terms of the mean score ($\bar{X}$ = 32.5 vs. $\bar{X}$ = 30.2). Teachers who had 10 or more but less than 10 years of teaching experience had more dispersed scores (SD = 2.863) than those with more than 20 years of teaching experience (SD = 2.121). However, there was a big difference between teachers who had less than 10 years of teaching experience and those who had 10 or more years of teaching experience in terms of correct item assignments. On average, the six teachers in the less experienced group only had about 27 items ($\bar{X}$ = 26.83) assigned correctly. Furthermore, there was a larger dispersion in their item assignments (SD = 6.462).

In this sample, nine teachers were teaching in public schools, and the other four were teaching or had taught in private schools. However, all four private school teachers were familiar to some degree with the indicators presented. Descriptive statistics show that teachers from public schools had a higher rate of correct item assignments than those from private schools. It may be because teachers from public schools were comparatively more familiar with the indicators. It may also be due to the fact that three of the four private schools teachers only had about one year of teaching experience. Additionally, responses from private school teachers were more dispersed. Table 19 presents information of correct item assignments from teachers when divided based on the type of schools where they were teaching.

Table 19. Type of Schools and Correct Assignments

Type of School	N	Min.	Max.	Mean	SD
Public	9	26	34	30.33	2.693
Private	4	15	33	26.00	8.083

Comparison of Empirical Methods and Judgmental Approach

Although results produced from both empirical methods highly agreed with each other (*r* = .93), teachers' judgments were not highly related with the results from either of the empirical methods. Correlations between teachers' judgments and the empirical methods were medium: *r* = .37 for judgmental approach and MH tests, and *r* = .32 for judgmental approach and LR procedures.

As far as the instructionally sensitive items detected, the MH tests and the LR procedures detected 19 identical items, with an exception that the LR procedure detected one more sensitive item. Although the orders of the sensitive items, when sorted according to their degrees of sensitivity, from the two empirical methods were not exactly the same, they were very close or similar. As for the judgmental approach, it is hard to identify an item simply as sensitive or as insensitive because the items' sensitivity was estimated by a continuum (i.e., an 11-point Likert scale was used with "0" representing totally insensitive and "10" representing totally sensitive). For the purpose of comparison, the first 20[8] items with the highest mean scores on sensitivity

[8] Items 13 and 30 had the same mean score of rating on sensitivity; both were ranked the 19th. Therefore, twenty, instead of nineteen, items were used for comparison to the sensitive items detected by the empirical methods.

rated by teachers were selected to be compared with the 19 sensitive items detected by empirical methods. Table 20 lists items detected as instructionally sensitive by both empirical methods and the judgmental approach. Items in this table were ordered based on their degrees of sensitivity according to different detection methods.

Table 20. Sensitive Items Detected by Empirical Methods and Judgmental Approach

	More Sensitive																		Less sensitive
LR	34	20	32	31	7	16	3	15	6	30	1	14	13	23	9	19	35	12	2
MH	34	20	32	7	16	31	3	13	15	6	30	14	9	23	1	12	19	35	2
Judgmental	1	15	7	31	34	11	4	2	3	10	20	5	35	8	19	17	6	21	30/13

Note: Items 30 and 13 had the same mean score of rating on sensitivity.

A close look reveals that 13 out of the 20 items that were ranked most sensitive using judgmental approach were also detected as instructionally sensitive by the two empirical methods. These thirteen items were: Items 1, 2, 3, 6, 7, 13, 15, 19, 20, 30, 31, 34, and 35. However, the ranking of sensitivity by teachers was very different from the ranking produced by the statistical methods.

The topics identified as sensitive to instruction by empirical and judgmental approaches are different, to some degree. The results of the empirical methods demonstrate that items belonging to the *Geometric Figures and Their Properties*, *Measurement and Estimation*, and *Statistics* topics were more likely to be sensitive. Table 21 lists the content areas tested by the detected sensitive items. Based on the information under Column “Percent Sensitive,” 100% of items under the benchmark of *Geometric Figures and Their Properties* were detected as instructionally sensitive; 83% items under the benchmark of *Measurement and Estimation* were detected as instructionally sensitive; and 60% items under the benchmark of *Statistics* were detected as instructionally sensitive. To compare results from the empirical methods and those from the judgmental approach, the 20 most sensitive items based on teachers' ratings on sensitivity were selected and examined to determine what common characteristics they had. Results of the judgmental approach demonstrate that items testing topics of *Number Sense, Computation*, *Measurement and Estimation*, and *Variable, Equations and Inequalities* were more likely to be thought of as sensitive to instruction. Table 22 lists the content areas tested by the items detected as instructionally sensitive by the judgmental approach.

Table 21. Topics Tested by Sensitive Items Detected by Empirical Methods

Standards	Benchmarks	Percent Sensitive	Indicators	Items
Standard 1 Number and Computation	**Benchmark 1 Number Sense**	20%	Ind. A1a: Generates and/or solves real-world problems using equivalent representations of rational numbers and simple algebraic expressions.	12
	Benchmark 4 Computation	50%	Ind. K2: performs and explains these computational procedures: d: adds, subtracts, multiplies, and divides fractions and expresses answers in simplest form.	13
			Ind. K5: finds percentages of rational numbers	3, 23
Standard 2 Algebra	**Benchmark 1 Patterns**	50%	Ind. K4: states the rule to find the nth term of a pattern with one operational change (addition or subtraction) between consecutive terms.	6, 30
	Benchmark 2 Variable, Equations, and Inequalities	43%	Ind. K7: knows the mathematical relationship between ratios, proportions, and percents and how to solve for a missing term in a proportion with positive rational number solutions and monomials.	20, 31
			Ind. A1: represents real-world problems using variables and symbols to write linear expressions, one- or two-step equations.	19
Standard 3 Geometry	**Benchmark 1 Geometric Figures and Their Properties**	**100%**	Ind. K3: identifies angle and side properties of triangles and quadrilaterals: d. rectangles have angles of 90°, opposite sides are congruent;	2
			g. trapezoids have one pair of opposite sides parallel and the other pair of opposite sides are not parallel.	16
	Benchmark 2 Measurement and Estimation	**83%**	Ind. K4: knows and uses perimeter and area formulas for circles, squares, rectangles, triangles, and parallelograms.	1, 34
			Ind. K6: uses given measurement formulas to find b. volume of rectangular prisms.	9, 14
			Ind. A1c: solves real-world problems by finding perimeter and area of two-dimensional composite figures of squares, rectangles, and triangles.	35
Standard 4 Data	**Benchmark 2 Statistics**	**60%**	Ind. K1: organizes, displays, and reads quantitative (numerical) and qualitative (non-numerical) data in a clear, organized, and accurate manner including a title, labels, categories, and rational number intervals using these data displays g. box-and-whiskers plots.	7,15
			Ind. A3: recognizes and explains a. misleading representations of data.	32

Note: "Percent Sensitive" refers to the ratio of the number of items detected as sensitive to the number of items used under a certain benchmark. Percentages for the most sensitive topics are in bold.

