DOCUMENT RESUME

| | | |
|---|---|---|
| ED 069 703 | | TM 002 155 |
| AUTHOR | Pyrczak, Fred, Jr. | |
| TITLE | Objective Evaluation of the Quality of Multiple-Choice Test Items. | |
| SPONS AGENCY | Office of Education (DHEW), Washington, D.C. Regional Research Program. | |
| BUREAU NO | BR-1-C-013 | |
| PUB DATE | Jun 72 | |
| GRANT | OEG-3-71-0109 | |
| NOTE | 43p. | |
| EDRS PRICE | MF-$0.65 HC-$3.29 | |
| DESCRIPTORS | College Students; Evaluation; *Item Analysis; Measurement Techniques; *Multiple Choice Tests; Statistical Analysis; Test Construction; Testing; *Test Validity | |

ABSTRACT

The basic objective of the study was to determine the validity of four new indices of item quality. Three of these were based on analyses of differential, empirical weights for item choices, and the fourth was designed to measure the relative attractiveness of distracters. A secondary objective was to ascertain the validity of the conventional discrimination indices. To attain these objectives, multiple-choice items designed to vary in quality with respect to nine common item-writing principles were prepared. The quality of each item was rated independently by three judges, and the average of their ratings was used as the criterion to determine the validity of the indices. The special test items were administered to a sample of college undergraduates, and the five indices were computed on the basis of their responses. The data were analyzed, and the conventional discrimination index was found to be a moderately valid measure of item quality. The weighted combination of the new indices also appeared to be valid. Because all of the new indices did not operate in the way expected, however, it is suggested that further research on them is necessary before they are considered for practical use in test-construction projects. (Author)

FINAL REPORT

PROJECT NO. 1-C-013
CONTRACT NO. OEG-3-71-0109

FRED PYRCZAK, JR.
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104

# OBJECTIVE EVALUATION OF THE QUALITY OF MULTIPLE-CHOICE TEST ITEMS

June 1972

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

Office of Education

National Center for Educational Research and Development
(Regional Research Program)

Final Report

Project No. 1-C-013
Contract No. OEG-3-71-0109

# OBJECTIVE EVALUATION OF THE QUALITY OF MULTIPLE-CHOICE TEST ITEMS

Fred Pyrczak, Jr.

University of Pennsylvania

Philadelphia, Pennsylvania 19104

June 1972

The research reported herein was performed pursuant to a contract with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
National Center for Educational Research and Development

## ABSTRACT

The basic objective of the study was to determine the validity of four new indices of item quality. Three of these were based on analyses of differential, empirical weights for item choices, and the fourth was designed to measure the relative attractiveness of distracters. A secondary objective was to ascertain the validity of the conventional discrimination index.

To attain these objectives, multiple-choice items designed to vary in quality with respect to nine common item-writing principles were prepared. The quality of each item was rated independently by three judges, and the average of their ratings was used as the criterion to determine the validity of the indices.

The special test items were administered to a sample of college undergraduates, and the five indices were computed on the basis of their responses.

The data were analyzed, and the conventional discrimination index was found to be a moderately valid measure of item quality. The weighted combination of the new indices also appeared to be valid. Because all of the new indices did not operate in the way expected, however, it is suggested that further research on them is necessary before they are considered for practical use in test-construction projects.

## ACKNOWLEDGMENTS

The writer is grateful to Frederick B. Davis, University of Pennsylvania, who provided invaluable assistance with every phase of this study.

Andrew Baggaley, James Diamond, and Dorothy Jones, all of the University of Pennsylvania, provided guidance with various aspects of this study.

Charlotte Croon Davis, Test Research Service; Gordon Fifer, Hunter College, City University of New York; and Mary B. Willis, American Institutes for Research, Palo Alto, served as judges of the quality of the test items used in this study.

The writer is also indebted to the administrators, staff, and students of the School of Education at the California State College, Los Angeles.

