INSTRUCTIONAL SENSITIVITY STATISTICS APPROPRIATE FOR OBJECTIVES-BASED TEST ITEMS*

Jacqueline B. Kosecoff

and

Stephen P. Klein

CSE Report No. 91
April 1974

Program for Research on Objectives-Based Evaluation
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California

*Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, April, 1974.

One of the typical steps in evaluating the quality of test items involves examining the degree to which student performance on the item is related to student performance on the total test. The basic assumption underlying this internal consistency approach to assessing item quality is that the total test score is the best available criterion of the degree to which students have mastered the content for which the test was designed. Thus, an item is considered "good" if it discriminates between high and low achievers in essentially the same way as does the total test.

Internal consistency indices of item discrimination, such as the point biserial correlation coefficient obtained between item and total test scores, have been used extensively in the construction of tests designed to make comparisons among students. Such indices are not maximally appropriate, however, for assessing item quality on measures designed to evaluate the effects of educational programs since discrimination indices are not uniquely sensitive to the effect of instruction. In other words, typical item discrimination in ices are so often influenced by a number of factors affecting test scores, such as general intellectual ability, that they may hide whether the item truly discriminates between those who have versus those who have not profited from the effects of instruction. This situation has given rise to a number of item sensitivity indices; that is, indices that reflect an item's sensitivity to instruction.

This paper describes several current attempts to provide some useful indices by which a test developer could judge the adequacy of his/her test items in terms of the extent to which the items reflect instruction. In addition, two new sensitivity indices are proposed, and the characteristics of these indices are compared to one another and to the traditional discrimination statistics.

It should be noted that efforts towards estimating item sensitivity have been associated almost exclusively with criterion-referenced testing situations (as compared to norm-referenced). This does not mean that item sensitivity indices are limited to situations where a test is to be interpreted using a criterion-referenced metric, but rather that criterion-referenced tests are thought to be more appropriate for the evaluation of instructional programs. Sensitivity indices should be associated with the question, "Can this item discriminate between learners and non-learners?" and not with whether the test is intended for criterion- or norm-referenced interpretation.

CURRENT SENSITIVITY INDICES

Cox and Vargas (1966) proposed a pretest-posttest difference sensitivity index that was obtained by computing "the percentage of students who pass the item on the posttest minus the percentage who pass the item on the pretest." Similar to the notion of raw gain, this index measures the percentage of students who had not mastered the item before instruction (at the pretest) but who had mastered the item after instruction (at the posttest). This index does not attend to how the item behaves with respect to the total test, to whether the item can discriminate between a group of students who actually learned and those who did not, or to corrections for guessing. Cox and Vargas correlated their index with the traditional discrimination index (top 27%--bottom 27%) and found rather low correlation coefficients suggesting fundamental differences between these indices.

Popham (1970) experimented with measuring changes which occur in items over an instructional period. He identified four possibilities: for any given learner, an item could be answered incorrectly on both the pre- and posttests (FF), correctly on the pretest but incorrectly on the posttest (PF),

incorrectly on the pretest but correctly on the posttest (FP), and correctly on both the pre- and posttests (PP). A situation characterized by a high percent of FP's was considered one reflecting learning, whereas a high percentage of PF's indicated negative learning.

Two statistics were explored. First for each item the percentage of students responding in each of the four ways was tabulated and items were ranked twice, according to highest percentage in the FP and PF categories respectively. When the two sets of rankings were compared, a negative correlation coefficient was obtained, suggesting a trend towards learning. Using this approach, an item is viewed as external to or independent of the total test. A second statistic, however, considered an item's homogeneity with the total test (i.e., an internal index). A 4xk Chi-square test (where 4 refers to the PP, PF, FP and FF categories and k to the number of items measuring the same objective) was conducted to measure the degree to which these items performed similarly with respect to the four possible response patterns. A non-significant test would indicate that all items performed similarly and thus reflected the effects of instruction in the same way. After field testing these statistics, Popham concluded that neither statistic represented an appropriate "red flag" for identifying items that fail to discriminate among learners.