Table 22. Topics Tested by Sensitive Items Detected by Judgmental Approach

Standards	Benchmarks	Percent Sensitive	Indicators	Items
Standard 1 Number and Computation	**Benchmark 1** **Number Sense**	**67%**	Ind. A1a: Generates and/or solves real-world problems using equivalent representations of rational numbers and simple algebraic expressions.	11,8
	Benchmark 4 **Computation**	**67%**	Ind. K2: performs and explains these computational procedures: b: multiplies and divides a four-digit number by a two-digit number using numbers from thousands place through thousandths place;	5
			c: multiplies and divides using numbers from thousands place through thousandths place by 10; 100; 1,000; .1; .01; .001; or single-digit multiples of each;	4
			d: adds, subtracts, multiplies, and divides fractions and expresses answers in simplest form.	13
			Ind. K5: finds percentages of rational numbers	3
Standard 2 Algebra	**Benchmark 1** **Patterns**	50%	Ind. K4: states the rule to find the nth term of a pattern with one operational change (addition or subtraction) between consecutive terms.	6, 30
	Benchmark 2 **Variable, Equations, and Inequalities**	**71%**	Ind. K7: knows the mathematical relationship between ratios, proportions, and percents and how to solve for a missing term in a proportion with positive rational number solutions and monomials.	20, 31
			Ind. K8: evaluates simple algebraic expressions using positive rational numbers.	10, 21
			Ind. A1: represents real-world problems using variables and symbols to write linear expressions, one- or two-step equations.	19
Standard 3 Geometry	**Benchmark 1** **Geometric Figures and Their Properties**	50%	Ind. K3: identifies angle and side properties of triangles and quadrilaterals: d. rectangles have angles of 90°, opposite sides are congruent;	2
	Benchmark 2 **Measurement and Estimation**	**60%**	Ind. K4: knows and uses perimeter and area formulas for circles, squares, rectangles, triangles, and parallelograms.	1, 34
			Ind. A1c: solves real-world problems by finding perimeter and area of two-dimensional composite figures of squares, rectangles, and triangles.	35
	Benchmark 3 **Transformational Geometry**	50%	Ind. A3: determines the actual dimensions and/or measurements of a two-dimensional figure represented in a scale drawing	17
Standard 4 Data	**Benchmark 2** **Statistics**	20%	Ind. K1: organizes, displays, and reads quantitative (numerical) and qualitative (non-numerical) data in a clear, organized, and accurate manner including a title, labels, categories, and rational number intervals using these data displays g. box-and-whiskers plots.	7,15

Note: "Ratio" refers to the ratio of the number of items detected as sensitive to the number of items used under a certain benchmark. Percentages for the most sensitive topics are in bold.

Summary

The purpose of this chapter was to present the results of instructionally sensitive items detected by using Mantel-Haenzsel tests, logistic regression and the judgmental method, to compare items detected by the three different methods, and to examine the degree of congruity among the three methods. To analyze the data of students' performance, descriptive statistics, frequency analysis, Mantel-Hanzsel tests, and logistic regression were utilized. To analyze the data of teachers' judgments, descriptive statistics, frequency analysis, and bivariate correlations were used. Bivariate correlations were also used to compare the results obtained from the three different methods. Based upon the results of these analyses, the next chapter discusses the results and presents implications of this study.

CHAPTER 5

DISCUSSION

This chapter interprets and discusses the findings in relation to the literature and presents implications for educational policy and practice, suggestions for future research, and limitations of the study.

Discussion of Results

The most important findings in this study include:

1) Judgmental methods and empirical methods are not highly correlated. Although the Mantel-Haenszel and logistic regression analyses are highly correlated, neither is highly correlated with teacher judgments. It appears teachers, at least when they receive the minimal training provided in this study, are not capable of making the required judgments regarding instructional sensitivity.
2) About half of the items in the test were detected as instructionally sensitive by empirical methods, but there are still many items that are not instructionally sensitive.

Comparison of MH tests and LR Procedure

The results produced by the MH tests and those by the LR procedure highly agreed with each other. Both methods identified the same 19 items as most instructionally sensitive. The effect sizes from both methods were highly correlated (r = .93). The strong relationship between the MH effect sizes and the LR effect sizes indicates that the rankings of degrees of sensitivity by both methods were very similar.

Despite the identicalness in items detected and the similarity in sensitivity rankings between the two methods, the logistic regression procedure is preferred and recommended. First, one of the disadvantages of the MH test is that it is less powerful in identifying non-uniform DIF (Rogers & Swaminathan, 1993). Results in this study supported Rogers' and Swaminathan's argument. Among the 20 sensitive items detected as statistically significant by the LR procedure, nineteen of them were sensitive due to the interaction. However, among the 19 sensitive items detected by the MH tests, only three were sensitive due to the interaction. Although Hidalgo and López-Pina (2004) proposed the modified MH procedure and reported similar power in detecting non-uniform DIF that the modified MH procedure has as the LR procedure, the modified MH procedure cannot be applied to this study because each student's membership changed across the items. Additionally, the LR procedure keeps the matching criterion (i.e., the proficiency θ) continuous, while the MH test has to "chop" the continuous variable into categories.

How Sensitive Is the Kansas Mathematics Interim Assessment?

Using the Second International Mathematics Study (SIMS), the U.S. eighth grade sample, Kao (1990) and Lehman (1986) implemented MIMIC models in their studies and only detected eight (5%) items sensitive to instruction. Pham (2009) used two methods (IRT models for the real TAKS data and a simulation study) to test whether the Texas Assessment of Knowledge and Skills (TAKS) exam measures student achievement in four different domains. The combined results of the two methods provided compelling evidence that the TAKS is instructionally insensitive. Compared to their findings, the number of sensitive items found in this study is considerably larger; about 54% of items were sensitive to instructions. This may suggest that students' performance on the seventh grade Kansas Interim Assessment in mathematics was influenced by differential instructional coverage. For the test-takers in this sample, the variation of their scores reflects their varied instructional experience more than general mathematical proficiency.

However, about half of the items still were not sensitive to instruction. Therefore, the questions will be: How much sensitivity would be enough for an achievement test? What would be the acceptable percentage of the detected sensitive items in a test? According to Popham (2003b), the number of sensitive items in a test is not the only criterion to conclude whether a test is instructionally sensitive or insensitive: “Instructional sensitivity is a necessary, but not sufficient condition for an NCLB test that’s going to benefit students” (p.7). He further emphasized that once the state’s NCLB tests’ instructional sensitivity has been assured, it is important to make sure that the tests measure significant skills and knowledge that children ought to be mastering.

Sensitive items detected by the empirical methods indicate that the *Geometric Figures and Their Properties*, *Measurement and Estimation*, and *Statistics* topics were more likely to be sensitive. In contrast, sensitive items detected by the judgmental approach indicate that topics of *Number Sense*, *Computation*, *Measurement and Estimation* and *Variable, Equations and Inequalities* were more likely to be sensitive to instruction. The only common topic detected by both methods is *Measurement and Estimation*. The exact reason for these topics or indicators to be more sensitive is not clear. However, one possible reason may be the clarity of the indicators. The three attributes to an instructionally sensitive test, proposed by Popham (2003b), are: 1) clarity of assessment targets, 2) a manageable number of assessment targets, and 3) instructionally informative results. If the indicators are stated with sufficient clarity, almost all the state’s teachers will be able to identify what the indicators really meant. Clearly described indicators can contribute to the teacher's instructional planning and effective instruction. In turn, the impact of effective instruction will be reflected by the items belonging to these indicators, if the items are instructionally sensitive.