This study was made possible by a grant from the Regional Research Program of the National Center for Educational Research and Development (Region III).

Fred Pyrczak
Principal Investigator

University of Pennsylvania
Philadelphia, Pennsylvania

CONTENTS

# LIST OF TABLES

# CHAPTER 1

## MEASUREMENT OF ITEM QUALITY

### The Problem

To write multiple-choice items of high quality for aptitude and achievement tests requires a thorough knowledge of the subject matter and skills that are to be measured, highly developed writing skills, ingenuity in conceiving and casting testable ideas into proper form, and psychological insight into the probable reactions of different groups of examinees to the items. Because item writing is so complicated a skill and because the component characteristics of items of high quality have never been adequately defined, satisfactory measurement of item quality has been, at best, difficult to achieve.

### The Background of the Problem

In the past, attempts to measure item quality have made use of subjective judgments, indices of item difficulty, and indices of item-choice correlation with a criterion variable (usually total score on the test in which the item is included). Commonly, only the correlation coefficient between the dichotomy of marking or not marking the keyed choice and the criterion variable has been computed. Subjective judgments have been less than satisfactory, partly because a clear indication of important points to be considered has not been available to the judges and partly because of the inherent unreliability of judgments of the type involved. Conventional item-analysis data, on the other hand, sometimes are helpful in detecting defective items and aid in the selection and revision of items for inclusion in the final version of a test. Despite these attempts to measure item quality, inspection of achievement and aptitude tests indicates that a relatively large number of faulty items are not identified during test construction. Consequently, it seems desirable to investigate systematically the validity of conventional item-analysis data for measuring item quality as well as to examine the effectiveness of some new methods for measuring item quality that have received little attention in the past.

### New Measures of Item Quality

Three new indices, suggested by Davis (1959), incorporate information provided by conventional item-analysis data with information on the choice-criterion coefficients for unkeyed choices in a manner designed to make the resulting indices particularly sensitive to specific aspects of item quality. These three indices, plus a fourth devised for use in this study, were determined for each of 54 specially prepared items in such a way that their usefulness in judging item quality could be estimated and compared directly with the usefulness of conventional item-analysis data. Three sets of judgments of the quality of the 54 items were used as criteria of item quality. These judgments were made with the aid of a guide list of critical points to be considered in evaluating the quality of multiple-choice items.

## The Purpose of the Study

The basic problem for study is, then, the measurement of the quality of multiple-choice test items. Specifically, the effectiveness of the best-weighted combination of the new indices is compared with that of conventional item-analysis data.

## CHAPTER 2

## FAULTS IN MULTIPLE-CHOICE ITEMS: A SURVEY OF THE LITERATURE

### The Emphasis upon Avoiding Faults in Multiple-Choice Items

Wesman (1971) suggests that during the last two decades the emphasis on item-writing principles in textbooks on educational measurement has increased greatly. The current emphasis on this topic is revealed by a recent survey of the literature by Masonis (1971), which resulted in a list of forty-seven principles for writing multiple-choice items. Violation of most of these principles leads, logically, to the construction of faulty items (i.e., items of low quality). It is interesting to note that the list contains several contradictions that result from disagreements among item-writing experts regarding principles. However, there appears to be widespread agreement among experts on many of these principles. For example, thirty-four writers suggest that "All options should be plausible for the uninformed student" (Masonis, 1971, p. 93). In the study reported here, special items were written that vary with respect to nine item-writing principles. Eight of these principles appear in the list compiled by Masonis, and six of them were suggested by nine or more writers. The principle that items should be unambiguous is the only principle used in this study that is not explicitly included in the list, but it is implied by several of the other principles. The effects of following these widely recommended principles and thus avoiding certain faults, however, has received relatively little attention in the literature.

