Detecting Item Sensitivity to Instruction:
A Comparison between Mantel-Haenszel and Logistic Regression Procedures

Jie Chen

Neal Kingston, Ph.D.

University of Kansas

ABSTRACT

The purpose of this study is to explore the relationship between students' instructional experiences and their performance on standardized achievement tests. Using a state interim assessment program, the study aims to detect items that are able to reflect the impact of effective instruction, to examine how students' performance on test items differ before and after instruction, and to explore what common characteristics the instructionally sensitive items have. Results show that more than half of the items in this test were sensitive to instruction. Items detected as sensitive by both the Mantel-Haenszel and logistic regression methods were identical. Items testing topics of Geometric Figures and Their Properties, Measurement and Estimation, and Statistics were more likely to be sensitive.

INTRODUCTION

Living in an era of test-based accountability systems, how do we hold accountability tests accountable? Court (2010) pointed out schools are perceived as better or worse based on their proficiency, readiness, or growth; teachers are believed to be more effective when their students have higher performance on high-stakes achievement assessments. He also specified that these accountability decisions are based on the assumption that test scores successfully reflect the effect of instruction. Therefore, it is important that accountability tests assess what is taught. However, 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.

Goe (2007) cautioned that 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. Therefore, the purpose of this study is to detect items, on a state achievement test, that are sensitive to instruction and to explore the relationship between students' instructional experiences and their performance on items.

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 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.

LITERATURE REVIEW

Standards-Based Assessment and Instructional Sensitivity

Standards-Based Assessment

Standards-based assessment is a comparatively new concept that is a key part 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, pp2-3) :

- 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).

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 or not.

Instructional Sensitivity and Standards-Based Tests

According to Popham (2001), the 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 know in terms of knowledge and skills, which they are expected to have mastered at a certain grade level.

In the era of standards-based reform, a standards-based assessment is the key component in 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).

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). Most researchers believed that OTL information from teachers works better to represent the instructional coverage for test items (Lehman, 1986; Kao, 1990).

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 (2010) 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, 2010). These tests measure what students bring to school, but not what they learn from school (Popham, 2010). 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.

Research Questions

In this study, the following research questions are addressed:

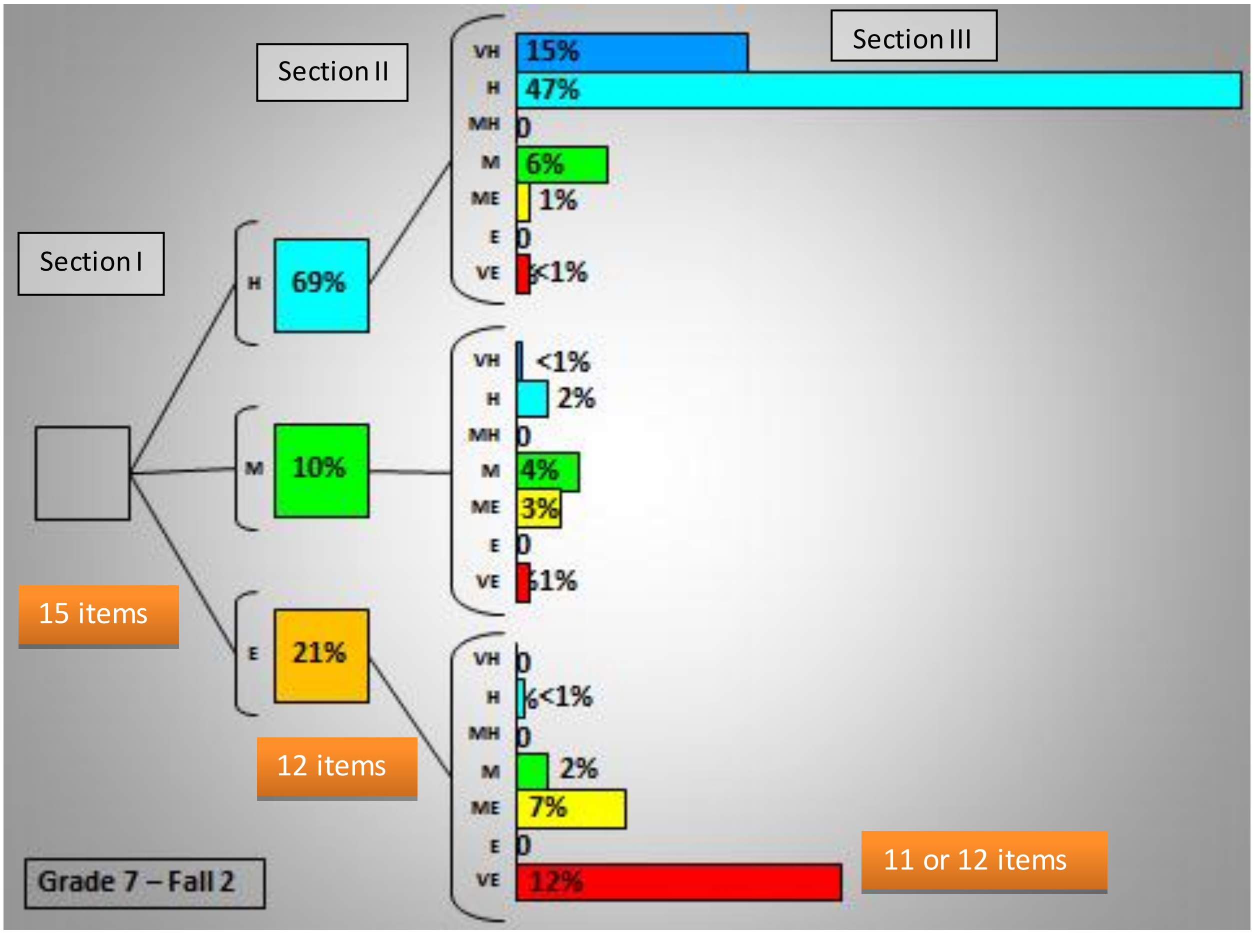
1. To what extent are the items in the state testing program sensitive to instruction?
2. Are item performance differences due to differences in curricular content covered in instruction?
3. How does the instruction in content, tested on the state testing program, influence students' performance?