Disappointingly, results in this study did not show strong association between sensitivity of items and clarity of indicators. The indicators that are more likely to be sensitive are not the ones ranked the highest on clarity by teachers. Generally speaking, most of these "sensitive" topics received more positive than negative comments on clarity, for example, Ind. K2b, Ind. K2c, and Ind. K2d (Standard I – Benchmark 4). Comments on the above indicators supported Popham's (2005) argument that an instructionally supportive accountability test should provide clear descriptions of what is to be assessed. The fact that, on average, items belonging to clearly stated indicators were likely to be sensitive confirms, though not convincingly, one of the rules proposed by Popham (2001), which ensures standards-based assessment to make meaningful contributions to improved instructional quality:

> Rule3: Create a sufficiently clear description of the knowledge and/or skills represented by the test so that teachers will have an understanding of the cognitive demands required for students' successful performance. (p. 6)

Popham (2003b) further states, "If an NCLB test contends that it measures 30 or 40 curricular targets, teachers will be unable to focus their instructional plans properly. An NCLB test that purports to measure too many content standards (or benchmarks, etc.) is certain to be instructionally insensitive" (p. 4). The items in Testing Window Two used for this study covered 15 indicators, with at least one item and at most three items under each indicator. Thus, the assessment targets are reasonable and manageable.

Comparison of Empirical and Judgmental Methods

The instructional sensitivity judgments made by curriculum experts did not agree with the results of empirical methods. The relationships between teachers' judgments and results obtained from the empirical methods were not strong. The correlation between teachers' judgments and results from the MH tests was r = .37; the correlation between teachers' judgments and results from the LR procedure was r = .32. The moderate relationship between empirical methods and judgmental methods indicates that when examining instructional sensitivity of a test and its items, curricular decisions should not merely rely on curriculum experts' judgment. It is advisable to make decisions based on results obtained from both methods.

Despite its major advantage, pointed out by Popham (2010b), that it can be implemented without great cost and biased items can be discarded before being used on an operational form of a significant test, the major disadvantage of the judgmental method is the subjectivity of the teachers' or curriculum experts' judgments. The data analysis procedures in this study show that adding judgments from one single teacher into the dataset often dramatically changed the results. The intra-class correlations among raters on sensitivity ratings were very low: r = .155 (when measures of absolute agreement were applied) and r = .182 (when measures of consistency were applied). The low intra-class correlation coefficients indicate poor agreement among raters. The intra-class correlations among raters on indicator clarity ratings were comparatively high: r = .339 (when measures of absolute agreement were applied) and r = .371 (when measures of consistency were applied). These intra-class correlation coefficients indicate better agreement among raters on indicator clarity ratings than on instructional sensitivity. The possible reasons why the judgmental approach did not provide consistent results in this particular study include: 1) limited training was provided to teachers regarding judging items as to their sensitivity to good instruction; 2) teachers were not randomly sampled to represent the larger population; 3) some teachers were not familiar enough with the state's curricular standards; and 4) other factors (e.g., teaching experience) that may influence teachers' judgments were not controlled for.

Due to the findings in this study that neither the intra-class correlation among raters on sensitivity rating nor the intra-class correlation on indicator clarity ratings reached moderate agreement (r = .5 ~ .6), the judgmental approach should be used with caution. Meticulous consideration and proper implementation are two elements to ensure the efficiency of the judgmental approach:

> My focus on empirical methodological issues does not diminish my belief that if judgmental detection of instructionally insensitivity is well conceived and properly implemented, such approaches can economically and efficiently increase the instructional sensitivity of important tests such as the accountability assessments now used throughout the United States. (Popham, 2010b, p. 4)

Future Research

This study suggests directions for future research. First, the OTL data can be collected and coded in a different way. The teachers' estimate of OTL information was used in this study, as it was in most previous studies. However, the shortcoming of teachers' information is that it may not reflect the fact that students in the same classroom may have different learning experiences. Students' OTL information may better reflect learning than teaching. In addition, OTL was typically measured as a dichotomous variable, when content coverage in teaching was the concern. However, when students' differences in motivation, cognitive ability and outside classroom experiences are considered, OTL can be an ordinal or a continuous variable. Therefore, it is worthwhile to explore different methods of collecting and coding OTL data in future studies.

Second, this study can be expanded from its focus on dichotomous items to polytomous items. Many studies (Kao, 1990; Kim, 1990; Lehman, 1986; Switzer, 1993; Yu, 2006), including this current study, on instructional sensitivity used multiple-choice items. However, many test items are open-ended questions requiring students to construct their own responses. In this case, ordinal grading will be applied, and the score for a given item will range from zero to the highest possible point. Although it will be more complicated to identify instructional sensitivity of items with ordinal grades, it is of importance to explore and develop methodologies to examine how sensitive the constructive response items are to instruction. Otherwise, educational specialists can hardly know whether tests containing open-ended questions are accountable in measuring how well students have been taught.

Third, results show that items testing topics of *Geometric Figures and Their Properties*, *Measurement and Estimation*, and *Statistics* are more likely to be sensitive to instruction than items testing other topics. Results also show that no strong association exists between sensitivity of items and clarity of indicators. In the future, this may be of some interest to explore the relationship between the detected sensitive items and the format of these items. The wording of the items may be a factor influencing items' sensitivity to instruction.

Further, it is common that students are nested in classes and classes are nested in schools. This nested nature of data calls for a construction of multilevel or hierarchical models. A mixed model, including measurements at the student level, at the classroom level, and at an even higher level, can be used to study instructional sensitivity. Thus, not only the relationship between students' performance and their instructional experience in the context of a specific classroom or school can be explored, but also the sample's representativeness of a larger population can be addressed.

Last, it should also be noted that the instruction's influence and effects will be more outstanding when it is accumulated. After the administration of the interim assessment, the teachers were asked to only check those indicators they taught in the current year prior to the assessment. It is possible that certain content was taught in previous years but not in the current year, and the teacher did not check it. However, the instruction in previous years would still affect students' performance. It is also possible that for certain content checked as taught in the current year, some students were only exposed to this content in the current year; but some were also exposed to the same content in previous years as well. In this case, the accumulative influence of instruction on students' performance should be taken into account. Future studies could control for the impact of instruction from the previous years, if there is any.

Policy Implications

The study offers some important policy and practice implications for educators and state assessment officials. First, although about half of the items in Kansas Interim Assessment on seventh grade mathematics were detected as instructionally sensitive, many other items were not sensitive. A further recognition of instructional insensitivity of many accountability tests needs to be achieved. If instructional sensitivity as a feature of tests or items is ignored, it will threaten the validity of many decisions that are made annually based on results from state assessments under NCLB (Polikoff, 2010). "Policymakers . . . should think about ways to ensure that NCLB and other standards-based assessments are actually sensitive to instruction" (Polikoff, 2010, p. 13). Also, test developers should attend to the common characteristics of the sensitive items detected in this study and develop better items that reflect the instruction received by students.

Second, both the low correlation between the judgmental and empirical methods and the subjectivity of the judgmental method indicate the necessity of employing empirical techniques in the detection of sensitive items. Although the judgmental approaches can be implemented before the items are used on a significant test, with acceptable cost, they are believed to lack the objectivity of more statistical approaches (Polikoff, 2010). Although the empirical methods cannot be implemented before the administration of the test to filter out potentially insensitive items, the examination of the sensitive items detected by these methods can provide important insights and useful information to test developers to create better items. The employment of both judgmental and empirical approaches can maximize favorable factors and minimize unfavorable ones. Accordingly, the state assessment officials are recommended to evaluate schools and teachers based on the full consideration of results produced from both methods. However, considering the reduction of cost (i.e., time and/or budget), empirical methods are advocated if the employment of both judgmental and empirical methods is not possible.

