

Instructional Sensitivity as a Psychometric Property of Assessments

Morgan S. Polikoff, *University of Southern California*

Standards-based reform, as codified by the No Child Left Behind Act, relies on the ability of assessments to accurately reflect the learning that takes place in U.S. classrooms. However, this property of assessments—their instructional sensitivity—is rarely, if ever, investigated by test developers, states, or researchers. In this paper, the literature on the psychometric property of instructional sensitivity is reviewed. Three categories of instructional sensitivity measures are identified—those relying on item or test scores only, those relying on item or test scores and teacher reports of instruction, and strictly judgmental methods. Each method identified in the literature is discussed alongside the evidence for its utility. Finally, recommendations are made as to the proper role of instructional sensitivity in the evaluation of assessments used under standards-based reform.

Keywords: assessment, instruction, NCLB, psychometrics, sensitivity

The success of schools under standards-based reform depends on the performance of students on state assessments of student achievement. The reform's theory of change suggests that, with high-quality content standards and aligned assessments, teachers will modify their instruction to align to the standards and assessments, and achievement will rise (Smith & O'Day, 1990). Dissatisfied with the performance of the No Child Left Behind Act (NCLB) to this point, researchers and policymakers have begun to advocate for and work toward common standards and assessments that would, presumably, be better aligned and more likely to encourage teachers to modify their instruction. However, there has been little discussion of the last part of the theory of change—that changes in instruction will result in changes in performance on assessments of student learning (Popham, 2007). Indeed, though millions of dollars are spent on developing and administering assessments of student learning each year across the 50 states (Goertz, 2005), there is little evidence that these assessments can detect the effects of high-quality instruction. Instead, it is assumed that assessments aligned to standards through one of the various alignment procedures will necessarily be able to detect instructional quality. This key omitted property of an assessment is called its instructional sensitivity, and it is the focus of this paper.

Research on instructional sensitivity as a psychometric property of assessments was prevalent around the birth of criterion-referenced testing in the 1960s and 1970s. It was believed to be a critical feature of criterion-referenced assessment (Haladyna & Roid, 1981). Paradoxically, however, as criterion-referenced assessment exploded in the last three decades, instructional sensitivity became assumed, rather

than studied. In what follows, I describe instructional sensitivity as a psychometric property and outline its origins in the criterion-referenced testing literature. Next, I identify and describe the various measures of instructional sensitivity and use the extant literature to identify the most useful measures. Three broad classes of instructional sensitivity indicators emerge—purely statistical methods based solely on test scores, instruction-focused methods based on a combination of test scores and teacher-reported instruction, and methods based on expert judgment. Finally, I discuss reasons why instructional sensitivity has lost appeal as a psychometric property and argue for its importance in the development and validation of criterion-referenced assessments used for standards-based reform.

Instructional Sensitivity: Meanings and Origins

Terminology

Before beginning, it is important to discuss key terminology. The term “instructional sensitivity” was chosen as the focal term for this analysis to refer to the extent to which student performance on a test or item reflects the instruction received (Kosecoff & Klein, 1974). Instead, the term “instructional validity” could have been chosen, as some authors use the terms interchangeably (e.g., D'Agostino, Welsh, & Corson, 2007). While instructional sensitivity is sometimes used to refer to the ability to detect differences in the *quality* of instruction (Popham, 2007), most uses of the term focus on instruction generally—implicitly, both content and quality (D'Agostino et al., 2007; Haladyna & Roid, 1981; Kosecoff & Klein, 1974; Muthen, Kao, & Burstein, 1991). Furthermore, wherever the term “instructional sensitivity” has been used, it has been to refer to a property of a test or item. In short, an instructionally sensitive test should be able to detect differences in instruction received by students.

Dr. Morgan Polikoff, University of Southern California – Rossier School of Education, Waite Phillips Hall 904D, Los Angeles, California 90089, United States; polikoff@usc.edu.

Instructional validity is a closely related term, which arose a few years later than sensitivity, in two parallel contexts. One usage of instructional validity arose in the context of the Debra P. v. Turlington case (1981) and focused on the extent to which schools were providing instruction in the content measured by the test (McClung, 1977). Instructional validity also arose in the context of the test score declines of the 1970s, with Feldhausen, Hynes, and Ames (1976) arguing that the declines were partly due to a lack of high-quality instruction in each tested objective—what they deemed instructional validity. In short, from the beginning “instructional validity” was a muddled term, at times referring to content only (as in the Debra P. case) and at others referring to content and quality of instruction. However, in both cases, instructional validity was focused on whether instruction was adequate to justify inferences and consequences made based on the results of a test. As D’Agostino et al. (2007) point out, the meaning of instructional validity quickly switched to a definition essentially identical to sensitivity: “the extent to which an assessment is systematically sensitive to the nature of instruction offered” (Yoon & Resnick, 1998, p. 2), allowing researchers the ability to use the terms interchangeably. Instructional validity was the more commonly used term throughout the 1980s and 1990s.

To further complicate matters, one could also speak of “curricular sensitivity” or “curricular validity.” The latter was sometimes used in the 1970s and 1980s, and it referred more to formal curriculum documents than the content or quality of instruction (McClung, 1977; Mehrens & Phillips, 1987). The term “instructional sensitivity” is used here because it is believed that the onus for evaluating instructional sensitivity/validity should properly be placed on the test or item, rather than the school or teacher. Furthermore, instructional sensitivity, which is based on an overall instructional effect, subsumes the narrower curricular sensitivity. Nonetheless, some of the studies cited in this analysis use different terminology, and some do not refer to instructional sensitivity or instructional validity at all. Wherever the study’s methods address the broad definition of instructional sensitivity offered above, its results are included.

Instructional sensitivity is also closely related to two other prominent educational concepts. The first is instructional alignment (Porter, 2002; Porter, Smithson, Blank, & Zeidner, 2007), or the extent to which the content of instruction matches up with the content of standards, assessments, or other materials. While Porter and colleagues use alignment as a teacher variable, making it a useful tool for evaluating sensitivity, alignment is perhaps more prominently used to refer to content overlap between standards and assessments (Webb, 1999). Yet another closely related term is opportunity to learn (OTL) (D’Agostino et al., 2007). OTL, originally introduced in cross-national studies of achievement, was a measure of the extent to which students had sufficient instruction and resources to learn some set of content (McDonnell, 1995). OTL was enshrined, with some controversy, in the Goals 2000 Act (Goals 2000: Educate America Act of 1994, 1994; McDonnell, 1995; Porter, 1995), though it has subsequently all but disappeared from the lexicon. The purpose of identifying these similar terms is to clarify that instructional sensitivity is the focus here, and the term that will be used throughout, even if the authors of the cited studies used a different term.

History

The origins of instructional sensitivity come from the early days of the development of criterion-referenced testing, when a number of authors identified the distinctions between criterion- and norm-referenced assessments (Cox & Vargas, 1966; Glaser, 1963; Haladyna & Roid, 1976; Popham & Husek, 1969). These authors argued that, by definition, norm-referenced assessments were useful for indicating the relative standing of individuals in a group and not for indicating an absolute level of achievement. Rather, they were designed to maximize variability among test takers in order to differentiate them. In contrast, criterion-referenced assessments were designed to be used to measure individual mastery over a particular domain (Glaser, 1963). Rather than differentiating among individuals, criterion-referenced tests were intended to differentiate between successive performances of one individual (Cox & Vargas, 1966).