Methods and Results

Data Description

The Kansas State Interim Assessments are multi-stage adaptive-designed computer adaptive tests. In total, ninety items were given to 5,510 seventh graders in the second testing window of the assessment on 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 1 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 1. 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 1) and of "high difficulty" ("H" in Figure 1) 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 1 presents the details of demographic information.

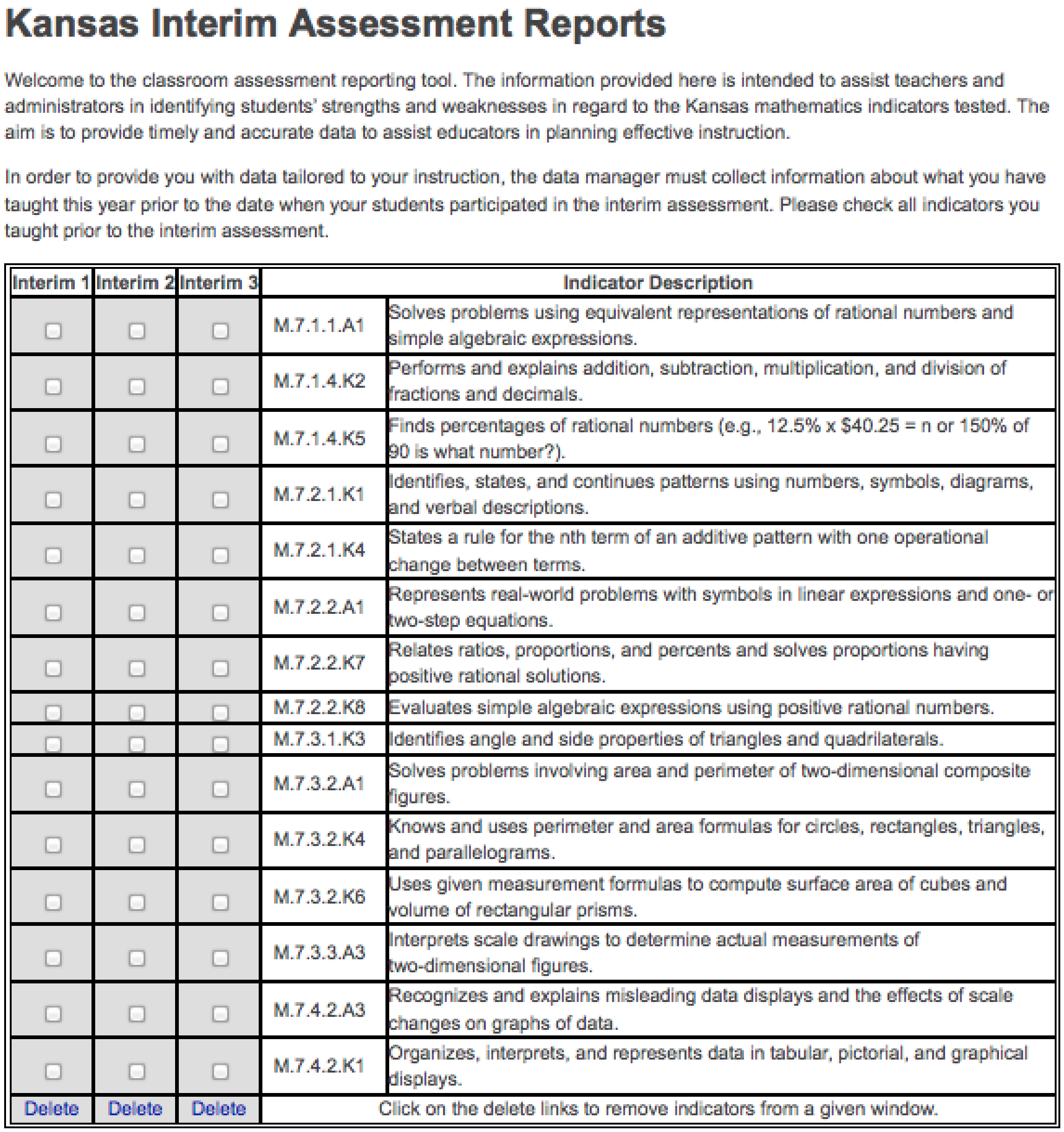
Table 1. The Demographics of the Student Sample