Limitations

Due to the non-experimental nature of the data and design, this study has several limitations that need to be considered:

First, the results reported here are limited in that a convenient sample of participants for the judgmental approach was used. Judgments made by this sample can hardly be generalized to a wider situation. Another convenient sample may provide different results. The sample's lack of representation of a larger population should be considered in interpreting the findings. Replication of this study with a variety of diverse samples will continue to strengthen the reliability and validity evidence for judgmental approaches. In addition, the training provided to teachers regarding item review for instructional sensitivity detection was limited. More training and practice should be provided to teachers before the actual review.

Second, more school and teacher factors need to be examined in order to understand the influences of instruction on students' performance. The factor examined in this study was limited to content coverage in the current year only. More factors, such as content coverage in previous years and extracurricular tutoring, are needed to get a better understanding of the relationship between instruction received and students' performance. Further, the items used in this study were limited to seventh grade math items only. Studies of this nature are encouraged to utilize items from other grades and subjects to provide more generalizable results.

Third, due to the fact that the student's membership (i.e., the instructed group or uninstructed group) changed over items, it was impossible to include all the test items in one model at one time to detect item sensitivity. One single item was analyzed at one time. For this reason, the MIMIC model that adjusts for measurement errors, but also requires all the test items to be included in one model to construct the baseline model, was not employed in this study. As a result, some relevant variables, such as demographic variables, were not controlled for as covariates in the measurement model used in the study. Also, it was impossible to detect the indirect influence from the grouping variable on an item through the latent variable (θ). Only the direct influence from the grouping variable on an item could be detected.

Lastly, this study only used items given to the students assigned to the top two pathways based on their performance in the previous sections in this multi-stage adaptive design computer test. The mathematics proficiency (θ) of these students was comparatively high: $-1.234 \leq \theta \leq 4$. Compared to the normal distribution of the θ with a range from -4 to 4, the restriction of the range of students' proficiency may have led to underestimation of the results.

Conclusion

The number of sensitive items found by the empirical methods in this study is large compared to previous studies. About 54% of items are sensitive to instruction. This may suggest that students' performance on the seventh grade Kansas Interim Assessment in mathematics is influenced by differential instructional coverage. For the test-takers in this sample, the variations of their scores reflect less general mathematics proficiency than that of their varied instructional experience. Further, results also show that all the detected items are in favor of the instructed group. In other words, students from the instructed group had a higher probability of succeeding on each of the detected items than those from the uninstructed group, after they were matched on proficiency. This finding indicates that the item performance differences are due to differences in curricular content covered in instruction, and instruction positively influences students' performance.

Although both the Mantel-Haenzsel tests and the logistic regression procedure detected the same items as instructionally sensitive, the logistic regression procedure is recommended by the researcher for two reasons. First, the LR procedure is more powerful in detecting items sensitive due to the interaction of students' instructional experience and their proficiency. Second, the LR procedure keeps the matching variable, student's proficiency, continuous, while the MH approach categorized this continuous variable.

The judgmental method is very subjective. Teachers had disagreement on most of the items, and their ratings were very dispersed on these items. For example, both Items 1 and 34 belong to the same indicator; all the teachers assigned Item 34 to the indicator correctly, but only five teachers assigned Item 1 to this indicator correctly. Item 1 was rated as the most sensitive item by the teachers but was assigned to the correct indicator by the fewest teachers. Adding one piece of information (i.e., judgments from one single teacher) would dramatically influence the results. For example, judgments from 11 teachers showed that 24 out of 35 items were rated above Point 7 (the mean value). When the sample size increased from 11 to 13, only 22 out of 35 items were rated above Point 7. In addition, the rankings of the items were also different when the sample sizes were different. Further, the relationships between the results obtained from the judgmental and the empirical methods changed considerably when the sample sizes were different: when n = 11, $r_{\text{judgmental \& MH}}$ = .32 and $r_{\text{judgmental \& LR}}$ = .28; when n = 13, $r_{\text{judgmental \& MH}}$ = .37 and $r_{\text{judgmental \& LR}}$ = .32. The moderate relationship between the judgmental method and the empirical methods also indicates that the judgmental method should be used with delicacy, and the instructional sensitivity decisions should not be made merely based on curricular experts' judgments.

Considering the significantly positive relationship between students' instructional experience and their performance on most test items, educators and policy makers may emphasize the importance of bolstering effective instruction and developing sensitive items. It is worthwhile to discover how to build accountability tests that will be instructionally sensitive, so that they can provide valid inferences about effective and ineffective instruction.

Reference

Airasian, P. W., & Madaus, G. F. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement, 20*(2), 103-118.

Baker, E. L. (2008). Empirically determining the instructional sensitivity of an accountability test: UCLA/CRESST.

Birnbaum, A. (1986). Test scores, sufficient statistics, and the information structure of tests. In F. M. Lord and M. R. Novick, *Statistical theories of mental test scores*, pp. 425-235. Reading, MA: Addison-Wesley.

Calfee, R. (1983). Establishing instructional validity for minimum competency programs. In G. F. Madaus& D. L. Stufflebeam (Eds.), *The Courts, Validity, and Minimum Competency Testing*. Hingham, MA: Kluwer-Nijhoff.

Clauser, B. E., &Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issue and Practice, 17*, 31-44.

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. Newbury Park, CA: Sage.

Cohen, D. K., & Hill, H. C. (1998). *Instructional policy and classroom performance: The mathematics reform in California* (No. CPRE-RR-39). Philadelphia: Consortium for Policy Research in Education.

Court, S. C. (2010). *Instructional Sensitivity of Accountability Tests: Recent Refinements in Detecting Insensitive Items*. Paper presented at the Council of Chief State School Officers' National Conference on Student Assessment, Detroit, MI.

Cox, R. C., & Vargas, J. S. (1966). *A comparison of item selection techniques for norm-referenced and criterion-referenced tests* (No. BR-5-0253-REPRINT-7). Pennsylvania: Learning Research and Development Center, Pittsburgh University.

D'Agostino, J. V., Welsh, M. E., & Corson, N. M. (2007). Instructional sensitivity of a state's standards-based assessment. *Educational Assessment*, *12*(1), 1-22.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives*, *8*(1), 1-44.

Educational Commission of the States. (2002). No Child Left Behind Issue Brief: A guide to standards-based assessment. Retrieved May 15, 2011 from http://www.ecs.org/clearinghouse/35/50/3550.pdf.

Fenwick, F. J. (2001). Using student outcomes to evaluate teaching: A cautious exploration. *New Directions for Teaching and Learning*, 63-74.

Jacob, B. A., & Lefgren, L. (2004). The impact of teacher training on student achievement: Quasi-Experimental evidence from school reform efforts in Chicago. *The Journal of Human Resources, 39*(1), 50-79.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rate using an effect size measure with the logistic regression procedure for DIF detection *Applied Measurement in Education*, *14*(4), 329-349.

Gierl, M. J., Khaliq, S. N., & Keith, B. (1999). *Gender differential item functioning in Mathematics and Science: Prevalence and Policy Implications*. Paper presented at the Annual Meeting of Improving Large-Scale Assessment in Education.

Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Gordon, J. V. (2008). *Performance on large-scale science tests: Item attributes that may impact achievement scores.* Unpublished Dissertation, Montana State University.

Haladyna, T. M. (1974). Effects of different samples on item and test characteristics of criterion-reference tests. *Journal of Educational Measurement, 11*(2), 93-99.