Both Cox and Vargus' difference index and Popham's Chi-square statistic were employed by Ozenne (1971) to initially select items for a criterion-referenced measure. Ozenne's major focus, however, was not with the instructional sensitivity of a single item but with the total test. Using analysis of variance techniques Ozenne proposed a model that accounted for the variability of subject responses under a variety of criterion-referenced test situations and, in turn, lead to an estimate of the total test's sensi-

tivity to instruction. Using a true experiment (instruction versus no instruction) Ozenne's work provides the most sensitive test of the effects of instruction. In this paper the major concern is not to validate the total test nor to measure the impact of an instructional sequence, but rather to approximate an item's sensitivity to instruction.

Roudabush (1973) has suggested still another sensitivity index that considers the item response patterns used by Popham but also provides a correction for guessing. This model borrows from a procedure described by Marks and Noll (1967) developed for use in a slightly different context. This model is used again to develop new indices in another section of this paper and is presented there in some detail.

In terms of Popham's response categories Roudabush defines an item sensitivity index as

$$s = \frac{\hat{FP}}{\hat{FP} + \hat{FF}}$$

where the ^ denotes that the percentage of responses falling into each category that have been corrected for guessing. That is, FF represents the "true" percentage of learners who did not master the item and FP the "true" percentage of learners who did not know the item at the pretest but mastered it by the posttest. This index measures the proportion of students that missed the item on the pretest and then correctly responded to it on the posttest after a correction for guessing is applied; it does not, however, measure the "gain' in learners from the pretest to the posttest. Once again, s is computed independently of the total test score and thus serves as an external (to the test) or test-independent sensitivity index. This procedure was applied to a criterion-referenced reading test. Roudabush (1973) concluded, "Using sensitivity to instruction as the major criterion for item selection leads to

choosing a different set of items than would ordinarily be choosen (p. 11)" (using the traditional indices).

To summarize, current efforts have focused on comparing an item's response pattern prior to and post instruction. In most cases an item is considered independent of the total test and the resulting statistic can be described as an external sensitivity index. Field testing these indices have not yet lead to a single index that reflects the effects of instruction. However, one consistent result has emerged: sensitivity indices tend to select different items than their traditional counterparts.

NEW SENSITIVITY INDICES

In this section two sensitivity indices will be developed. The first statistic, an internal sensitivity index (ISI) measures an item's performance within the context of the total test, comparing how a given item and the entire test discriminate among learners. The second statistic, an external sensitivity index (ESI) measures an item independently of the total test, providing an estimate of an individual item's ability to assess learning. A correction for guessing (used as well by Roudabush) is provided for the ESI. The development of these statistics were guided by three criteria. An item sensitivity statistic must:

a. optimize ease in computation,

b. provide unique information, and

c. be relatively consistent with other general indices of item quality.

Internal Sensitivity Index (ISI).

Consider the pattern for pre- posttest performance among students who correctly responded to item i as depicted in Table 1. The total sample ($n_1 + n_2 + n_3 + n_4 = N_i$) represents the number of students who passed item i

at the posttest. The number of scores falling into cell (1, 1) reflects the frequency of students failing both the pretest and the posttest among those who correctly responded to item i. This is an undesirable outcome since item i has failed to identify a non-learning situation: students have remained non-masters after instruction and yet they correctly responded to item 1 on the posttest. Scores falling into cell (1,2), on the other hand, suggest a more desirable outcome. In this case students who correctly responded to item i on the posttest were non-masters before instruction and have reached mastery by the posttest. Cells (2, 1) and (2, 2) are situations in which students were already masters prior to instruction. In cell (2,1) students who had previously mastered the material based on a pretest, responded correctly to item i on the posttest but failed the total posttest indicating non-mastery or negative learning. Hopefully such situations are rare, particularly when pre- and posttests are close together in time providing little opportunity for forgetting. Finally, scores falling into cell (2,2) suggest that students who correctly answered item i on the posttest were able to demonstrate mastery both prior to and following instruction. Although this pattern is not undesirable in terms of item i's sensitivity, teaching already acquired skills is certainly questionable.