### The Effects of Faults on Scores on Multiple-Choice Tests

Some of the research on test-wiseness provides data on the extent to which examinees use certain kinds of faults to advantage in determining their responses to multiple-choice items. One approach that has been used to measure the variation due to the advantageous use of faults involves a comparison of the total scores obtained by groups of examinees on sets of items that are designed to measure the same points but which vary with respect to their quality. Millman and Setijadi (1966) used this approach to determine the extent to which test-wiseness exists in samples of American and Indonesian students. They used multiple-choice items with plausible distracters and multiple-choice items with implausible distracters. In general, the latter were easier than items with plausible distracters. Furthermore the difference in performance on the two types of items was greater for Americans, who were known to have had more experience responding to multiple-choice items, than for Indonesians. Although no tests of statistical significance were conducted, Millman and Setijadi suggest that the type of fault examined may have a differential effect on the performance of examinees with varying levels of test-taking experience.

Another approach that has been used to measure test-wiseness requires the construction of items that deal with very obscure or fictitious material and the incorporation of certain faults that examinees may use to raise their scores on the test above the level that most likely would be expected to occur as a result of chance alone. For example, some of the items used by Slakter et al. (1970b) to measure test-wiseness included one option each that resembled the stem of the question. The items dealt with fictitious content so that examinees could not answer the questions on the basis of knowledge. A test-wise examinee, in terms of these items, was defined as one who had a tendency to select options that resemble the stems. Significant over all differences were found among examinees in grades five through eleven on test-wiseness items that contained four types of faults, including the one described above. An important limitation of studies that measure test-wiseness in the manner just described is that the results may be appropriately generalized only to performance on tests that are extremely difficult, which is not typical of most tests used in educational situations.

In general, both approaches to the study of the particular aspect of test-wiseness under consideration have indicated that an important source of variation in test scores may be attributable to faults that are present in test items. These studies, however, only have been concerned with types of faults that may aid examinees in determining the keyed choices to multiple-choice items. It should be noted that some of the most serious faults in items make it more difficult for an examinee to select the correct choice even when he has a substantial amount of information about the point being tested. For example, an ambiguity in the stem of an item may mislead and cause a knowledgeable examinee to select an incorrect choice. In such items, there is no response that clearly should be chosen on the basis of the principles of test-wiseness alone. The study that is presented in this report is concerned with the identification of both types of faults in multiple-choice items.

Another limitation of these test-wiseness studies is that, in a strict sense, the results appropriately may be generalized only to items with faults that are similar in nature and degree. Wesman (1971) suggests that the limited generalizability of item-writing studies probably is responsible for the paucity of research in this area. The seriousness of this limitation with respect to one particular fault was demonstrated by Chase (1964). He found that when responding to very difficult items in which one choice in each item was longer than the others, examinees tended to choose the extra-long choices only when these choices were three times as long as the others. When these choices were only one-and-a-half to two times as long as the other choices, the extra-long choices did not appear to affect examinees' performance. Furthermore, the tendency to select choices that were three times as long disappeared when each of the difficult items with an extra-long choice was preceded by very easy items in which the extra-long choices were clearly incorrect. Thus, this study indicates that the widely-recommended principle that keyed choices should be no longer than the distracters may be an important principle in terms of its effect on examinee performance only under certain circumstances.

Despite the limitations discussed above, a sufficient number of such studies (e.g., Chase, 1964; Millman, 1966; Slakter, 1970; and Wahlstrom and Boersma, 1968) have identified variation in test scores apparently attributable to certain kinds of faults in test items to warrant the hypothesis that careful analysis of the responses of examinees may aid in the measurement of item quality. This hypothesis is consistent with much of the literature concerning the uses of the conventional discrimination index, which is reviewed in a later section.