In light of this purpose, a number of authors commented that the item statistics used for traditional norm-referenced assessments were not appropriate for criterion-referenced assessments (Haladyna & Roid, 1981). For instance, one traditional item statistic is the item-total correlation. Items not sufficiently correlated with the total score are deleted for norm-referenced assessment. However, the discarded items may cover important topics in a criterion-referenced situation, in which case the assessment would no longer be representative of the domain (Cox & Vargas, 1966). Another item statistic is the discrimination index (or IRT discrimination index), which is the difference in item performance between higher and lower performers. For criterion-referenced items, however, traditional discrimination is not as important as discrimination between masters and nonmasters, or those who have received instruction and those who have not (Brennan, 1972; Helmstadter, 1974; Kosecoff & Klein, 1974; Popham & Husek, 1969). A third item statistic is its difficulty—very easy or hard items are often removed from assessments to eliminate floor and ceiling effects. But for criterion-referenced assessments, difficult items may simply indicate that the topic has not yet been taught—when the topic is taught, the difficulty of the item should decrease. Overall, psychometricians agreed that traditional norm-referenced item selection techniques were inappropriate for the criterion-referenced tests (Cox & Vargas, 1966; Haladyna & Roid, 1981; Kosecoff & Klein, 1974; Popham, 1971; Popham & Husek, 1969).

To address this shortcoming of traditional techniques, psychometricians set out to develop more appropriate item selection techniques for criterion-referenced assessments. The indices they created were all based around the same premise—item selection techniques for criterion-referenced assessments should be able to identify the effects of some treatment, to see how much some individual had learned (Brennan, 1972; Cox & Vargas, 1966; Helmstadter, 1972; Kosecoff & Klein, 1974; Popham, 1971; Roudabush, 1974). This property of items and assessments was called instructional sensitivity. It was seen as the key psychometric feature of assessments for measuring the achievement of some criterion, and several indices and techniques were proposed. Furthermore, early research established that tests differed markedly in their sensitivity to instruction—so much so that research findings not accounting for sensitivity were likely to be substantially biased (Walker & Shaffarzik, 1974).

Before the early 1980s, the focus of psychometricians working on instructional sensitivity was the creation of statistics, like the traditional item difficulty and item discrimination indices, which could be calculated using only data from assessments. Around the early 1980s, new methods were brought to light. These methods included the gathering of data on the content of teachers' instruction and relating these data to students' learning of that content. At this point, the earlier indices, which ignored the content of instruction, disappeared, and the term "instructional sensitivity" began to be replaced by instructional validity or OTL. Nonetheless, the techniques used in analyses of instructional validity remain applicable to the investigation of sensitivity, so those techniques are discussed here.

Measuring Instructional Sensitivity

In this section, I introduce the indicators of instructional sensitivity that have appeared in the literature. Before introducing the indicators, however, it is important to point out one limitation that applies to *all* of the studies presented. Because *none* of the studies considered the content and quality of instruction on some external measure (e.g., having experts observe, code the content, and rate the quality of teachers' instruction), it is impossible to say whether a finding of low or no sensitivity in any particular study is due to a poor-quality test that is actually insensitive to instruction or to poor-quality instruction, so that the test results actually reflect the instruction received by students. In contrast, a finding of high sensitivity indicates both good instruction and also a high-quality test sensitive to that instruction. Clearly, the goal is always to have instruction of maximum effectiveness, and to design a test to capture the effects of instruction.

Instructional Sensitivity as an Item Statistic

The first wave of instructional sensitivity indices were, like traditional item statistics, based on item responses. The majority required both pretest and posttest data, ideally, but not necessarily, from the same students or students randomly selected from a certain pool. The first indicator was the Pretest-Posttest Difference Index (PPDI), proposed by Cox and Vargas (1966). The most conceptually simple of the indices, the PPDI is the proportion of students passing the item on the posttest minus the proportion of students passing the item on the pretest.

$$PPDI = \text{Diff}_{\text{post}} - \text{Diff}_{\text{pre}}. \quad (1)$$

In other words, the PPDI represents the proportion of students who have mastered the item content during the period of instruction.

Three closely related indices were also proposed. The Percent of Possible Gain (PPG) was designed to correct for the fact that large gains are impossible on items that are easy at pretest (Brennan & Stolurow, 1971).

$$PPG = (\text{Diff}_{\text{post}} - \text{Diff}_{\text{pre}}) / (1 - \text{Diff}_{\text{pre}}). \quad (2)$$

The PPG seemed superior to the PPDI because it accounted for the potential for items to show improvement (Haladyna & Roid, 1981). The Brennan Index was analogous to the PPDI, except mastery and nonmastery groups were used instead of post- and pretest groups (Brennan, 1972).

$$\text{Brennan} = \text{Diff}_{\text{masters}} - \text{Diff}_{\text{nonmasters}}. \quad (3)$$

		Posttest	
		Incorrect $X_{\cdot 0}$	Correct $X_{\cdot 1}$
Pretest	Incorrect $X_{0 \cdot}$	X_{00}	X_{01}
	Correct $X_{1 \cdot}$	X_{10}	X_{11}

FIGURE 1. Contingency table of pretest and posttest item performance.

This technique was seen as superior because it took total test performance into account, in addition to individual item data. However, the Brennan Index required the establishment of some a priori mastery threshold. A third alternative was the ZDIFF (Haladyna & Roid, 1981), which was analogous to the PPDI, except that ZDIFF was a normalized difference between item response theory (IRT) difficulty statistics on the pre- and posttests. An obvious advantage of ZDIFF is that it can be used on tests scored using IRT, which is the case for most large-scale assessments today.

The next group of sensitivity indices introduced was a set based on a contingency table of correct and incorrect responses, like the one shown in Figure 1. Contingency table methods were seen as more appropriate because they used the patterns in item response across pre- and posttests. The proportion in each cell represents the proportion of respondents who answered the item right or wrong at pretest/posttest. The first contingency table index was a traditional phi coefficient (Popham, 1971), which was seen as useful because it took into account all four possible pretest-posttest answer patterns.

$$\Phi = (X_{00}X_{11} - X_{01}X_{10}) / \sqrt{(X_{1 \cdot}X_{0 \cdot}X_{\cdot 1}X_{\cdot 0})}. \quad (4)$$

Hsu (1971) proposed an identical sensitivity index (Φ^*), except that the phi coefficient was calculated using right (1) and wrong (0) answers, along with mastery (1) and non-mastery (0) based on some preset mastery threshold. The relationship between Φ^* and Φ was analogous to the relationship between Brennan and PPDI; that is, Φ^* was seen as potentially superior because it accounted for both item and test performance.

An alternative contingency table sensitivity index was proposed by Roudabush (1974), based on work done at McGraw Hill. This Guessing-Corrected Sensitivity Index (GCSI, my term) accounted for guessing by assuming a fixed guessing parameter p that did not vary across items. Furthermore, it was assumed that there was no forgetting, meaning that any individuals in X_{10} had guessed. Therefore, the guessing parameter p was equal to $X_{10} / (X_{00} + X_{10})$.

$$GCSI = (X_{01} - X_{10})(X_{01} + X_{00}) / ((X_{01} - X_{10})(X_{01} + X_{00}) + (X_{10} + X_{00})^2). \quad (5)$$

The last contingency table index was the External Sensitivity Index (ESI), proposed by Kosecoff and Klein (1974).

The ESI was designed to evaluate the item's sensitivity independently from that of the test. They also offered an ESI with a guessing correction, which they called ESI*.

$$ESI = (X_{01} - X_{00}) / (X_{00} + X_{01} + X_{10} + X_{11}), \quad (6)$$

$$ESI^* = (X_{00} - X_{10})((X_{01} - X_{10}) - (X_{00} + X_{10})) / (X_{00} * (X_{00} + X_{01} + X_{10} + X_{11})). \quad (7)$$

The authors described the ESI as being similar to a phi coefficient. The denominator in both ESI and ESI* was the full set of test takers. The ESI* and GCSI were the only two indices proposed that accounted for guessing.