Insert Table 1 about here

To investigate the effects of instruction we need only study those students who fail the pretest (i.e., the students who are non-masters with respect to the total test prior to instruction).* With respect to posttest

*Note that this model assumes that a definition of mastery can be established. Some guidelines for mastery testing are put forth by Harris, 1974. In this same paper Harris also sets a precedent for considering selected portions of the data (as we do later in an alternate ISI).

TABLE 1

Distribution of Students Responding Correctly to Item i in terms of Pre- and Posttest Performance

	Fail Posttest	Pass Posttest	marginals
Fail Pretest	n_1 (1,1)	n_2 (1,2)	n_1+n_2
Pass Pretest	n_3 (2,1)	n_4 (2,2)	n_3+n_4
marginals	n_1+n_3	n_2+n_4	$n_1+n_2+n_3+n_4 = N_1$

where

n_1 = observed frequency of students who answered item i correctly on the posttest but failed the pre- and posttest

n_2 = observed frequency of students who answered item i correctly on the posttest but failed the pretest and passed the posttest

n_3 = observed frequency of students who answered item i correctly on the posttest but passed the pretest and failed the posttest

n_4 = observed frequency of students who answered item i correctly on the posttest and passed the pre- and posttests

scores, the proportion of students correctly responding to item i who failed the pretest but passed the posttest minus the proportion of students giving the correct response to item i who fail both the pre- and posttests provides a measure of an item's sensitivity to instruction. That is, a sensitivity index should discriminate among students (correctly answering item i) who were non-masters before instruction and masters after instruction. In formula notation this statistic can be expressed as:

$$(1) \qquad ISI = \frac{n_2 - n_1}{n_1 + n_2 + n_3 + n_4} = \frac{n_2 - n_1}{N}$$

If a passing score on the test is equated with mastery of the associated instructional objectives, then the ISI provides a measure of an item's ability to discriminate between those who have and have not profited from instruction.

The possible scores on the ISI range from -1 to +1. A score of -1 occurs when all students fail both the pre- and posttests but correctly respond to item i on the posttest. Certainly such an item is not sensitive to instruction and does not discriminate between masters and non-masters in a desirable fashion. On the other hand, a score of +1 is obtained when all students who properly answer item i on the posttest fail the pretest but pass the posttest. This is the ideal situation; item i can discriminate between students who are non-masters prior to instruction and masters after instruction.* Any scores in cells (2,1) and (2,2) (i.e., n_3 and/or $n_4 \neq 0$) will force the ISI to be less than +1. This is also a desirable property as students falling into

*It should be noted that the satisfaction derived from an index value of +1 is directly related to N_i (the number of students who passed item i on the posttest). It is possible that only one student passes item i at the posttest ($N_i = 1$) and that he (she) was a non-master at pretest and a master at posttest. In such a case ISI = +1, but in view of the value of +1, there is little cause for celebration.

this category are by definition masters prior to instruction and therefore should be directed to other instructional activities (rather than repeating already mastered materials).

External Sensitivity Index (ESI)

The ESI attends to item quality from a test-independent point of view. Once again let us turn to 4 possible categories of response to item 1 across pre- and posttest. The model for this approach depicted in table 2 closely resembles that for the ISI; however, like Roudabush and Popham, we now consider the responses to item i on pre- and posttest independent of total test performance. The total sample ($n_1+n_2+n_3+n_4=N_i$)* now represents all learners tested and the scores falling into cell (1,2), for example, reflect the frequency of students who miss item i on the pretest but pass item i on the posttest.

Insert Table 2 about here

The derivation of the ESI is analogous to that of ISI. Once again we are only concerned with students who were non-masters (in terms of item i) at the pretest, that is, those students falling into cells (1,1) and (1,2).