Logically, if faults irrelevant to the points being tested account for some of the variation in responses to test items, tests composed of faulty items should be less valid than those composed of faultless items. Studies of test-wiseness generally have been concerned with the extent to which the trait exists among examinees and with its correlates (such as sex and grade) rather than the effects of such faults on the characteristic reliability and validity of tests. To the best of this writer's knowledge, only two studies have been conducted to determine the effects of various faults on these test characteristics. Dunn and Goldstein (1959) found that tests composed of items containing cues to the correct choice, extra-long correct choices, and inconsistencies in grammar between the stem and incorrect choices are less difficult than identical tests that do not have these characteristics. The presence or absence of these characteristics did not significantly affect the reliability or validity of any of the tests used in their study. Board and Whitney (1972), on the other hand, obtained somewhat different results in an unpublished investigation of the effects of four types of faults on test items. In general, they found that the faults that they examined benefited poorer students more than better students, that significantly lower reliability coefficients were obtained as a result of three types of faults, and that significantly lower validity coefficients occurred as a result of all four types of faults. The differences in the studies cited above suggest that the conditions under which faults affect test validity and reliability are not fully understood.

In spite of the contradictory evidence regarding the effect of item faults on test reliability and validity, certain principles of item writing are widely recommended by test-construction experts. The stress placed upon following these principles, in fact, may be justified solely in terms of their effect on the public acceptance of multiple-choice tests. For instance, items that have not been written in accordance with established item-writing principles are a source of concern to subject-matter specialists and scholars, such as Hoffmann (1962). Because a great deal of the criticism of multiple-choice test items springs from a lack of scholarly precision in writing and editing them, it appears important to conduct studies leading to the validation of conventional measures of item quality and to the development of new and, it is hoped, better measures.

## Identification of Faulty Items by Means of Conventional Item-Analysis Data

In formal test-construction projects, drafts of test items usually are administered to samples of examinees representative of those with whom the items ultimately are to be used. On the basis of examinee responses, estimates are made of each item's difficulty, of the attractiveness of each choice, and of the ability of each choice to discriminate among examinees of high and low ability in the trait to be measured. Many methods for arriving at these estimates have been proposed. The merits and deficiencies of the various estimates as well as their relationships to each other and to over all test characteristics have received a great deal of attention in the literature. Some of these considerations are discussed in the section of Chapter 3 that describes the conventional index of item quality used in this study. The basic purpose of this section of the report, however, is to review the literature that deals explicitly with the use of conventional indices to identify faults in individual test items.

If, as suggested in the previous section, some of the variation in test scores is attributable to faults in test items, the presence or absence of faults should affect the difficulty levels of individual test items. The use of item-difficulty indices to detect faulty items, however, is not straightforward because of two factors. First, a considerable amount of variation in difficulty indices normally is expected to occur as a result of the levels of abilities in examinees with respect to the points being tested. Furthermore, some types of faults, such as the presence of an ambiguity in the stem of an item, are likely to increase item difficulty while others, such as the inclusion of implausible distracters, are likely to decrease an item's difficulty from what it otherwise would be. In light of these considerations, it is not surprising that the use of information on item difficulty as an aid in detecting faults is not recommended in the literature.

Indices of choice attractiveness, on the other hand, apparently are more helpful in the process of identifying faulty items. Specifically, it has been suggested that distracters that are chosen by very few or none of the examinees should be regarded as implausible and be replaced (e.g., Adams, 1964, p. 357; Ahmann and Glock, 1971, p. 192; Henryssen, 1971, pp. 136-137; and Thorndike and Hagen, 1969, p. 127). Index 3, one of the new indices of item quality investigated in this study, is designed to provide an over all indication of the quality of each item with respect to the relative attractiveness of its distracters.

Apparently, indices of the extent to which items discriminate between those high in ability and those . w in ability on some criterion variable also may be used in identifying faulty items. Numerous writers suggest that items that discriminate poorly should be inspected closely for possible deficiencies (e.g., Anastasi, 1968, pp. 170-171; Davis, 1949, pp. 26-27; and Gulliksen, 1950, p. 365). To aid in the process of inspecting questionable items, it is commonly recommended that separate tabulations be made of the number of high-ability examinees and low-ability examinees who marked each choice. Illustrations of how faults may be detected in this manner are presented in the literature for items that have an unnecessary similarity between the keyed choice and the stem (Ahmann and Glock, 1971, pp. 193-194); that have distracters that may be too close in meaning to the keyed choice (Henryssen, 1971, pp. 136-137); that are tricky (Ebel, 1965, p. 369); and that are designed poorly (Ebel, 1965, p. 371).