Helmstadter (1974) used Bayesian methods to propose three indices of effectiveness, B1, B2, and B3. Haladyna and Roid (1976, 1981) argued that these were indices of sensitivity because they required pretest and posttest scores. B1 was described as the probability that a student has knowledge, given that the student gets the item right. B2 was described as the probability that a student does not have knowledge, given that the student gets the item wrong. B3 was described as the probability of making a correct classification to mastery or nonmastery based on the item response.

$$B1 = (Diff_{post} * Diff_{com}) / ((Diff_{post} * Diff_{com}) + (Diff_{pre} * (1 - Diff_{com}))), \quad (8)$$

$$B2 = ((1 - Diff_{pre})(1 - Diff_{com})) / ((1 - Diff_{pre}) * (1 - Diff_{com}) + ((1 - Diff_{post})(Diff_{com}))), \quad (9)$$

$$B3 = Diff_{post} * Diff_{com} + (1 - Diff_{pre}) * (1 - Diff_{com}). \quad (10)$$

Here, a subscript of post refers to the posttest, pre refers to the pretest, and com refers to the combined sample. Particularly B3 was seen as potentially useful, because choosing items high on B3 would ensure a test well-suited for correctly classifying mastery and nonmastery students.

Finally, two variants on the traditional point biserial correlation were proposed. The first was the combined-samples point-biserial (COMPBI), where the data from the pre- and posttests were combined and the traditional point biserial correlation was calculated (Haladyna, 1974). The COMPBI was proposed because Haladyna viewed the PPDI as being attenuated in cases where the sample did not contain appreciable numbers of both masters and nonmasters. The second was the Internal Sensitivity Index (ISI), proposed by Kosecoff and Klein (1974). The purpose of the ISI was to measure the sensitivity of the item in the context of the overall test. The ISI was based on the contingency table in Figure 2, which shows the test pass-fail performance at pre- and posttest of all individuals who correctly answered item i at the posttest.

$$ISI = (X_{01} - X_{00}) / (X_{00} + X_{01} + X_{10} + X_{11}). \quad (11)$$

The authors viewed the ISI as being similar to a point biserial correlation.

		Posttest	
		Fail	Pass
Pretest	Fail	X_{00}	X_{01}
	Pass	X_{10}	X_{11}

FIGURE 2. Contingency table of pretest and posttest total score pass/fail for those passing a particular item at posttest.

Comparisons of item statistics. In light of the large number of proposed item statistics for evaluating item sensitivity, it is useful to summarize the findings about the relative merits of the indices. A number of studies provide comparisons of two or more indices (Haladyna, 1974; Haladyna & Roid, 1976, 1981; Helmstadter, 1972, 1974; Hsu, 1971; Kosecoff & Klein, 1974). There are also a number of studies that compare the sensitivity indices to other item statistics (e.g., Popham, 1971); here, however, the focus is on comparisons among sensitivity indices.

The earliest comparison of item sensitivity indices compared the Φ^* with the PPDI using pre- and posttest data from a group of second through sixth graders (Hsu, 1971). The goal was to examine the effects of sample characteristics on the relationship between the indices. Hsu found that the indices were highly correlated across items under all circumstances—almost always greater than .70, and often above .90—including when the items ranged in difficulty. He recommended either index to select items for criterion-referenced tests, cautioning that the sample characteristics might affect sensitivity.

In the first of Helmstadter's two studies comparing sensitivity indices (1972), he compared the COMPBI to the PPDI, based on a sample of 28 statistics students. He found that the two indices were correlated .78 across items, higher than the correlation of either index with a traditional discrimination index calculated on the posttest sample. He recommended the use of the more "conceptually satisfying" (p. 5) COMPBI, since both indices required the same data. In a second study (1974), Helmstadter introduced the Bayesian indices and compared them to the PPDI and COMPBI using pre-post measures on 43 statistics students and 55 psychology students. He found that the COMPBI and the PPDI were quite highly correlated ($r \geq .90$) across items, and that both were somewhat highly correlated with the B3 index (average $r \sim .68-.69$). Any of those three indices were deemed appropriate for use.

The ESI, ESI*, and ISI were compared with each other and with the Φ^* index in a study by Kosecoff and Klein (1974). Two separate data sets were used—one from a seven-item test given to a graduate experimental design class, and the other from a 70-item test given to a ninth-grade mathematics class. All three of the ESI, ESI*, and ISI were deflated with high levels of pretest mastery. Thus, a modified set of indicators was proposed and used. The modified indicators took the pretest mastery students out of the denominators of the respective formulae. The modified ISI was shown to be highly correlated

with the Φ^* ($r = .83$) across items on the 70-item test. Neither the ESI nor ESI* indicators were highly correlated with either the ISI or Φ^* ($r < .35$), perhaps because both ISI and Φ^* account for both item and total test performance, while ESI and ESI* use only item-level data. The authors concluded that the ISI was as appropriate as the Φ^* for selecting items.

Finally, three studies were conducted by Haladyna and Roid (Haladyna, 1974; Haladyna & Roid, 1976, 1981) to examine various indices. In the first study (Haladyna, 1974), 189 undergraduate education students completed pre- and posttests. The PPDI and COMPBI indices were compared, with results revealing high correlations (median $r = .75$) across the items on the multiple forms. Either index was recommended for use. In a second study (Haladyna & Roid, 1976), COMPBI, PPDI, ZDIFF, and the three Bayesian Indices were compared, using a pre-post assessment of 250 dental students. The correlations among PPDI, COMPBI, and ZDIFF were almost perfect ($r \geq .94$) across items, and the correlations of all three with B1 were also large, nearly .80 or greater. The B2 and B3 indices were less highly correlated with the other four, with correlations between .30 and .65. The authors conducted simulations comparing PPDI and B1, finding that B1 was generally higher than PPDI for items with high posttest difficulty; they interpreted this to mean that B1 was somewhat better at identifying sensitivity on highly difficult items at posttest. However, they recommended more work to examine this finding, and concluded that PPDI was the most conceptually simple way to measure instructional sensitivity.

In a third study (Haladyna & Roid, 1981), the authors used seven data sets with pre- and posttests and varying subject characteristics and instructional effectiveness to compare seven sensitivity indices and several difficulty and discrimination indices. The seven sensitivity indices were PPDI, PPG, COMPBI, ZDIFF, and the three Bayesian indices. As in previous studies, the PPDI, COMPBI, and ZDIFF were highly correlated with one another ($r \geq .75$) across items. The PPG and B3 were less highly correlated with these three ($r \geq .62$). The B1 and B2 had some correlations with the other indices as low as .38. For all seven data sets, the median intercorrelations among the four non-Bayesian indices were .58 or greater. Based on the results of the analyses, the authors eliminated the indices one at a time. The PPG was eliminated due to sensitivity to pretest difficulty. The ZDIFF was eliminated because it was the most complex to calculate and did not provide better information. The three Bayesian indices were eliminated due to sensitivity to various pretest or posttest conditions. The COMPBI was chosen as the second best option, because it was sensitive to restricted score ranges, and it was more difficult to compute than PPDI. Thus, the authors chose the PPDI, because of its ease of computation and interpretation, as well as it being the most consistently highly related to the other indices.

Conclusions about sensitivity indices. The sensitivity indices described here were conceived of to address perceived shortcomings in traditional item statistics. To ensure that the instructional sensitivity indices were not simply duplicating traditional item statistics, a number of researchers compared the two types of indices. For instance, Helmstadter (1972) correlated the posttest only point biserial (POSTBI), a traditional discrimination index, with the PPDI and COMPBI, finding that the correlations with POSTBI were less than .50.

Cox and Vargas (1966) compared a traditional upper-lower item discrimination index with PPDI, finding correlations of .37 to .40. Haladyna (1974) compared pretest-only point biserial to POSTBI and PPDI, finding correlations of .57 and .31, respectively. The most complete analysis was conducted by Haladyna and Roid (1976, 1981), who compared a number of difficulty and discrimination indices to the major sensitivity indices. They found that the sensitivity indices were not systematically related to the difficulty or discrimination indices, and argued that “instructional sensitivity is a unique item concept discernably related to item difficulty” (1981, p. 50).