	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	142	4.1
						Hispanic	471	13.7
						White	2577	74.8
						multiracial	138	4.0
						Pacific Islander	7	.2
Missing	6	.2		2134	61.9		6	.2
Total	3446	100.0		3446	100.0		3446	

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 would be presented after the teacher's login. Figure 2 shows how the screen actually appeared to the teachers.

Figure 2. 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.	

The teacher was expected to click the check box beside any indicators that were taught prior to the interim assessment. The teacher should consider whether the instruction provided prior to an assessment for each indicator was adequate for students to be able to successfully answer all potential items assessing that indicator. 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 2 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 2. 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 2 continued)

Item	**Group**	**n** **(total = 3446)**	$\bar{\theta}$ **(SD)** **(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 2 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 criminated 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.

In order to answer the 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 $P(U=1)=\frac{e^z}{1+e^z}$. 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: $Z = \beta_0 + \beta_1\theta + \beta_2G$; 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: $Z = \beta_0 + \beta_1\theta + \beta_2G + \beta_3(\theta*G)$. By comparing the fit of Model 3 (i.e., the χ^2 statistics) 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 1 to Model 2 indicates instructional sensitivity of an item due to the interaction between a student's proficiency 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 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 3 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 3. Report of Logistic Regression Model Comparisons: χ^2, $\Delta\chi^2$, R^2, and ΔR^2

	χ^2(*df*=1)	***R*^{2}**	**χ^2(*df*=1)**	***R*^{2}**	**χ^2(*df*=1)**	***R*^{2}**	**$\Delta\chi^2$(*df*=2)**		**$\Delta\chi^2$(*df*=1)**		**$\Delta\chi^2$(*df*=1)**		**Δ*R*^{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 3, the χ^2 statistics (refer to columns labeled as "$\boldsymbol{\chi^2}$ **(*df* = 1) Model 1**", "$\boldsymbol{\chi^2}$ **(*df* = 1) Model 2**" and "$\boldsymbol{\chi^2}$ **(*df* = 1) Model 3**") and the Nagelkerke R^2 effect size (refer to columns labeled as "$\boldsymbol{R^2}$ **M1**", "$\boldsymbol{R^2}$ **M2**" and "$\boldsymbol{R^2}$ **M3**") for each model were reported first. Next, the pair-wise comparison of χ^2 statistics between each two models (refer to two columns labeled as "$\boldsymbol{\Delta\chi^2}$") 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 sensitive to instruction. 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.