Haladyna, T., & Roid, G. (1981). The role of instructional sensitivity in the empirical review of criterion-referenced test items. *Journal of Educational Measurement, 18*(1), 39-53.

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

Hanna, G. S., & Bennett, J. A. (1984). Instructional sensitivity expanded. *Educational and Psychological Measurement, 44*(3), 583-596.

Helmstadter, G. C. (1974). *A comparison of Bayesian and traditional indexes of test item effectiveness.* Paper presented at the National Council on Measurement in Education.

Herman, J. L., & Klein, D. C. D. (1997). Assessing Opportunity to Learn: A California Example. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Hidalgo, M. D., & López-Pina. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903-915.

Kao, C.-F. (1990). *An investigation of instructional sensitivity in mathematics achievement test items for U.S. eighth grade students.* Doctor of Philosophy Dissertation, University of California, Los Angeles.

Kim, S.-W.(1990). *Gender and OTL effect on mathematics achievement for U.S. SIMS 12th grade students.* Doctor of Philosophy Dissertation, University of Kansas, Los Angeles.

Lehman, J. D. (1986). *Opportunity to learn and differential item functioning.* Doctor of Philosophy Dissertation, University of California, Los Angeles.

Linn, R. L. (1983). Curricular validity: Convincing the courts that it was taught without precluding the possibility of measuring it. In G. F. Madaus & D. L. Stufflebeam (Eds.), *The Courts, Validity, and Minimum Competency Testing*. Hingham, MA: Kluwer-Nijhoff.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22,* 719-748.

Mazor, K. M., Clauser, B. E. & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement, 54*, 284-291.

Mehrens, W. A., & Phillips, S. E. (1987). Sensitivity of item difficulties to curricular validity. *Journal of Educational Measurement, 24*(4), 357-370.

Mehrens, W. A., & Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. *Journal of Educational Measurement, 23*(3), 185-196.

Miller, M. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement*, *25*(3), 205-219.

Muthén, B. O. (1989). Using item-specific instructional information in achievement modeling. *Psychometrika*, *54*(3), 385-396.

Muthén, B. O. (1988a). *Instructionally sensitive psychometrics: Applications to the second international mathematics study* (No. CSE 286): UCLA Center for Research on Evaluation, Standards, and Student Testing.

Muthén, B. O. (1988b). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 213-238). Hillsdale, NJ: Erlbaum.

Muthén, B. O. (1987). *Using item specific instructional information in achievement molding* (No. 143). Los Angeles, CA: Center for the Study of Evaluation; Center for Research on Evaluation, Standards, and Student Testing.

Muthén, B. O., Kao, C.-F., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28*(1), 1-22.

Mehréns, W. A., & Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. *Journal of Educational Measurement, 23*(3), 185-196.

Niemi, D., Wang, J., Steinberg, D. H., Baker, E. L., & Wang, H. (2007). Instructional sensitivity of a complex language arts performance assessment. *Educational Assessment, 12*(3&4), 215-237.

Perkins, K. (1984). *A comparison of instructional sensitivity indices* (No. 143).

Penfield, R. D. (2007). *Differential item functioning analysis system (DIFS 4.0) user's manual*.

Peyton, V. D. (2000). *Differential item functioning of the parental modernity inventory: A method comparison.* Unpublished Dissertation, University of Kansas, Lawrence.

Pham, V. H. (2009). *Computer modeling of the instructionally insensitive nature of the Texas Assessment of Knowledge and Skills (TAKS) Exam.* PhD Dissertation, The University of Texas at Austin, Austin, Texas.

Phillips, S. E., & Mehrens, W. A. (1988). Effects of curricular differences on achievement test data at item and objective levels. *Applied Measurement in Education, 1*(1), 33-51.

Phillips, S. E., & Mehrens, W. A. (1987). Curricular differences and unidimensionality of achievement test data: An exploratory analysis. *Journal of Educational Measurement, 24*(1), 1-16.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement, Issue and Practice, 29*(4), 3-14.

Popham, W. J. (2010a). Instructional sensitivity. In W. J. Popham (Ed.), *Everything school leaders need to know about assessment*. Thousand Oaks, CA: Sage.

Popham, W. J. (2010b). *Empirically snaring instructionally insensitive items*. Paper presented at the American Educational Research Association, Denver, Colorado.

Popham, W. J. (2007a). Accountability tests' instructional insensitivity: The time bomb ticketh. *Education Week*. Retrieved from http://www.edweek.org/login.html?source=http://www.edweek.org/ew/articles/2007/11/14/12popham.h27.html&destination=http://www.edweek.org/ew/articles/2007/11/14/12popham.h27.html&levelId=2100

Popham, W. J. (2007b). *Instructional insensitivity of tests: Accountability's dire drawback.* Paper presented at the American Educational Research Association, Chicago, Illinois.

Popham, W. J. (2006). *Determining the instructional sensitivity of accountability tests.* Paper presented at the Large-Scale Assessment Conference, San Francisco, California.

Popham, W. J. (2003a). Living with your NCLB tests: schools' ability to meet expectations will depend on tests' instructional sensitivity - or dying. *School Administrator*. Retrieved from http://findarticles.com/p/articles/mi_m0JSD/is_11_60/ai_111270454/

Popham, W. J. (2003b). *Are your state's NCLB tests instructionally insensitive? Here's how to tell!* Paper presented at the National School Board Association.

Popham, W. J. (2001). *Standards-based assessment: Solution or charade?* Paper presented at the American Educational Research Association, Seattle, Washington.

Popham, W. J. (1995). *Classroom assessment: What teachers need to know*. Needham Heights, Massachusetts: Allyn and Bacon.

Popham, W. J., Keller, T., Moulding, B., Pellegrino, J., & Sandifer, P. (2005). Instructionally supportive accountability tests in Science: A viable assessment option? *Measurement, 3*(3), 121-179.

Popham, W. J., & Kaase, K. (2009). *Detecting instructionally insensitive items in accountability tests: Methodological advances*. Paper presented at the Association for Educational Assessment, Malta, Europe.

Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*(1), 1-9.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 105-116.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement, 33*(2), 215-230.

Shannon, G. A., & Cliver, B. A. (1987). An application of item response theory in the comparison of four conventional item discrimination indices for criterion-referenced tests. *Journal of Educational Measurement, 24*(4), 347-356.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika* 58, 159-194.

Simpson, R. L., LaCava, P. G., & Graner, P. S. (2004). The No Child Left Behind Act: Challenged and implications for educators. *Intervention in School and Clinic, 40*(2), 67-75.

Swaminathan, H., & Rogers, H. J. (1990).Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Switzer, D. M. (1993). *Differential item functioning and opportunity to learn: Adjusting the Mantel-Hansel Chi-square procedure.* Doctor of Philosophy, University of Illinois, Urbana-Champaign.

van der Linden, W. J. (1981). A latent trait look at pretest-posttest validation of criterion-referenced test items. *Review of Educational Research, 51*(3), 379-402.

Walker, D. F. (1983). What constitutes curricular validity in a high-school-leaving examination? In G. F. Madaus& D. L. Stufflebeam (Eds.), *The Courts, Validity, and Minimum Competency Testing*. Hingham, MA: Kluwer-Nijhoff.

Wiley, D. E., & Yoon, B. (1995). Teacher reports on opportunity to learn: Analyses of the 1993 California Learning Assessment System (CLAS). *Educational Evaluation and Policy Analysis, 17*(3), 355-370.