While the item statistics described here were created for use in identifying sensitive items for building criterion-referenced tests, few studies actually used them for that purpose. One notable study (Crehan, 1974) used the PPDI and a modified Brennan Index to select items for a test, and compared the validity and reliability of the test with those of alternate forms constructed using items selected via other techniques (including teacher ranking, POSTBI, and random selection). The Brennan Index was modified by using posttest responses from a group that was not tested before instruction (i.e., the pretest and posttest group were not the same). The sample used for item selection was different from the sample used to calculate validity and reliability. The author also created a validity index, which they defined as the number of instructed students who passed the assessment and the number of uninstructed students who failed the assessment, as a proportion of the total number of test takers. Validity estimates were ranked, and the resulting ranks were subjected to nonparametric ANOVAs within each item pool. In five of six cases, the use of the PPDI and modified Brennan Index was found to result in tests with significantly higher validity indices than the other four selection methods. Perhaps not surprisingly, the Brennan and PPDI indices that were designed to estimate the effects of instruction were better able to distinguish instructed and noninstructed students than was a traditional item statistic. There was no difference in reliability across the selection techniques.

Based on the literature described above, several findings about the utility of the various indicators are possible.

- (1) The PPDI is straightforward to implement and understand, even for nonexperts in measurement, and results suggest that it performs well across a variety of circumstances.
- (2) Two studies suggest the ZDIFF is probably at least as useful as the PPDI. Since most current test developers use IRT methods, ZDIFF would be easy to implement.
- (3) The COMPBI performs as well as the PPDI in most cases, and it could be used interchangeably with the PPDI.
- (4) The B1, B2, B3, Φ^* , ISI, ESI, ESI*, and PPG indices do not have enough evidence to support their use.
- (5) Limited evidence suggests the use of PPDI in selecting items for tests results in a better ability to separate instructed and noninstructed students.

Certainly, there are limitations with the use of the sensitivity indices described here. The major limitation is that the methods are all quite dated, and they are based on classical test-theory techniques that are no longer used in large-scale testing. While the authors of the studies cited here often identified difficulties in estimating the various indices, modern computing capabilities erase this concern. A second limitation is that, unless the pre-post indices are calculated for both a treatment and a control group (i.e., a group that receives instruction and a group that does not), there will be threats to the internal validity of the inferences drawn. That is, it

will not be possible to tell whether the apparent instructional sensitivity is actually a result of high-quality instruction or, for instance, the effects of maturation. Another important limitation is that the sensitivity indices described above all require the use of pre- and posttest data or posttest data from an equivalent control group, which can be difficult to obtain.

One potential solution to some of these limitations is to use more modern IRT techniques to evaluate items for sensitivity. Though such methods have not been extensively studied to date, it is easy to see potential applications of IRT to sensitivity measures. One example is the use of ZDIFF, described above. For another example, differential item functioning (DIF) analyses (Holland & Wainer, 1993) could be conducted to compare instructed students to uninstructed students. An instance of the DIF methodology that used instructional data is discussed in the next section (Masters, 1988). If the collection of data on instruction was undesirable, another option would be to give the same test to a group of students at the beginning of a grade and a different group of students at the end of the grade. Or, a researcher examining sensitivity to a particular curriculum or intervention that was experienced by a subset of the population could simply give the test to treatment and control students simultaneously. In either case, a DIF analysis could then be used to indicate whether the items were sensitive to whatever the instruction was that students experienced. Such an analysis would have the advantage of providing tests of significance. However, it would not be an effective approach in the case where all items were sensitive or all items were not, because DIF analyses identify items operating differentially, controlling for total score. Furthermore, it would be hard to describe the instruction students received without analyzing data on it. Still, this approach would be at least as easy to do as ZDIFF, and could provide useful information.

Instructional Sensitivity Based on Instruction-Focused Methods

Early in the 1980s, researchers at Michigan State's Institute for Research on Teaching and UCLA's Center for Research on Evaluation, Standards, and Student Testing began work on a set of instructional sensitivity methods that differed from the previous, purely statistical approaches. Again, while these researchers were investigating sensitivity as defined in this paper, they generally used other terms, such as instructional validity. The most compelling rationale for this new line of work was presented by Airasian and Madaus (1983). They argued that the statistical methods of detecting instructional sensitivity touched on instruction only indirectly, claiming, "only when the post hoc methods of analysis are linked to specific information about instructional emphasis in the schools and classrooms studied can achievement differences be definitively linked to instructional factors" (p. 111). In large part, this dissatisfaction with the statistical sensitivity approaches was due to the increasing weight and consequences attached to assessments (Baker & Herman, 1983; Linn, 1983). Researchers argued that new approaches were necessary, ones in which tests were compared with the actual instruction students received in school, in order to ensure that students had an OTL the material tested.

In light of this new perspective on instructional sensitivity, researchers began formulating new methods of detecting sensitivity in assessments. A number of techniques were used, techniques that differed not only on their measures

of instruction, but also on the analytic methods. The measures of instruction used were reform-oriented instructional practice surveys (Wiley & Yoon, 1995), proportional or time measures of curriculum coverage (Brutten, Mouw, & Perkins, 1992; Hanson, McMorris, & Bailey, 1986; Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002), teacher-rated yes-or-no topic coverage (Masters, 1988; Muthen, 1987, 1988; Muthen et al., 1995; Muthen et al., 1991; Yoon, Burstein, & Chen, 1989; Yoon, Burstein, & Gold, 1991; Yoon & Resnick, 1998), a content-by-cognitive demand taxonomy (Freeman et al., 1983; Gamoran, Porter, Smithson, & White, 1997; Mehrens & Phillips, 1987), and observations or expert-rated alignment (D'Agostino et al., 2007; Greer, 1995). Analytical techniques ranged from simple comparisons of means (Wiley & Yoon, 1995) and ANOVAs (Brutten et al., 1992) to hierarchical linear modelings (HLMs) (D'Agostino et al., 2007; Yoon & Resnick, 1998) and modified IRT models (Muthen, 1987, 1988; Muthen et al., 1995; Muthen et al., 1991). However, all had the same basic approach—obtain some measure of the instruction to which students are exposed, and correlate the reported instruction with student performance on the assessments. In this section, the different methods for detecting item instructional sensitivity based on item statistics and teacher instruction are briefly discussed. To conclude, suggestions are made about the apparent utility of the various approaches.

Comparison of instruction-focused measures of sensitivity.

The most basic data on instruction that can be reported are teacher survey responses about generic instructional practices, including questions about the frequency with which teachers engage in various kinds of practices, assign certain kinds of problems, and lead certain kinds of in-class activities. Yoon and colleagues (Wiley & Yoon, 1995; Yoon & Resnick, 1998) used such survey data to evaluate the sensitivity of test scores. Wiley and Yoon (1995) used data from a teacher survey of their use of reform-oriented instructional practices to evaluate the reform-oriented California Learning Assessment System. Teachers first indicated how often they used particular reform-oriented methods emphasized on the assessment (e.g., identifying and using appropriate information for solutions) or taught broad content areas (e.g., algebraic reasoning strategies). The authors then separated the teachers based on their responses to each question into "high" (i.e., frequent) and "low" (i.e., infrequent) groups. Then, the authors calculated the difference in average student scores for students taught by teachers in the "high" group versus students taught by teachers in the "low" group. They found what they described as consistently positive evidence of sensitivity, especially in fourth grade, where 20 of the 24 low-high differences were significant, and all in the expected direction (i.e., more "reform-oriented" practices were associated with better student performance). However, the magnitudes of OTL effects were small—generally .20 points or less on the six-point test scale (no standardized effect sizes were offered).