The proportion of students who were non-masters at the pretest but masters on the posttest minus the proportion of students who were non-masters at the pretest and remained non-masters at the posttest provides a second, test-independent measure of an item's sensitivity. In formula notation this

*Note that in this model N= the total number of students tested while in the model for the ISI, the denominator N_i = the number of students passing item i on the posttest.

TABLE 2

Response for Students Responding
to Item i Across Pre and Posttests

	Fail i on Posttest	Pass i on Posttest	Marginals
Fail i on Pretest	n_1 (1,1)	n_2 (1,2)	n_1+n_2
Pass i on Pretest	n_3 (2,1)	n_4 (2,2)	n_3+n_4
Marginals	n_1+n_3	n_2+n_4	$n_1+n_2+n_3+n_4+N$

n_1 = observed frequency of students who missed item i on the pretest and the posttest

n_2 = observed frequency of students who missed item i on the pretest but responded correctly on the posttest

n_3 = observed frequency of students who responded correctly to item i on the pretest but missed it on the posttest

n_4 = observed frequency of students who answered item i correctly on the pre- and posttest

statistic can be expressed as:

(2) $$ESI = \frac{n_2 - n_1}{n_1 + n_2 + n_3 + n_4} = \frac{n_2 - n_1}{N}$$

Comparing the formulas for the ISI and ESI, it is clear that these indices do not differ in computational form; however, each utilizes different types of frequencies (i.e., different definitions for n_1, n_2, $n_3 + n_4$) and consequently provides different kinds of information. The ISI measures item quality from the perspective of the total test's discriminating power while the ESI offers an individual estimate of how an item reflects learning.

Like the ISI, the values of the ESI can range from -1 to +1. A score of -1 would occur when no one learned; that is, each student failed item i on both the pretest and the posttest. Such a result suggests that either instruction failed to benefit any of the students or more realistically that the item fails to discriminate among learners. A score of +1 on the other hand is obtained when all students fail item i on the pretest but pass item i on the posttest. This is the ideal situation; item i shows maximum change in the direction of learning. Finally, any scores in cells (2,1) and (2,2) (i.e., n_3 and/or $n_4 \neq 0$) will lower the absolute value of the ESI.

Correction for Guessing for External Sensitivity Index. The ESI can be further redefined to correct for guessing. Traditionally a predetermined correction for guessing based on an item's format (e.g., the number of distractors in a multiple-choice test) is universally applied to all similarly formatted items in a given test. In this section an alternate formula for estimating the probability of guessing the correct response for a particular item is derived based on Marks and Noll.[1] This correction, based on the frequencies displayed in Table 2 rather than on item format, test length or

[1]Mark's and Noll's correction method was also applied by Roudabush.

other considerations, can assume different values for each item. Using this correction we can solve for the expected frequencies (or true values) of the cells in Table 2 and can derive an ESI that reflects any biases due to guessing.

We begin our derivation by making the folloiwng assumptions:

a. There is a non-zero probability, p, that a student who does not know the answer will guess correctly, where p is derived from observed data rather than a predetermined value based on the item's format.

b. Scores are independent from pre- to posttest (e.g., there is no systematic bias due to recollection of responses on the pretest).

c. There is no systematic forgetting between pretest and the posttest, and therefore $E(2,1) = E(n_3) \equiv 0$.

When deriving p, the probability of guessing the correct answer, we will refer to Table 2 and its notation. In addition, the following notation will be employed:

ν_1 = the true frequency of cell $(1,1)$[2] ; the number of students who legitimately did not know the answer to item i at both pre- and posttests (i.e., students who did not learn)

ν_2 = the true frequency of cell (1,2); the number of students who legitimately did not know the answer to item i at the pretest but then learned by the posttest (i.e., students who learned)

ν_3 = the true frequency of cell (2,1); the number of students who legitimately knew the answer to item i at the pretest but not at the posttest (note that according to assumption 3, we expect to find zero students in this cell, that is $E(n_3) = \nu_3 = 0$)