Woods, C. M., & Grimm, K. (2011). Testing for nonuniform differential item function with multiple indicator multiple cause models.

Yoon, B., & Resnick, L. B. (1998). *Instructional validity, opportunity to learn and equity: New Standards Examinations for the California Mathematics Renaissance* (No. CSE 484). Los

Angeles, California: National Center for Research on Evaluation, Standards, and Student Testing.

Yu, L. (2006). *Using a differential item functioning (DIF) procedure to detect differences in opportunity to learn (OTL).* Master of Science, The Pennsylvania State University, University Park.

Yu, L., Lei, P.-W., & Suen, H. K. (2006). *Using a Differential Item Functioning (DIF) procedure to detect differences in Opportunity to Learn (OTL).* Paper presented at the American Educational Research, San Francisco, California.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, Canada: Department of National Defense.

Appendix A

Overview of the Mathematics Interim Assessment (Seventh Grade)

2010-11	Mathematics		
Testing Windows	Fall 1 September 15 – October 29	Fall 2 October 30 – December 31	Winter January 1 – February 28
Test Format	Multiple choice		
Test Sessions	1 per testing window		
Sessions Length	Suggested: 45-60 minutes		
# Indicators Tested	15		
# Questions/Indicator	2-4		
# Questions/Test	38		

Note: Adapted from Kansas Interim Assessment examiner’s manual, 2010.

Appendix B

Indicator Representation on the Interim Assessment (Grade 7)

Indicator Number	Grade 7 Indicator Descriptions	Number of Items on Each Interim Assessment	Number of Items on Each Summative Assessment
M.7.1.1.A1	Solves problems using equivalent representations of rational numbers and simple algebraic expressions.	3	6
M.7.1.4.K2	Performs and explains addition, subtraction, multiplication, and division of fractions and decimals.	4	8
M.7.1.4.K5	Finds percentages of rational numbers (e.g., 12.5% x $40.25 = n or 150% of 90 is what number?).	2	5
M.7.2.1.K1	Identifies, states, and continues patterns using numbers, symbols, diagrams, and verbal descriptions.	2	4
M.7.2.1.K4	States a rule for the nth term of an additive pattern with one operational change between terms.	2	5
M.7.2.2.A1	Represents real-world problems with symbols in linear expressions and one- or two-step equations.	3	6
M.7.2.2.K7	Relates ratios, proportions, and percents and solves proportions having positive rational solutions.	4	8
M.7.2.2.K8	Evaluates simple algebraic expressions using positive rational numbers.	2	5
M.7.3.1.K3	Identifies angle and side properties of triangles and quadrilaterals.	3	7
M.7.3.2.A1	Solves problems involving area and perimeter of two-dimensional composite figures.	2	5
M.7.3.2.K4	Knows and uses perimeter and area formulas for circles, rectangles, triangles, and parallelograms.	2	5
M.7.3.2.K6	Uses given measurement formulas to compute surface area of cubes and volume of rectangular prisms.	2	4
M.7.3.3.A3	Interprets scale drawings to determine actual measurements of two-dimensional figures.	2	4
M.7.4.2.A3	Recognizes and explains misleading data displays and the effects of scale changes on graphs of data.	2	5
M.7.4.2.K1	Organizes, interprets, and represents data in tabular, pictorial, and graphical displays.	3	7

Appendix C
Tested Indicators and Their Corresponding Items

Standards	*Benchmarks*	*Indicators*	*Items*
Standard 1 Number and Computation	Benchmark 1: Number Sense	**Ind. A1a: Generates and/or solves real-world problems using equivalent representations of rational numbers and simple algebraic expressions.**	8, 11, 12
	Benchmark 4: Computation	Ind. K2: performs and explains these computational procedures: a: adds and subtracts decimals from ten millions place through hundred thousandths place;	
		b: multiplies and divides a four-digit number by a two-digit number using numbers from thousands place through thousandths place;	5
		c: multiplies and divides using numbers from thousands place through thousandths place by 10; 100; 1,000; .1; .01; .001; or single-digit multiples of each;	4, 33
		d: adds, subtracts, multiplies, and divides fractions and expresses answers in simplest form.	13
		Ind. K5: finds percentages of rational numbers.	3, 23
Standard 2 Algebra	Benchmark 1: Patterns	**Ind. K1: identifies, states, and continues a pattern presented in various formats including numeric (list or table), algebraic (symbolic notation), visual (picture, table, or graph), verbal (oral description), kinesthetic (action), and written using these attributes: a: counting numbers including perfect squares, cubes, and factors and multiples (number theory);**	28
		b: positive rational numbers including arithmetic and geometric sequences (arithmetic: sequence of numbers in which the difference of two consecutive numbers is the same, geometric: a sequence of numbers in which each succeeding term is obtained by multiplying the preceding term by the same number)	29
		Ind. K4: states the rule to find the nth term of a pattern with one operational change (addition or subtraction) between consecutive terms.	6, 30
	Benchmark 2: Variable, Equations, and Inequalities	**Ind. K7: knows the mathematical relationship between ratios, proportions, and percents and how to solve for a missing term in a proportion with positive rational number solutions and monomials.**	20, 31
		Ind. K8: evaluates simple algebraic expressions using positive rational numbers.	10, 21
		Ind. A1: represents real-world problems using variables and symbols to write linear expressions, one- or two-step equations.	18, 19, 24

Appendix C. continued

Standards	*Benchmarks*	*Indicators*	*Items*
Standard 3 Geometry	Benchmark 1: Geometric Figures and Their Properties	**Ind. K3: identifies angle and side properties of triangles and quadrilaterals:** a. sum of the interior angles of any triangle is 180°;	
		b. sum of the interior angles of any quadrilateral is 360°;	
		c. parallelograms have opposite sides that are parallel and congruent;	
		d. rectangles have angles of 90°, opposite sides are congruent;	2
		e: rhombi have all sides the same length, opposite angles are congruent;	
		f. squares have angles of 90°, all sides congruent;	
		g. trapezoids have one pair of opposite sides parallel and the other pair of opposite sides are not parallel.	6
	Benchmark 2: Measurement and Estimation	**Ind. K4: knows and uses perimeter and area formulas for circles, squares, rectangles, triangles, and parallelograms.**	1, 34
		Ind. K6: uses given measurement formulas to find a. surface area of cubes,	
		b. volume of rectangular prisms.	9, 14
		Ind. A1c: solves real-world problems by finding perimeter and area of two-dimensional composite figures of squares, rectangles, and triangles.	27, 35
	Benchmark 3: Transformational Geometry	**Ind. A3: determines the actual dimensions and/or measurements of a two-dimensional figure represented in a scale drawing**	17, 22
Standard 4 Data	Benchmark 2: Statistics	**Ind. K1: organizes, displays, and reads quantitative (numerical) and qualitative (non-numerical) data in a clear, organized, and accurate manner including a title, labels, categories, and rational number intervals using these data displays** a. frequency tables;	
		b. bar, line, and circle graphs;	
		c. Venn diagrams or other pictorial displays;	
		d. charts and tables;	
		e. stem-and-leaf plots (single);	
		f. scatter plots;	26
		g. box-and-whiskers plots.	7, 15
		Ind. A3: recognizes and explains **a. misleading representations of data;**	32
		b. the effects of scale or interval changes on graphs of data sets.	25

Note: The 15 tested indicators are in bold.