In a later analysis, Yoon and Resnick (1998) created indices of teacher reform-oriented instruction based on teacher survey responses (e.g., frequency of lectures, performance assessments, and portfolio assignments) and examined achievement differences on a reform-oriented assessment using HLM. They found a great degree of sensitivity to differences in students' exposure to reform-oriented practices, such that between-classroom differences of .10 to .25 standard deviations were entirely explained by teacher use of those

practices. Both studies' results lead to the conclusion that surveys of instructional methods can be useful when the methods surveyed are tightly aligned to the methods assessed.

One step higher, in terms of data complexity, are reports of time-on-task or proportion of time spent covering particular curricula or curricular units. These are not tied to specific topics, but rather focus on the extent to which lessons have been taught or curricular materials have been used in a class. In one example (Ruiz-Primo et al., 2002), science units from the Full Option Science System (FOSS) textbook series were analyzed. Four types of student assessments (immediate, close, proximal, and distal) were administered, with the close and proximal assessments administered pre and post. Changes in student scores were correlated with estimates of the proportion of the FOSS unit activities covered. The authors found that the assessments were sensitive to amount of instruction—units that were covered the most also saw the greatest gains in students' scores—though effect sizes were not reported. In another study (Hanson et al., 1986), classes were grouped according to the number of units completed in a beginning reading program. Components of variance analyses were run on student test results; more than one-third and as much as 70% of the variance in achievement was due to the amount of the program taught. Furthermore, the technique helped identify certain "most sensitive" items to further increase the sensitivity of the assessment. In contrast to these two studies, a study of sensitivity using ANOVAs and the number of completed lessons found that the test was insensitive to the amount of student exposure to an English as a second language (ESL) curriculum (Brutten et al., 1992). It is not clear why there were differences in findings among these studies, but there are at least two possibilities. The first is that the ESL curriculum in the third study was ineffective, while the other curricula were effective. The second is that the test used in the third study (Brutten et al., 1992) was unable to detect the effects of instruction, while the other tests were. As with all cases where the assessment is found to be insensitive to instruction, there is no way to tell for sure which is true.

Another type of instructional data for sensitivity analysis is yes-no responses as to whether particular topics were taught. These OTL indicators were first used in international comparative studies to tie student scores to OTL across nations (McDonnell, 1995). These data have the advantage of being directly tied to individual items, and they can be analyzed using many techniques. The earliest analyses were based on the International Study of Achievement in Mathematics, with only graphical and correlational methods used (Husen, 1967). Teachers were asked whether the content of particular items was covered by the students taking the test. Then, average OTL was correlated with aggregated student performance, with correlations between .40 and .98. Teachers were also asked about their relative emphasis on tested topics, and graphical comparisons with student performance were displayed. Across populations, content emphasis was more highly related to achievement than any other measured variable.

A second way of analyzing yes-no OTL data was employed by Masters (1988) in examining the instructional sensitivity of the Pennsylvania mathematics and reading assessments. A representative sample of Pennsylvania teachers completed a survey indicating whether they covered the state's instructional objectives. Then, a DIF analysis of student test results

was conducted, comparing students in schools where *all* surveyed teachers reported covering the target objective to students in schools where *none* of the surveyed teachers reported covering the target objective. There was no DIF found for the reading items studied (a purposeful selection of items), while there was significant DIF on 12 of 21 mathematics items. Sensitivity effects were small to moderate in magnitude, with some standardized *p*-value differences as large as .18.

Using another IRT-based analysis method, Muthen and colleagues (Muthen, 1987, 1988; Muthen et al., 1991; Muthen et al., 1995) refined techniques for analyzing student test scores with data from the Second International Mathematics Study (SIMS), accounting for OTL. Muthen's (1987) IRT estimation method involved allowing the IRT difficulty parameter to vary across levels of OTL (in this case, OTL or not). Roughly half of SIMS items appeared to be sensitive to OTL (just 10–15% exhibiting strong OTL sensitivity); however, the larger effects appeared to be attributable more to OTL effects on the overall trait level rather than OTL effects on the probability of correctly responding to an individual item (Muthen, 1988). Of the few items that were highly sensitive, these were often definitional items (e.g., identifying an acute angle), where having OTL turned the problem into a very straightforward response (Muthen, 1988).

Yoon and colleagues used similar teacher response data and different analytical techniques to assess the effects of OTL in the Mathematics Diagnostic Testing Program (Yoon et al., 1989; Yoon et al., 1991). Again, teacher responses indicating yes-no content coverage were used. The authors calculated differences in students' *p*-values on particular items across years (e.g., comparing fourth grade students' *p*-values for a particular item in 1988 and the next fourth grade cohort's *p*-values on the same item in 1989). These differences were then compared with teachers' reports of change in content coverage across years. Few items and topics showed systematic differences in *p*-values across years, and there was no systematic relationship found between content coverage and student performance, indicating little if any instructional sensitivity (Yoon et al., 1989, 1991). Again, it is impossible to say whether the items were found to be insensitive to instruction because they were insensitive, or because the instruction was poor quality.

Somewhat more complex in terms of data requirements are measures that capture the content and cognitive demand of teachers' instruction. These methods were pioneered at Michigan State University's Institute for Research on Teaching in the early 1980s (e.g., Freeman et al., 1983) and are widely used today as part of the Surveys of Enacted Curriculum (SEC) (Porter, 2002). The methods include teacher self-reports of the amount of focus and cognitive demand level (e.g., memorize, make connections) on a set of fine-grained topics in a content taxonomy. The methods can be applied using taxonomies in English language arts and reading, mathematics, and science, and, combined with content analyses of assessments and standards using the same framework, can be used to compute indices of instructional alignment (Porter, 2002; Porter et al., 2007). These measures have been used to study the relationship between what is taught and what is tested. In one study (Gamoran et al., 1997), students' growth in achievement was regressed on course type and content alignment using HLM. Instructional alignment with the assessment was highly predictive of student achievement growth—along with binary course indicators, it explained

more than 70% of the between-class variance in achievement growth. Furthermore, with alignment in the model, the type of class no longer significantly predicted between-class differences in achievement. The results suggest that the particular test studied was highly sensitive to instructional differences among classes.

A modified framework was used in another study to examine the sensitivity of item difficulty on the Stanford Achievement Test to differences in textbook content coverage, often a proxy for instruction, among three popular mathematics textbooks used in one district (Mehrens & Phillips, 1987). The authors correlated *p*-values for students in different schools with differences in textbook content coverage of the textbooks they used. Most textbooks covered almost all items on the Stanford Test—80% or more when aggregated across grades 5–6 in all domain-textbook combinations except conceptual understanding for two of the texts. The authors determined that the small differences among textbooks were not significantly correlated with item *p*-values in 35 of 36 cases, suggesting that the Stanford Test was not sensitive to the minor differences in textbook content coverage. In this case, the finding of no sensitivity might be explained by the overall high content coverage of the textbooks. Alternatively, it may be the case that the way the textbooks are used, which was not studied, is what makes instruction effective. Or, the Stanford Test might not be sensitive to instruction.

The final and most complex method of collecting instructional data for analyzing instructional sensitivity is through observation and/or expert judgment. This method requires establishing a coding scheme, training coders, and coding for reliability. Two studies have used expert rating—one of teacher-reported instructional coverage (D'Agostino et al., 2007), and one of classroom observations (Greer, 1995)—to investigate instructional sensitivity. In Greer's (1995) study of three districts, third grade students' scores on the new Illinois Goal Assessment Program (IGAP) reading assessment were regressed on a set of more than 50 classroom interaction variables, such as the number and type of teacher-initiated questions and the nature of the feedback given from the teacher to students (p. 14). These variables were obtained from detailed observations undertaken by the research team. After controlling for entering ability and the identity of previous teachers, Greer found that 11 of the instruction variables were related to IGAP scores, but that no two districts shared any significant instructional activities. For instance, in one district, the number of background questions asked was a significant positive predictor; in the other districts it was not significant.