ν_4 = the true frequency of cell (2,2); the number of students who knew the answer to item i at both the pre- and posttests (i.e., students who always knew)

We are now ready to compute the expected cell frequencies and the value of p. Consider the cell (1,1). The observed frequency n_1 can be entirely

---

[2] In probabilitic terms, n_1 is the expected value of ν_1.

accounted for by those students who did not learn (ν_1) and guessed wrong on item i twice (on the pre- and posttests). The probability of guessing correctly at the posttest is p, and consequently the probability of making a bad guess is 1-p. Applying the multiplication rule for probability we have the probability of guessing wrong twice is $(1-p)^2$. Therefore, the observed n_1 can be expressed mathematically as:

$$n_1 = (1-p)^2\nu_1 \tag{3}$$

Equations for n_2,n_3 and n_4 can be derived using similar reasoning. The observed frequency in cell (1,2) can be explained by students who learned but guessed wrong on item i on the pretest $[(1-p)\nu_2]$ plus students who did not learn and guessed unsuccessfully on item i on the posttest $[(p(1-p)\nu_1]$. That is,

$$n_2 = (1-p)\ \nu_2 + p(1-p)\ \nu_i \tag{4}$$

Students falling in cell (2,1) are those who did not learn but guessed successfully on the pretest and unsuccessfully on the posttest $[p(1-p)\nu_1]$. Recall that we have assumed that students do not forget during instruction and consequently that the situation of knowing item i before instruction but not after instruction is impossible ($\nu_3 \equiv 0$). Therefore we have

$$n_3 = p(1-p)\nu_1 \tag{5}$$

Finally, the observed frequency in cell (2,2) can be accounted for by a combination of students who always knew (ν_4), students who learned and guessed correctly on item i on the pretest (ν_2), and students who never learned but guessed correctly twice ($p^2\nu_1$). This yields

$$n_4 = \nu_4 + p\nu_2 + p^2\nu_1 \tag{6}$$

From equations (3) and (5) we can solve for p.

(3): $n_1 = (1-p)^2\nu_1$

(5): $n_1 = p(1-p)\nu_1$

and therefore

(7) $$p = \frac{n_3}{n_1+n_3}$$

Proceeding in a similar fashion we can use equations (1) through (5) to find the following expected cell frequencies:

(8) $$\nu_1 = \frac{(n_1+n_3)^2}{n_1}$$

(9) $$\nu_2 = \frac{(n_2-n_3)(n_1+n_3)}{n_1}$$

(10) $$\nu_3 \equiv 0$$

(11) $$\nu_4 = n_4 - \frac{n_2 n_3}{n_1}$$

(12) $$\nu_1+\nu_2+\nu_3+\nu_4 = N \equiv n_1+n_2+n_3+n_4$$

A corrected external sensitivity index can then be computed:

(13) $$ESI^* = \frac{\nu_2-\nu_1}{N} = \left(\frac{n_1+n_3}{n_1 N}\right)[(n_2-n_3)-(n_1+n_3)]$$

Parallels to Traditional Indicators

The internal and external sensitivity indices have many similarities to traditional item statistics. First, both sensitivity indices range from -1 to 1 as do the item discrimination index (top 27% - bottom 27%) and the correlation coefficients.

Second, the categorical distribution underlying the sensitivity indices is structurally similar to the reliability coefficient (in that it can be viewed as the fraction of true outcomes to total outcomes for a particular definition of desirable performance). Students falling in the fail-pass (FP) and fail-fail (FF) categories (i.e., who fail the pretest) have scores that

can be influenced by instruction; these sources of score distribution can be compared with true score variation. Students falling in the pass-fail (PF) and pass-pass (PP) categories (i.e., who are masters prior to instruction) cannot be influenced by instruction; these sources of score distribution can be compared with error variation. Finally, the FP, FF, PF, and PP categories represent all possibilities for score distribution and can be compared with total score variation. Therefore, $\frac{FP+FF}{N}$, the proportion of score distribution that can be instructionally influenced, parallels the proportion of true-to-total score variation, that is, the reliability coefficient.