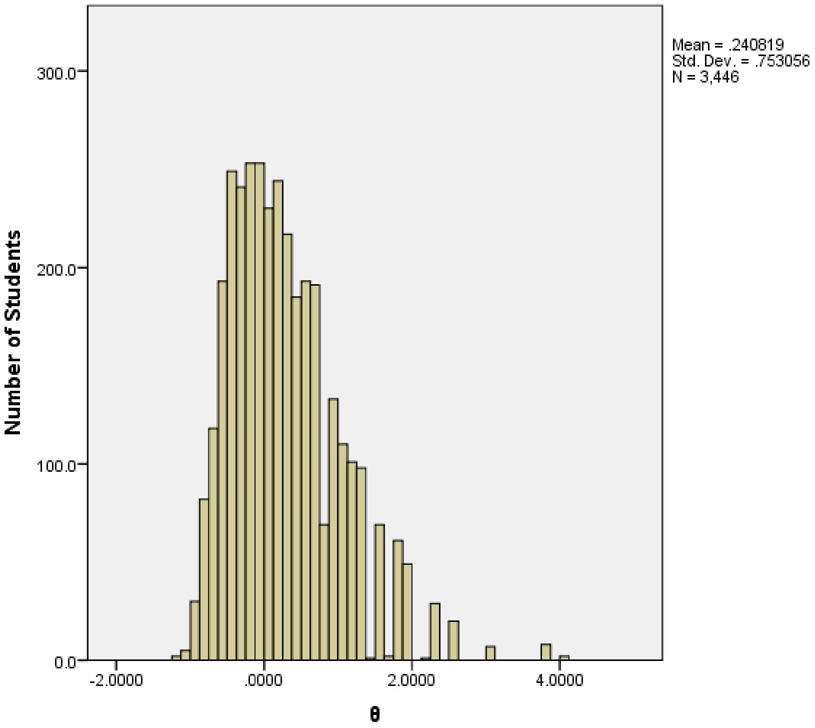
Mantel-Haenszel Tests

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 1, students who were given the items presented in the first 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 4 is a summary of descriptive statistics of θ, and Figure 3 graphically shows how θ for the sample used in this study was distributed.

Table 4. Descriptive Statistics of Math proficiency (θ)

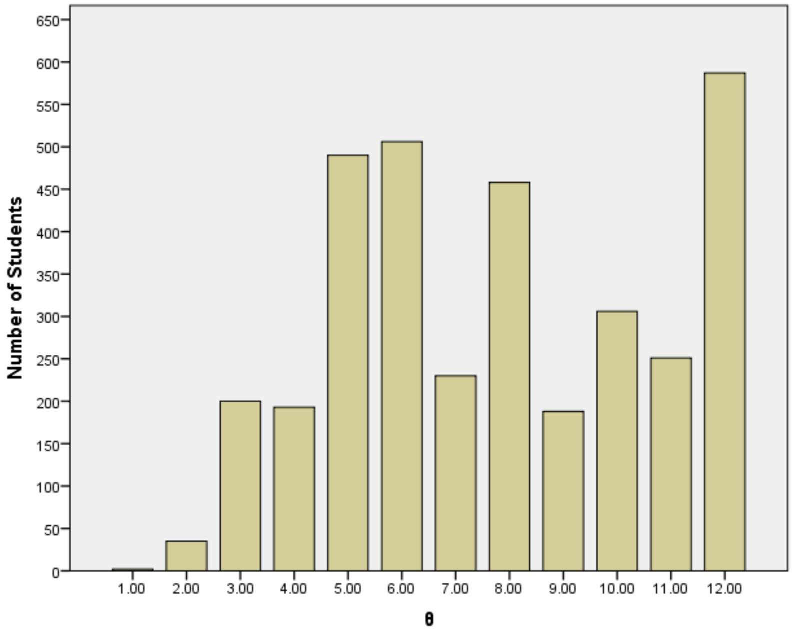
	N	**Min.**	**Max.**	**Mean**	**SD**	**25th Percentile**	**50th Percentile**	**75th Percentile**
θ	3,446	-1.234	4.000	.241	.75	-.352	.128	.640

Figure 3. Distribution of 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 4.

Figure 4. 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).

$$\alpha_i = \frac{p_{ri}}{q_{ri}} \Big/ \frac{p_{fi}}{q_{fi}} = \frac{a_i}{b_i} \Big/ \frac{c_i}{d_i} = \frac{a_i d_i}{b_i c_i} \; i = 1, 2, . . . , k$$

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.

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 focal group, and with values greater than 1.0 indicating that the item is in favor of the reference 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}. The values are then asymptotically normally distributed 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 is used as the measure of effect size. Table 5 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 5. Results of Mantel-Haenszel Tests

Item	MH χ^2 (*df* = 1)	MH 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
33	.947	-.181	.17
10	1.945	-.158	.11
25	2.021	-.133	.09
29	2.837	-.133	.08
22	2.322	-.124	.08
24	2.012	-.122	.08
27	1.528	-.104	.08
18	1.195	-.086	.08
21	.202	-.070	.13
28	.004	-.021	.15
26	.110	.053	.13
17	1.602	.121	.09
8	2.751	.144	.08
11	4.367	.212	.10

As shown in Table 5, 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 6) 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 7 were ordered based on the effective sizes of the two methods, respectively.