Last, in a study by D'Agostino et al. (2007), teachers were asked how much *emphasis* they put on the instructional objectives measured on the fifth-grade mathematics Arizona Instrument to Measure Standards (AIMS) using a four-point scale, like an extended version of the yes-or-no OTL items. Teachers were also asked about two objectives that were covered on the state test. They were asked to identify sample items reflecting the objectives and the key pieces of knowledge and skills they expected students to learn. Then, two content experts rated the *alignment* between the teachers' responses and the way the objectives were defined on the state test, using a four-point alignment scale. Finally, an HLM was used to examine the effects of emphasis, alignment, and their interaction on students' AIMS scores, controlling for prior achievement. Alignment and the emphasis/alignment interaction were significant predictors of achievement. The

authors concluded that the test was sensitive to instruction; however, since emphasis alone was not a significant predictor, the authors concluded that the assessment was sensitive to a narrow form of instruction—the type emphasized on the assessment.

Conclusions about instruction-focused measures of sensitivity. It is difficult to generalize from the list of studies described here. Each measured different facets of instruction, and each used different analytical techniques to accomplish their measurement. Nevertheless, there are a few trends that merit emphasis.

- (1) Measures of instruction based on teacher surveys of generic instructional practices may find that assessments are instructionally sensitive, particularly if the methods on the survey are closely aligned with the methods tested. However, the results of such analyses are probably not as useful for informing policy as more detailed surveys that ask about specific topics or more specific teaching methods.
- (2) Measures of instruction that focus on time-on-task or the proportion of a curriculum covered may be more useful than surveys of generic instructional practices. Again, these are most useful when the goal is evaluating the sensitivity of an assessment to a particular intervention or curriculum. It is also important to capture time-on-task in any instruction-related instructional sensitivity analysis, as it may be an important predictor of gains.
- (3) The yes/no OTL responses appear limited in terms of their utility in evaluating sensitivity. While some analyses identified effects using these OTL responses (Husen, 1967; Masters, 1988), most did not (Muthen, 1987, 1988; Muthen et al., 1991; Muthen et al., 1995; Yoon et al., 1989; Yoon et al., 1991). Furthermore, research using other techniques (e.g., D'Agostino et al., 2007; Gamoran et al., 1997) suggests that both content and pedagogy matter—*what* is taught, and *how* it is taught.
- (4) Measures of instruction based on classroom observations or detailed teacher reports about content taught are the most promising for measuring instructional sensitivity. It is critical for these techniques that actual instruction be examined rather than textbook use or other proxies. The best strategies are to develop and use consistent measures of instruction and alignment across studies, so that results can be compared. While observations are limited by high costs in terms of time and expertise, teacher surveys of instructional coverage are low-cost and reliable (Porter et al., 2007), though they focus on content only.

Instructional Sensitivity Based on Expert Judgment

The third type of instructional sensitivity measure proposed in the literature is the judgmental approach. Advocates of instructional sensitivity as an item or test characteristic have often advocated the use of judgment in concert with the statistical or instruction-focused measures of sensitivity described here (e.g., Haladyna & Roid, 1981). Some have gone further, arguing that empirical measures are “the least important kind of evidence to be considered in the construction of criterion-referenced tests. Far more important are the judgment of content specialists and data bearing on the content validity of test scores” (Hambleton, Swaminathan, Algina, & Coulson, 1978, p. 38; see also Millman, 1974; Popham, 2007). These

researchers have therefore developed a set of judgment-based sensitivity analyses.

Expert judgment methods. One group of approaches based on expert judgment uses expert rating of the alignment between the assessment and some set of objectives. For instance, Rovinelli and Hambleton (1977) discussed an approach where content experts evaluated each item on a three-point scale: +1 = definite feeling that an item is a measure of an objective, 0 = undecided about whether the item is a measure of an objective, and -1 = definite feeling that an item is not a measure of an objective. The authors then developed an index based on judges' responses and proposed that test developers could set a threshold of acceptable congruence based on the index. Two other techniques were also proposed. In the first, content specialists were asked to use a rating scale to rate the appropriateness of each item for measuring the particular objective it was written for. In the second, a matching task was proposed whereby experts were presented with a list of objectives and items and asked to match them. Then, a contingency table or chi-square analysis could be used to assess the overlap. There are numerous examples of expert-judgment techniques like these for determining the overlap of assessments and objectives or instruction. While these have been called tools to investigate sensitivity (Hambleton et al., 1978), without some data on student performance on those items it is hard to see how these techniques can be called investigations of item sensitivity. Rather, they are more like investigations of alignment, which could subsequently be used to conduct analyses of sensitivity with the collection of test data. A similar method is the Webb alignment procedure used by states to establish the alignment of standards and assessments (Webb, 1999).

One alternative judgment approach for identifying sensitivity has recently been proposed and endorsed by Popham (2007), in a departure from his previous support for empirical measures. Popham argued that tests used for accountability were inadequate, because they were insensitive to differences in instructional effectiveness. Little evidence was provided for this claim, except to say that "students' performances . . . are more heavily influenced by [their] socioeconomic status than by the quality of teachers' instructional efforts" (p. 147). Insensitive assessments undermine a fundamental assumption of accountability—that teachers and districts have it within their power to work more effectively and raise achievement. This leads to a pernicious cycle of teaching to the test that still results in poor test outcomes. Thus, Popham proposed the evaluation of the instructional sensitivity of standards-based assessments in order that assessments might be constructed that were more sensitive. He advocated the use of judgmental evidence, rather than empirical evidence, for practical reasons such as the difficulty and specificity of conditions under which empirical data must be collected. Empirical evidence, he suggested, should be of a confirmatory nature and, while desirable, is not necessary.

The judgmental approach he advocated is much like the panels used to conduct standard setting procedures. First, a collection of 15–20 content specialists and educators would be assembled and trained. Next, each panelist would review the items and rate the assessment on four 10-point scales. The first scale was the "number of curricular aims assessed," where a higher score indicated that the assessment measures a number of standards appropriate to be taught by teach-

ers in a year. The second scale was the "clarity of assessment targets," where higher scores were awarded to assessments offering clear frameworks for teachers to help them understand what was being assessed. The third scale was the number of "items assessed per curricular aim," where higher scores were awarded to assessments on which there were sufficient items for each objective that teachers could receive meaningful, reliable feedback about their performance in teaching each objective. The fourth and final scale was the assessment's "item sensitivity," where participants rated each item on: (1) whether a student's likelihood of responding correctly would be affected by her family's socioeconomic status; (2) whether a student's likelihood of responding correctly would be mostly determined by her innate aptitude; and (3) whether a majority of students would get the item correct if their teacher provided reasonably effective instruction on it. Together, responses on the four facets would provide a summary measure of the instrument's instructional sensitivity. Popham did not provide evidence that any sensitivity analyses using the techniques had been conducted, nor evidence that experts could make the sorts of judgments he recommended.

Conclusion about judgmental methods. While Popham offered no standards for evaluating the results of such a sensitivity review, such standards could likely be established. In all likelihood, the standards would parallel those recommended as best practice in establishing performance standards (Hambleton & Pitoniak, 2006). These would include explanation of the procedures for selecting participants and adequate description of training and procedures (AERA, APA & NCME, 1999). A bigger question is whether respondents could be trained to make the distinctions Popham suggests. Some evidence suggests that trained reviewers have difficulty detecting technical flaws in items (Engelhard, Davis, & Hansche, 1999), of which sensitivity may be a type. In any event, careful empirical work would be needed to support the claim that educators and content experts could discern instructionally sensitive items from those that are not instructionally sensitive. This would probably involve correlating results from the judgmental methods with measures of instructional quality from some external measure, along with statistical measures or instruction-focused methods discussed here. If this work were done, it is possible that the judgments as to the instructional sensitivity of items and instruments resulting from these panels would be as valid as the standard-setting judgments made in establishing NCLB proficiency levels. However, it is impossible to endorse these methods without evidence as to their feasibility and effectiveness—evidence that has not yet been produced.