Finally, if one were to search for specific parallels to the ISI and ESI among traditional indices, the point biserial discrimination index and the phi coefficient respectively seem the most appropriate candidates. Both the ISI and point biserial measure the extent to which an item performs in concert with the total test. In the same fashion both the ESI and phi coefficient (between two items) measure the extent to which two items share similar response patterns. Computationally, the ESI can be thought of as a phi coefficient between item i on the pretest and item i on the posttest.

DATA APPLICATION

The ISI and ESI and two traditional indices (phi and point biserial) were computed using two sets of test data. The first was a 7-item multiple choice test measuring knowledge of Campbell and Stanley's research designs. This test, designed by the authors, was administered to their graduate level introductory statistics courses prior to and after instruction. The second data source was a 70-item multiple-choice test administered to 115 students before and after they received a ninth-grade mathematics program. The test used for this purpose was developed by a school district and was designed to

assess student performance on those objectives that the district considered to be most important at that grade level. For both tests a score above the test mean was considered to indicate mastery. (These levels reflect the test developers' suggestions, Harris (1974) has presented some guidance for establishing mastery levels).

The results of these efforts are displayed in tables 3 through 8. Because of the manageable number of items in the first 7-item test, a complete listing of intermediate results is provided for this measure in tables 3 through 6. Table 3 presents the item response patterns for the computation of the ISI. Each 2x2 matrix is analogous to Table 1, the numbers inside each cell are the n's for a given item. Similarly, Table 4 presents the analogue of Table 2, giving both the n's and ν's required for the computation of the ESI and ESI*. In Table 5, the values of the relevant statistics are displayed for each item.

A review of the values for the various indices reveals that the values of the ESI (both corrected and uncorrected for guessing) are quite low and that the ESI corrected for guessing is generally lower than its non-corrected counterpart. The ISI values are typically higher than the ESI and tend to parallel the point biserial and phi coefficients.

On the whole, the average sensitivity indices are quite low, suggesting at first glance that the test items were not particularly sensitive to instruction. However, upon a second, more careful inspection of the data, and in specific, the response patterns in tables 3 and 4, an alternate explanation emerges. We note that many students were masters prior to instruction as evidenced by the sizable frequencies in cells (2,2) (i.e., the values of n_4 were large). Frequently, as many as half the students demonstrated mastery of the materials at the pretest. Consequently, even though the difference between cells (1,1) and (1,2) was considerable (i.e., item i discriminated among

TABLE 3

Item-Response Patterns for Computation of Internal Sensitivity Index/7-item test*

item 1

Total Pretest Score	Total Posttest Score: fail	Total Posttest Score: pass
fail	1	11
pass	0	19

N=31

item 2

Total Pretest Score	Total Posttest Score: fail	Total Posttest Score: pass
fail	0	15
pass	0	23

N=38

item 3

Total Pretest Score	Total Posttest Score: fail	Total Posttest Score: pass
fail	2	21
pass	1	26

N=50

item 4

Total Pretest Score	Total Posttest Score: fail	Total Posttest Score: pass
fail	2	21
pass	1	26

N=50

item 5

Total Pretest Score	Total Posttest Score: fail	Total Posttest Score: pass
fail	0	19
pass	1	25

N=47

item 6

Total Pretest Score	Total Posttest Score: fail	Total Posttest Score: pass
fail	1	20
pass	0	26

N=47

item 7

Total Pretest Score	Total Posttest Score: fail	Total Posttest Score: pass
fail	0	19
pass	1	25

N=45

*Cells contain number of students passing each item on the postest

TABLE 4

Item Response Patterns for Computation of External Sensitivity Index/7-item test (numbers in parentheses correspond to ν's)

PRETEST \ POSTTEST	incorrect	correct
incorrect	10 (36.1)	9 (0)
correct	9 (0)	22 (13.9)