Appendix D

ITEM SENSITIVITY REPORT FORM

Impact of Instructional Sensitivity on Test Items on High-Stakes Achievement Tests

INTRODUCTION

Instructional sensitivity refers to the degree to which students' performance on a test accurately reflects the quality of instruction. An instructionally sensitive item should vary in difficulty when responded by students with different instructional experiences. In other words, students who are taught the content tested should have a higher probability of responding to an item correctly than those who were not taught. Put simply, instructionally sensitive items should reflect the impact of effective instruction on students' performance. Detection of test items in terms of their instructional sensitivity will provide educators and administrators with new insights and knowledge for both improving achievement measurement process and taking action to bolster the practice of instruction in school or districts.

Please review carefully each of the multiple-choice items from the 2010-2011 Kansas Interim Assessment on Mathematics and provide your feedback based on the following guiding questions. The data collected from you will ONLY be used for this study. Your name and any other identifying information will be excluded from all reports based on analysis of your responses. Your cooperation will be highly appreciated!

Background Information

1) Gender: M ( ) F ( )

2) Are you currently teaching 7th grade math? Yes ( ) No ( )
If not, when did you recently teach 7th grade math? Year _________

3) How long have you been teaching math? __________________________

4) How long have you been teaching math in Kansas? ___________________________

---

1. Read the test items carefully. Based on the 10-point scale below, indicate how sensitive each of the test items is. Respond by assigning items under each ranking number.

	Totally Insensitive										*Totally sensitive*
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Item #											

2. For purposes of a teacher's instructional planning, how clearly are this state test's assessment targets (the skills and knowledge it measures) described? Are they stated with sufficient clarity that almost all of the state's teachers can identify what the benchmark and/or indicator really means? Respond by circling one number from the number line given below and providing your rationale.

Standards	*Benchmarks*	*Indicators*	*Totally Unclear* … *Extremely Clear*	*Comments*
Standard 1 Number and Computation The student uses numerical and computational concepts and procedures in a variety of situations	**Benchmark 1: Number Sense** The student demonstrates number sense for rational numbers, the irrational number pi, and simple algebraic expressions in one variable in a variety of situations.	Ind. A1a: Generates and/or solves real-world problems using equivalent representations of rational numbers and simple algebraic expressions.	*0 1 2 3 4 5 6 7 8 9 10*	
	Benchmark 4: Computation The student models, performs, and explains computation with rational numbers, the irrational number pi, and first-degree algebraic expressions in one variable in a variety of situations.	Ind. K2: performs and explains these computational procedures: a: adds and subtracts decimals from ten millions place through hundred thousandths place;	*0 1 2 3 4 5 6 7 8 9 10*	
		b: multiplies and divides a four-digit number by a two-digit number using numbers from thousands place through thousandths place;	*0 1 2 3 4 5 6 7 8 9 10*	
		c: multiplies and divides using numbers from thousands place through thousandths place by 10; 100; 1,000; .1; .01; .001; or single-digit multiples of each;	*0 1 2 3 4 5 6 7 8 9 10*	
		d: adds, subtracts, multiplies, and divides fractions and expresses answers in simplest form.	*0 1 2 3 4 5 6 7 8 9 10*	
		Ind. K5: finds percentages of rational numbers	*0 1 2 3 4 5 6 7 8 9 10*	

Note: To save space, the size of the table presented here is smaller than the actual table presented to the teachers for item review.

Standards	*Benchmarks*	*Indicators*	*Totally Unclear* *Extremely Clear*	*Comments*
Standard 2 Algebra The student uses algebraic concepts and procedures in a variety of situations	**Benchmark 1: Patterns** The student recognizes, describes, extends, develops, and explains the general rule of a pattern in a variety of situations.	Ind. K1: identifies, states, and continues a pattern presented in various formats including numeric (list or table), algebraic (symbolic notation), visual (picture, table, or graph), verbal (oral description), kinesthetic (action), and written using these attributes: a: counting numbers including perfect squares, cubes, and factors and multiples (number theory);	*0 1 2 3 4 5 6 7 8 9 10*	
		b: positive rational numbers including arithmetic and geometric sequences (arithmetic: sequence of numbers in which the difference of two consecutive numbers is the same, geometric: a sequence of numbers in which each succeeding term is obtained by multiplying the preceding term by the same number)	*0 1 2 3 4 5 6 7 8 9 10*	
		Ind. K4: states the rule to find the nth term of a pattern with one operational change (addition or subtraction) between consecutive terms.	*0 1 2 3 4 5 6 7 8 9 10*	
	Benchmark 2: Variable, Equations, and Inequalities The student uses variables, symbols, rational numbers, and simple algebraic expressions in one variable to solve linear equations and inequalities in a variety of situations.	Ind. K7: knows the mathematical relationship between ratios, proportions, and percents and how to solve for a missing term in a proportion with positive rational number solutions and monomials.	*0 1 2 3 4 5 6 7 8 9 10*	
		Ind. K8: evaluates simple algebraic expressions using positive rational numbers.	*0 1 2 3 4 5 6 7 8 9 10*	
		Ind. A1: represents real-world problems using variables and symbols to write linear expressions, one- or two-step equations.	*0 1 2 3 4 5 6 7 8 9 10*	

Standards	*Benchmarks*	*Indicators*	*Totally Unclear* … *Extremely Clear*	*Comments*
Standard 3 Geometry The student uses geometric concepts and procedures in a variety of situations	**Benchmark 1: Geometric Figures and Their Properties** The student recognizes geometric figures and compares their properties in a variety of situations.	Ind. K3: identifies angle and side properties of triangles and quadrilaterals: a. sum of the interior angles of any triangle is 180°;	*0 1 2 3 4 5 6 7 8 9 10*	
		b. sum of the interior angles of any quadrilateral is 360°;	*0 1 2 3 4 5 6 7 8 9 10*	
		c. parallelograms have opposite sides that are parallel and congruent;	*0 1 2 3 4 5 6 7 8 9 10*	
		d. rectangles have angles of 90°, opposite sides are congruent;	*0 1 2 3 4 5 6 7 8 9 10*	
		e: rhombi have all sides the same length, opposite angles are congruent;	*0 1 2 3 4 5 6 7 8 9 10*	
		f. squares have angles of 90°, all sides congruent;	*0 1 2 3 4 5 6 7 8 9 10*	
		g. trapezoids have one pair of opposite sides parallel and the other pair of opposite sides are not parallel.	*0 1 2 3 4 5 6 7 8 9 10*	
	Benchmark 2: Measurement and Estimation The student estimates, measures, and uses measurement formulas in a variety of situations.	Ind. K4: knows and uses perimeter and area formulas for circles, squares, rectangles, triangles, and parallelograms.	*0 1 2 3 4 5 6 7 8 9 10*	
		Ind. K6: uses given measurement formulas to find a. surface area of cubes,	*0 1 2 3 4 5 6 7 8 9 10*	
		b. volume of rectangular prisms.	*0 1 2 3 4 5 6 7 8 9 10*	
		Ind. A1c: solves real-world problems by finding perimeter and area of two-dimensional composite figures of squares, rectangles, and triangles	*0 1 2 3 4 5 6 7 8 9 10*	
	Benchmark 3: Transformational Geometry The student recognizes and performs transformations on two- and three- dimensional geometric figures in a variety of situations.	Ind. A3: determines the actual dimensions and/or measurements of a two-dimensional figure represented in a scale drawing	*0 1 2 3 4 5 6 7 8 9 10*	