Discussion

The purpose of this paper was to discuss instructional sensitivity as a psychometric property of items and assessments. To address this topic, I introduced and discussed the meaning of the term "instructional sensitivity." Next, I identified the various methods of estimating instructional sensitivity, grouping them into three major categories: statistical, instruction-focused, and judgmental. Within each category, the proposed methods were described and compared, and recommendations were made as to their utility. Thus far, while there have been comparisons of methods within categories, there has been no comparison across categories. As described

previously, the most promising item statistics are the PPDl and ZDIFF. The most promising instruction-focused techniques involve the use of teacher-reported content on detailed taxonomies and the observation and/or expert judgment approaches to evaluating instruction. The judgmental approach may prove useful, but there is no current evidence to support its use.

Even among the most promising methods, each has certain drawbacks. For the item statistics, these appear to have disappeared from wide use due to three issues. First, and most importantly, they are disconnected from instruction (Hambleton et al., 1978; Popham, 2007). This is the likely reason behind their replacement with instruction-focused methods. Second, the empirical results based on the statistical indices varied widely across settings and samples (Haladyna & Roid, 1981), suggesting an undesirable instability in the indices that may or may not have been attributable to the indices themselves. Because no work connected the sensitivity indices with measures of instructional quality, it is impossible to say whether the differences found in the sensitivity indices across tests were due to the assessments themselves or to the actual quality of instruction. Third, from a practical standpoint, the most useful indices required pre- and posttest data, which may have been expensive and unpalatable or raised red flags about testing effects on student performance.

In spite of these concerns, the available evidence, which is dated, suggests that the PPDl and ZDIFF are useful for detecting the effects of instruction. Researchers and evaluators should pursue the use of these indicators when pretests are feasible. To aid in implementation, random assignment of students to pre- and posttests can help ensure there are no testing effects. States should also consider evaluating items using these statistics when possible. Items can be embedded in assessments in consecutive years (e.g., students from a particular cohort take items as part of their third grade test one year and fourth grade test the next year), and the ZDIFFs or other indices calculated. The moderate degree of redundancy in state standards should ensure that all students will still be tested on material covered in the standards (Porter, Polikoff, & Smithson, 2009).

Though the instruction-focused techniques replaced the item statistics in research and in practice, it does not appear that they have caught on to any great extent, either. To be sure, there are a few instances of examinations of state assessments using the instruction-focused techniques (D'Agostino et al., 2007; Greer, 1995), but these are not the norm. The major limitation of the instruction-focused methods is that these approaches require more effort for teachers and researchers than statistical approaches. A second concern is that, while there are promising results in some instruction-focused sensitivity studies, these have been sporadic. The number of studies indicating large sensitivity effects (D'Agostino et al., 2007; Gamoran et al., 1997; Greer, 1995; Hanson et al., 1986; Ruiz-Primo et al., 2002; Yoon & Resnick, 1998) seems no greater than the number of studies with small or zero effects (Brutten et al., 1992; Mehrens & Phillips, 1987; Wiley & Yoon, 1995; Yoon et al. 1989; Yoon et al., 1991), though again it is possible that these differences are due to either the measures themselves or the quality of the instruction received. Likely the best way to evaluate whether the instruction or the test is the source of the insensitivity is to administer the assessment in many different settings, particularly ones where the quality or content of instruction is suspected or

known to differ. If sensitivity varies across settings, then poor instruction is the likely culprit. In contrast, if sensitivity is universally low no matter the setting, then the test is probably to blame. Despite the limitations of instruction-focused methods, it seems clear that, with high-quality data such as that obtained in detailed observations or teacher-reported content, measuring instructional sensitivity is possible. Work on the most useful instruction-focused sensitivity measures should move forward, and future work on sensitivity should include measures of instruction.

As for judgmental approaches, there is not enough evidence to support their use. Popham's (2007) approach may one day be useful, but substantial development work must be done first. Even if the approach is well supported by evidence, however, it may have limitations. Most importantly, judgmental methods run the risk of being too far removed from teachers' instruction. It is very likely that, in order to be accepted as useful, judgmental approaches would need to be validated against instruction-focused approaches. Additionally, judgmental approaches will be seen by some as lacking the objectivity of more statistical approaches. Detailed recording of methods and results will be necessary to help fight this charge.

Of course, two possibilities have been ignored to this point. The first possibility is that instructional sensitivity is a worthless concept that is unimportant in test development. However, the evidence described here contradicts this option, as some tests and items are clearly more sensitive than others. More importantly, if schools and individuals are to be held accountable for performance on assessments, we must be sure that it is within their power to improve results on those assessments, rather than relying on assumptions. The second possibility is that instructional sensitivity cannot be adequately measured by any of the techniques described here. Again, this seems unlikely. There is good evidence that several of the approaches presented here are quite useful for identifying sensitive items and assessments. To be sure, the evidence is limited by the relative dearth of research on sensitivity as compared to other features of assessments (e.g., difficulty, discrimination). Nevertheless, it is apparent that measures of sensitivity exist, and the techniques described here present a promising starting point.

Given this conclusion, why has instructional sensitivity not been deemed as important in test development as other test features? The most likely explanation begins with the replacement of statistical approaches to sensitivity with instruction-focused approaches due to a perceived lack of relationship with instruction. Once the instruction-focused approaches became the most favored, the early analyses found small effects because of poor specification of the instructional variables. Furthermore, OTL became a politically sensitive term when it was included in Goals 2000, and even by the mid-1990s high-quality work on instruction-focused techniques was still only beginning. Finally, the perceived burden of the various instruction-focused approaches was a challenge. As NCLB demanded more tests, the traditional principles of test development won out over the difficult-to-measure and seemingly less important sensitivity techniques.

Based on the work described here, there is ample evidence that instructional sensitivity is an important facet of any criterion-referenced assessment. It is even more apparent that sensitivity has been largely ignored as a feature of tests or items worth studying. This seems a grievous oversight, one

that threatens the validity of the thousands of decisions that are made annually based on results from state assessments under NCLB. Policymakers in particular should attend to these conclusions and think about ways to ensure that NCLB and other standards-based assessments are actually sensitive to instruction. Well-developed IRT techniques make statistical analysis of sensitivity using ZDIFF quite straightforward, and such methods could be used right away at little cost. While these techniques would have the advantage of focusing in on individual items, they would remain disconnected from actual instruction, which is a shortcoming.

The other best and most feasible method for evaluating sensitivity seems to be the correlation of student achievement gains with teacher-reported content coverage (Gamoran et al., 1997). For these methods, it is important to capture content at the fine-grained level of detail. The SEC, which can be used for these purposes, are low-cost and reliable for teachers to fill out (Porter, 2002; Porter et al., 2007), and teacher samples need not be overly large (just 48 in Gamoran et al., 1997). Furthermore, while SEC data are usually reported at the level of whole tests, they can also be reported at the level of individual items or subscales (Porter, 2002), so finer-grained analyses of sensitivity could be conducted than were done in the study by Gamoran and colleagues. In addition, the data on the SEC could be useful for states in examining other policy-relevant issues, such as the extent to which teachers are aligning their instruction with state standards and assessments.

States and researchers could implement one or both of these techniques immediately for the cost of teacher incentives and a statistician, and within one year's time there could be extensive data on the sensitivity of each state's standards-based assessments. Items shown to be highly sensitive, even in just a few settings, could be kept and used to guide further item development, while items that were insensitive across a variety of settings could be discarded. If the broad suggestions described here are followed, there is every reason to think that assessments can be built that will be fairer and will have greater power to detect instructional effects where they exist and greater precision in their classification of teacher and school effectiveness.