N=50

item 1

PRETEST \ POSTTEST	incorrect	correct
incorrect	9 (16.0)	28 (33.3)
correct	3 (0)	10 (.67)

N=50

item 2

PRETEST \ POSTTEST	incorrect	correct
incorrect	0 (0)	3 (3)
correct	0 (0)	47 (47)

N=50

item 3

PRETEST \ POSTTEST	incorrect	correct
incorrect	0 (0)	16 (16)
correct	0 (0)	34 (34)

N=50

item 4

PRETEST \ POSTTEST	incorrect	correct
incorrect	2 (12.5)	6 (7.5)
correct	3 (0)	39 (30)

N=50

item 5

PRETEST \ POSTTEST	incorrect	correct
incorrect	2 (4.5)	4 (2.5)
correct	1 (0)	16 (.50)

N=50

item 6

PRETEST \ POSTTEST	incorrect	correct
incorrect	2 (12.5)	31 (45)
correct	1 (0)	16 (.50)

N=50

item 7

TABLE 5
Item Statistics for 7-item Test

	Pretest	Posttest	ESI	ESI*	ISI	(item i with pass/fail posttest) PHI	(item i with posttest) PBIS
Item 1	$\bar{x}$=.62 SD=.49 N=50	$\bar{x}$=.62 SD=.49 N=50	-.72	-.02	.32	.15	.55
Item 2	$\bar{x}$=.26 SD=.44 N=50	$\bar{x}$=.76 SD=.43 N=50	.35	.38	.40	.45	.63
Item 3	$\bar{x}$=.94 SD=.24 N=50	$\bar{x}$=1.00 SD=0.0 N=50	.06	.06	.38	undefined	undefined
Item 4	$\bar{x}$=.68 SD=.47 N=50	$\bar{x}$=1.00 SD=0.0 N=50	.32	.32	.38	undefined	undefined
Item 5	$\bar{x}$=.84 SD=.37 N=50	$\bar{x}$=.90 SD=.30 N=50	-.10	.08	.42	.48	.51
Item 6	$\bar{x}$=.34 SD=50 N=50	$\bar{x}$=.94 SD=.24 N=50	.81	.58	.40	.65	.54
Item 7	$\bar{x}$=.88 SD=.33 N=50	$\bar{x}$=.90 SD=.30 N=50	-.20	.04	.42	.48	.58
Total	$\bar{x}$=4.56 SD=1.26 N=7	$\bar{x}$=6.21 SD=1.00 N=7	$\bar{x}$=.03 SD=.57 N=7	$\bar{x}$=.21 SD=.22 N=7	$\bar{x}$=.39 SD=.03 N=7	$\bar{x}$=.44 SD=.16 N=7	$\bar{x}$=.56 SD=.04 N=7

TABLE 6

Alternate Sensitivity Indices
(Adjusted for Masters Prior to Instruction)

	ISI	ESI*	ESI
Item 1	.83	-1.00	-.05
Item 2	1.00	.41	.51
Item 3	.83	1.00	1.00
Item 4	.83	1.00	1.00
Item 5	1.00	-.25	.5
Item 6	.90	.43	.88
Item 7	1.00	0.67	.33

TABLE 7

Summary Results for 70-item Test

	x	SD	N
Pretest	15.61	7.18	115
Posttest	31.08	11.90	115
ESI*	-.40	.30	70
ESI	-.18	.28	70
ISI	.12	.22	70
PHI (item with pass/fail on posttest)	.31	.16	70
P-BIS (item with posttest score)	.36	.16	70

TABLE 8

Correlations between Traditional and Sensitivity Indices, 70-item test

	ISI	ESI*	ESI	PHI	PBIS
ISI	1.00	-.07	-.22^{+}	.83	.82^{++}
ESI*		1.00	.88^{++}	.34^{++}	.32^{++}
ESI			1.00	.23^{+}	.21^{+}
PHI				1.00	.97^{++}
PBIS					1.00

learners and nonlearners) the large frequencies in cells (2,2) tended to reduce this effect.