Standards	*Benchmarks*	*Indicators*	*Totally Unclear* *Extremely Clear*	*Comments*
Standard 4 Data The student uses concepts and procedures of data analysis in a variety of situations	**Benchmark 2: Statistics** The student collects, organizes, displays, and explains numerical (rational numbers) and non-numerical data set in a variety of situations with a special emphasis on measures of central tendency.	Ind. K1: organizes, displays, and reads quantitative (numerical) and qualitative (non-numerical) data in a clear, organized, and accurate manner including a title, labels, categories, and rational number intervals using these data displays a. frequency tables;	*0 1 2 3 4 5 6 7 8 9 10*	
		b. bar, line, and circle graphs;	*0 1 2 3 4 5 6 7 8 9 10*	
		c. Venn diagrams or other pictorial displays;	*0 1 2 3 4 5 6 7 8 9 10*	
		d. charts and tables;	*0 1 2 3 4 5 6 7 8 9 10*	
		e. stem-and-leaf plots (single);	*0 1 2 3 4 5 6 7 8 9 10*	
		f. scatter plots;	*0 1 2 3 4 5 6 7 8 9 10*	
		g. box-and-whiskers plots.	*0 1 2 3 4 5 6 7 8 9 10*	
		Ind. A3: recognizes and explains a. misleading representations of data;	*0 1 2 3 4 5 6 7 8 9 10*	
		b. the effects of scale or interval changes on graphs of data sets.	*0 1 2 3 4 5 6 7 8 9 10*	

3. Please classify each item into the following categories.

Standards	*Benchmarks*	*Indicators*	*Items*
Standard 1 Number and Computation The student uses numerical and computational concepts and procedures in a variety of situations	**Benchmark 1: Number Sense** The student demonstrates number sense for rational numbers, the irrational number pi, and simple algebraic expressions in one variable in a variety of situations.	Ind. A1a: Generates and/or solves real-world problems using equivalent representations of rational numbers and simple algebraic expressions.	
	Benchmark 4: Computation The student models, performs, and explains computation with rational numbers, the irrational number pi, and first-degree algebraic expressions in one variable in a variety of situations.	Ind. K2: performs and explains these computational procedures: a: adds and subtracts decimals from ten millions place through hundred thousandths place;	
		b: multiplies and divides a four-digit number by a two-digit number using numbers from thousands place through thousandths place;	
		c: multiplies and divides using numbers from thousands place through thousandths place by 10; 100; 1,000; .1; .01; .001; or single-digit multiples of each;	
		d: adds, subtracts, multiplies, and divides fractions and expresses answers in simplest form.	
		Ind. K5: finds percentages of rational numbers	
Standard 2 Algebra The student uses algebraic concepts and procedures in a variety of situations	**Benchmark 1: Patterns** The student recognizes, describes, extends, develops, and explains the general rule of a pattern in a variety of situations.	Ind. K1: identifies, states, and continues a pattern presented in various formats including numeric (list or table), algebraic (symbolic notation), visual (picture, table, or graph), verbal (oral description), kinesthetic (action), and written using these attributes: a: counting numbers including perfect squares, cubes, and factors and multiples (number theory);	
		b: positive rational numbers including arithmetic and geometric sequences (arithmetic: sequence of numbers in which the difference of two consecutive numbers is the same, geometric: a sequence of numbers in which each succeeding term is obtained by multiplying the preceding term by the same number)	
		Ind. K4: states the rule to find the nth term of a pattern with one operational change (addition or subtraction) between consecutive terms.	
	Benchmark 2: Variable, Equations, and Inequalities The student uses variables, symbols, rational numbers, and simple algebraic expressions in one variable to solve linear equations and inequalities in a variety of situations.	Ind. K7: knows the mathematical relationship between ratios, proportions, and percents and how to solve for a missing term in a proportion with positive rational number solutions and monomials.	
		Ind. K8: evaluates simple algebraic expressions using positive rational numbers.	
		Ind. A1: represents real-world problems using variables and symbols to write linear expressions, one- or two-step equations.	

Standards	*Benchmarks*	*Indicators*	*Items*
Standard 3 **Geometry** The student uses geometric concepts and procedures in a variety of situations	**Benchmark 1: Geometric Figures and Their Properties** The student recognizes geometric figures and compares their properties in a variety of situations.	Ind. K3: identifies angle and side properties of triangles and quadrilaterals: a. sum of the interior angles of any triangle is 180°;	
		b. sum of the interior angles of any quadrilateral is 360°;	
		c. parallelograms have opposite sides that are parallel and congruent;	
		d. rectangles have angles of 90°, opposite sides are congruent;	
		e: rhombi have all sides the same length, opposite angles are congruent;	
		f. squares have angles of 90°, all sides congruent;	
		g. trapezoids have one pair of opposite sides parallel and the other pair of opposite sides are not parallel.	
	Benchmark 2: Measurement and Estimation The student estimates, measures, and uses measurement formulas in a variety of situations.	Ind. K4: knows and uses perimeter and area formulas for circles, squares, rectangles, triangles, and parallelograms.	
		Ind. K6: uses given measurement formulas to find a. surface area of cubes,	
		b. volume of rectangular prisms.	
		Ind. A1c: solves real-world problems by finding perimeter and area of two-dimensional composite figures of squares, rectangles, and triangles.	
	Benchmark 3: Transformational Geometry The student recognizes and performs transformations on two- and three- dimensional geometric figures in a variety of situations.	Ind. A3: determines the actual dimensions and/or measurements of a two-dimensional figure represented in a scale drawing	
Standard 4 **Data** The student uses concepts and procedures of data analysis in a variety of situations	**Benchmark 2: Statistics** The student collects, organizes, displays, and explains numerical (rational numbers) and non-numerical data set in a variety of situations with a special emphasis on measures of central tendency.	Ind. K1: organizes, displays, and reads quantitative (numerical) and qualitative (non-numerical) data in a clear, organized, and accurate manner including a title, labels, categories, and rational number intervals using these data displays a. frequency tables;	
		b. bar, line, and circle graphs;	
		c. Venn diagrams or other pictorial displays;	
		d. charts and tables;	
		e. stem-and-leaf plots (single);	
		f. scatter plots;	
		g. box-and-whiskers plots.	
		Ind. A3: recognizes and explains a. misleading representations of data;	
		b. the effects of scale or interval changes on graphs of data sets.	

Appendix E

CETE mailing address: Angela Broaddus, Center for Educational Testing and Evaluation, Joseph R. Pearson Hall, 1122 West Campus Rd., Room 748, Lawrence, KS 66045

CETE fax number: 785-864-2916

CENTER FOR EDUCATIONAL TESTING AND EVALUATION

CONFIDENTIALITY AGREEMENT

Test security and student confidentiality are of utmost importance to the Kansas State Department of Education. As a participant in this test item review for the Kansas Math Interim Assessment research project, you have access to materials that must be kept secure. Please treat all materials as confidential.

You are asked not to reproduce any materials, directly or indirectly, and not to disclose the content of these materials. The Kansas State Department of Education takes pride in ensuring equity for all students. Therefore, please do not put any Kansas student at an unfair advantage by sharing information learned with your district colleagues.

We are certain that you share our concern that all potential assessment materials be handled in a professional, secure, and confidential manner, and we ask for your adherence to these guidelines by signing below.

__________________________ __________________________
Participant Name (please print) Date

____
Participant Signature

____
School or Organization