Acknowledgments

The author wishes to thank Andrew Porter and three anonymous reviewers for their helpful comments on earlier drafts of this paper.

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C050041-05 to the University of Pennsylvania. The opinions expressed are those of the author and do not represent views of the U.S. Department of Education.

References

AERA, APA, & NCME (1999). *Standards for educational and psychological tests*. Washington, DC: American Educational Research Association.

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

Baker, E. L., & Herman, G. L. (1983). Task structure design: Beyond linkage. *Journal of Educational Measurement*, 20(2), 149–164.

Brennan, R. L. (1972). A generalized upper-lower item discrimination index. *Educational and Psychological Measurement*, 32(2), 289–303.

Brennan, R. L., & Stolurow, L. M. (1971). *An empirical decision process for formative evaluation: Research Memorandum No. 4*. Cambridge, MA: Harvard University.

Brutten, S. R., Mouw, J. T., & Perkins, K. (1992). Measuring the instructional sensitivity of ESL reading comprehension items. *Southern Illinois Working Papers in Linguistics and Language Teaching*, 1, 1–9.

Cox, R. C., & Vargas, J. S. (1966, April). *A comparison of item-selection techniques for norm referenced and criterion referenced tests*. Paper presented at the Annual Conference of the National Council on Measurement in Education, Chicago, IL.

Crehan, K. D. (1974). Item analysis for teacher-made mastery tests. *Journal of Educational Measurement*, 11(4), 255–262.

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

Debra P. v. Turlington (1981). 644 F.2d 397 (U.S. Ct. App. 5th Cir.).

Engelhard, G., Davis, M., & Hansche, L. (1999). Evaluating the accuracy of judgments obtained from item review committees. *Applied Measurement in Education*, 12(2), 199–210.

Feldhausen, J. F., Hynes, K., & Ames, C. A. (1976). Is a lack of instructional validity contributing to the decline of achievement test scores? *Educational Technology*, 16(7), 13–16.

Freeman, D. J., Belli, G. M., Porter, A. C., Floden, R. E., Schmidt, W. H., & Schwille, J. R. (1983). The influence of different styles of textbook use on instructional validity of standardized tests. *Journal of Educational Measurement*, 20(3), 259–270.

Gamoran, A., Porter, A. C., Smithson, J., & White, P. A. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis*, 19(4), 325–338.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 18, 519–521.

Goals 2000: Educate America Act of 1994, Pub. L. No. 103–227, H.R. 1804, 103rd Cong., 2d Sess. (1994).

Goertz, M. E. (2005). Implementing the No Child Left Behind Act: Challenges for the states. *Peabody Journal of Education*, 80(2), 73–89.

Greer, E. A. (1995). *Examining the validity of a new large-scale reading assessment instrument from two perspectives*. Urbana, IL: Center for the Study of Reading.

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

Haladyna, T. M., & Roid, G. H. (1976, April). *The quality of domain-referenced test items*. Paper presented at the Annual Conference of the American Educational Research Association, San Francisco, CA.

Haladyna, T. M., & Roid, G. H. (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., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement*, 4th Ed. (pp. 89–115). Westport, CT: Praeger.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, 48(1), 1–47.

Hanson, R. A., McMorris, R. F., & Bailey, J. D. (1986). Difference in instructional sensitivity between item formats and between achievement test items. *Journal of Educational Measurement*, 23(1), 1–12.

Helmstadter, G. C. (1972, September). *A comparison of traditional item analysis selection procedures with those recommended for tests designed to measure achievement following performance oriented instruction*. Paper presented at the Convention of the American Psychological Association, Honolulu, HI.

- Helmstadter, G. C. (1974, April). *A comparison of Bayesian and traditional indexes of test-item effectiveness*. Paper presented at the Annual Conference of the National Council on Measurement in Education, Chicago, IL.
- Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Mahwah, NJ: Lawrence Erlbaum.
- Hsu, T. (1971, April). *Empirical data on criterion-referenced tests*. Paper presented at the Annual Conference of the American Educational Research Association, New York.
- Husen, T. (1967). *International study of achievement in mathematics: A comparison of twelve countries*. New York: John Wiley.
- Kosecoff, J. B., & Klein, S. P. (1974, April). *Instructional sensitivity statistics appropriate for objectives-based test items*. Paper presented at the Annual Conference of the National Council on Measurement in Education, Chicago, IL.
- Linn, R. L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement, 20*(2), 179–189.
- Masters, J. R. (1988, April). *A study of the differences between what is taught and what is tested in Pennsylvania*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.
- McClung, M. S. (1977). Competency testing: Potential for discrimination. *Clearinghouse Review, 11*, 439–448.
- McDonnell, L. M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis, 17*(3), 305–322.
- Mehrens, W. A., & Phillips, S. E. (1987). Sensitivity of item difficulties to curricular validity. *Journal of Educational Measurement, 24*(4), 357–370.
- Millman, J. (1974). Criterion-referenced measurement. In J. W. Popham (Ed.), *Evaluation in education: Current applications* (pp. 311–397). Berkeley, CA: McCutchan.
- Muthen, B. O. (1987). *Using item-specific instructional information in achievement modeling*. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing.
- Muthen, B. O. (1988). *Instructionally sensitive psychometrics: Applications to the second international mathematics study*. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing.
- Muthen, B. O., Huang, L., Jo, B., Khoo, S., Goff, G. N., Novak, J. R., et al. (1995). Opportunity-to-learn effects on achievement: Analytical aspects. *Educational Evaluation and Policy Analysis, 17*(3), 371–403.
- Muthen, 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.
- Popham, J. W. (1971). Indices of adequacy for criterion-reference test items. In J. W. Popham (Ed.), *Criterion-referenced measurement: An introduction* (pp. 79–98). Englewood Cliffs, NJ: Educational Technology Publications.
- Popham, J. W. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan, 89*(2), 146–155.
- Popham, J. W., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*(1), 1–9.
- Porter, A. C. (1995). The uses and misuses of opportunity-to-learn standards. *Educational Researcher, 24*(1), 21–27.
- Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3–14.
- Porter, A. C., Polikoff, M. S., & Smithson, J. (2009). Is there a de facto national intended curriculum? Evidence from state content standards. *Educational Evaluation and Policy Analysis, 31*(3), 238–268.
- Porter, A. C., Smithson, J., Blank, R., & Zeidner, T. (2007). Alignment as a teacher variable. *Applied Measurement in Education, 20*(1), 27–51.
- Roudabush, G. E. (1974, April). *Item selection for criterion-referenced tests*. Paper presented at the Annual Conference of the American Educational Research Association, New Orleans, LA.
- Rovinelli, R. J., & Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal of Educational Research, 2*(1), 49–60.
- Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. P. (2002). On the evaluation of systematic science education reform: Search for instructional sensitivity. *Journal of Research in Science Teaching, 39*(5), 369–393.
- Smith, M. S., & O'Day, J. (1990). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing* (pp. 233–267). London: Taylor & Francis.
- Walker, D. F., & Shaffarzick, J. (1974). Comparing curricula. *Review of Educational Research, 44*(1), 83–111.
- Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states*. Research Monograph No. 18. Madison, WI: National Institute for Science Education.
- 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., Burstein, L., & Chen, Z. (1989). *Patterns in teacher reports of topic coverage and their effects on math achievement: Comparisons across years*. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing.
- Yoon, B., Burstein, L., & Gold, K. (1991). *Assessing the content validity of teachers' reports of content coverage and its relationship to student achievement*. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing.
- Yoon, B., & Resnick, L. B. (1998). *Instructional validity, opportunity to learn, and equity: New standards examinations for the California mathematics renaissance*. Los Angeles, CA: Center for the Study of Evaluation.