In order to detect item sensitivity in this situation an alternate form of the indices was utilized in which the scores of students demonstrating mastery at the pretest were not taken into account in solving for the sensitivity indices. In computational terms, the values of n_3 and n_4 were removed from the denominator and the formulae for these alternate indices became*

$$ISI = (n_2 - n_1)/(n_1 + n_2) \quad (n_1, n_2 \text{ defined in Table 1}) \tag{14}$$

$$\text{(uncorrected)} \quad ESI = (n_2 - n_1)/(n_1 + n_2) \quad (n_1, n_2 \text{ defined in Table 2}) \tag{15}$$

$$\text{(corrected)} \quad ESI^* = (\nu_2 - \nu_1)/(\nu_1 + \nu_2) \quad (\nu_1, \nu_2 \text{ defined in (8) (9)}) \tag{16}$$

These values are presented in Table 6. The consistently high values for the alternate ISI confirm our suspicion that items were artificially deflated by a high proportion of prior masters and were indeed sensitive to instruction. On the other hand, the greatly varying values for the ESI tend to reduce our confidence in this statistic.

Inspection of Table 7 reveals a similar pattern in the 70-item exam. The values of the ESI and ESI* are quite low and vary considerably while the ISI values are higher, more consistent and tend to parallel the values for the phi and phis coefficients.

In Table 8, correlations between the various indices are presented for the 70-item test. (Correlations could not be computed for the 7-item test as N=7). The ISI was significantly correlated with both the p-biserial and phi coefficient. It would appear then that these 3 indices would tend to select many of the same items as "good" or bad. In contrast the correlates

*The use of partial data is not new to psychometrics. Harris (1974), for example, also considers selected data in his discussion of technical characteristics of mastery tests.

for the ESI with the phi and point biserial although significant, were rather small, suggesting that this index would not give the same judgment of an item as the traditional statistics. Apparently considering an item independently of the total test leads to very different results than viewing an item in terms of total test performance. Perhaps an item considered as a single, independent measure is not powerful and/or stable enough to discriminate among those who have and have not profited from instruction.

CONCLUSIONS

Two types of sensitivity indices were developed in this paper, one internal to the total test and the second external. To evaluate the success of these statistics we considered the three criteria suggested for a satisfactory index of item quality. The ISI appears to meet these demands. Certainly it is easily computed. In addition its moderately positive correlations with other traditional statistics confirms that the ISI provides unique information and yet is not inconsistent with these indices. However, when there are a large number of masters at the pretest an alternate form of the ISI is sometimes necessary to demonstrate item sensitivity. Finally, the theoretical construction of the ISI is both intuitively understandable and similar in form to other statistics. The ESI, on the other hand, does not fair as well as its internal counterpart. Although computationally simple it fails to demonstrate any consistent correlations with the traditional indices, suggesting a rather random statistic. Perhaps a single item (or an item viewed independently of the total test) is not sufficient to provide a stable, reliable measure of the effects of instruction.

In summary, the ISI appears to provide a suitable measure of an item's ability to distinguish between those who have and have not benefited from

instruction. Further, the most appropriate approach for evaluating item quality is an examination of the item in context with total test performance.

References

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.

Cox, R. C., & Vargas, J. S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1966.

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Marks, E., & Noll, G. A. Procedures and criteria for evaluating reading and listening comprehension tests. Educational and Psychological Measurement, 1967, 27, 335-348.

Ozenne, D. G. Toward an evaluation methodology for criterion-referenced measures: Test Sensitivity. CSE Report No. 72. Los Angeles: Center for the Study of Evaluation, University of California, Oct., 1971.

Popham, W. J. Indices of adequacy for criterion-referenced test items. A symposium presentation at the joint session of the National Council for Measurement in Education and the American Educational Research Association, New Orleans, 1973.

Roudabush, G. E. Item selection for criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, 1973.