Table 6. A Comparison of Results Obtained from Both 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	33	-.181	-
24	.002	-	10	-.158	-
4	.001	-	25	-.133	-
22	.001	-	29	-.133	-
25	.001	-	22	-.124	-
27	.001	-	24	-.122	-
28	.001	-	27	-.104	-
29	.001	-	18	-.086	-
33	.001	-	21	-.070	-
17	.001	-	28	-.021	-
8	.001	-	26	.053	-
18	.001	-	17	.121	-
21	0	-	8	.144	-
26	0	-	11	.212	-

Table 6 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 or not. 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:

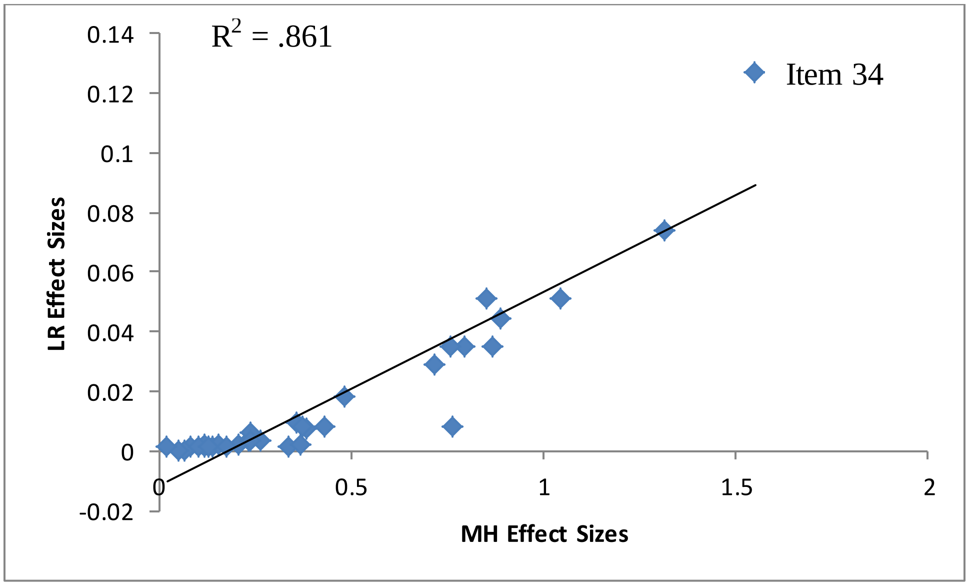
First, both methods detected 19 items in common which were instructionally sensitive. Logistic Regression procedures detected one more sensitive item (Item 11) in addition to the 19 items the methods detected in common. However, the degrees of sensitivity for items detected were not exactly the same when ranked by the two methods, although the rankings were similar. Table 7 lists the sensitive items detected by both methods, where items with the same rankings were highlighted; 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 7. Ranking of Sensitivity by 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 method were highly correlated (*r* = .928), which means the two methods agreed with each other to a high degree in detecting instructionally sensitive items. Figure 5 is the scatter plot showing the relationship between the effect sizes from both methods.

Figure 5. 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, nineteen 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 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.

Fifth, items testing topics of *Geometric Figures and Their Properties*, *Measurement and Estimation*, and *statistics* were more likely to be sensitive. Table 8 listed the content areas tested by the detected sensitive items. Based on the information under Column "Ratio", two out of two items under the benchmark of *Geometric Figures and Their Properties* were detected as instructionally sensitive; five out of six items under the benchmark of *Measurement and Estimation* were detected as instructionally sensitive; and three out of five items under the benchmark of *statistics* were detected as instructionally sensitive.

Table 8. Content Characteristics of Sensitive Items

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.

Conclusion

The number of sensitive items found in this study is considerably large. About 54% of items were sensitive to instruction. This may suggest that students' performance on seventh grade Kansas Interim Assessment on mathematics were 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 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, and instruction positively influenced 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 researchers 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. Among the 20 sensitive items detected 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. Second, the LR procedure keeps the matching variable, student's proficiency, continuous, while the MH approach categorized this continuous variable.

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.

Reference

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.

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.

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., & 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.

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.

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.

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.

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.

Penfield, R. D. (2007). DIFAS 4.0 user’s manual.

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.

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

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

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.

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.

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.

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.