Factors other than the presence or absence of faults may cause discrimination indices to vary from item to item. Misinformation on the part of examinees has been cited widely as one such factor (e.g., Anastasi, 1968, p. 170; Davis, 1951, p. 306; Ebel 1965, p. 372). Thus, despite the numerous individual illustrations in the literature showing how the discrimination index may be used to identify items with faults, their over all effectiveness as measures of item quality is not clear. One of the contributions of this study is that such a determination is made.

## CHAPTER 3

## A STUDY OF THE VALIDITY OF MEASURES OF ITEM QUALITY

### Questions to Be Answered

Multiple-choice items of low quality can be found in standardized achievement and aptitude tests despite the emphasis on avoiding faults in the literature and the widespread use of the conventional discrimination index in item-selection and revision procedures. In light of this fact, a clear need exists to investigate systematically the validity of the conventional measure of item quality as well as to determine the usefulness of some promising new measures of item quality. This study was undertaken to accomplish these objectives. The conventional discrimination index was expressed in terms of the Davis Discrimination Index. Three of the new indices selected for investigation are based on choice-weight scores and a fourth measures the relative attractiveness of distracters. All five indices were determined for each item in two parallel forms of a 27-item arithmetic reasoning test. The criterion for determining the validity of the indices was the average rating of each item's quality by three expert judges.

The data described above were obtained in order to answer the following specific questions:

1a. What is the validity of the conventional discrimination index for measuring item quality in each random half of the group of examinees on each form of the test?

1b. For each form, is the average of the validity coefficients obtained in the two halves of the examinees significantly different from zero?

2a. What is the validity of the best-weighted combination of the new indices for measuring item quality in each half of the sample of examinees on each form of the test?

2b. What is the relative contribution of each new index to the measurement of item quality in each half of the examinees on each form?

2c. What is the cross-validated multiple-correlation coefficient between the weighted composite of the new indices and the criterion in each half of the examinees on the two forms?

2d. For each form. is the average of the two cross-validated coefficients significantly different from zero?

3. For each form, is the average validity coefficient for the conventional index for both halves of the examinees significantly different from the average cross-validated multiple-correlation coefficient for the two halves?

## Construction of the Special Multiple-Choice Items

For the purposes of this study, it was necessary to write items that would be heterogeneous with respect to their quality. First, nine commonly recognized characteristics of items of high quality were identified, as follows:

1. Presence of an adequate keyed choice;
2. Absence of distracters that can be defended as adequately correct because of ambiguities in expressing the meanings of the stem and the choices;
3. Absence of distracters that can be defended as adequately correct when the stem and choices are unambiguous in meaning;
4. Absence of ambiguity caused by the use of a negative or double negatives;
5. Absence of distracters that are implausible because of a lack of homogeneity with each other and with the keyed choice;
6. Absence of distracters that are implausible when all choices are relatively homogeneous and the presence of naturally attractive distracters;
7. Absence of an extra-long or precisely worded keyed choice;
8. Absence of logically overlapping distracters;
9. Presence of grammatical agreement of the stem with the choices.

Next, two arithmetic-reasoning items were written to conform to the specifications represented by each of the nine characteristics. Thus, eighteen items of high quality were made available.

Then, two arithmetic-reasoning items were written in such a way as to make them slightly faulty with respect to each of the nine characteristics. Thus, eighteen items of medium quality were made available.

Finally, two arithmetic-reasoning items were written in such a way as to make them seriously faulty with respect to each of the nine characteristics.

The faults that were incorporated into the items needed to be of such a nature that they would not adversly affect the examinees' motivation and acceptance of the tests as legitimate measures of arithmetic-reasoning ability. Hence, there was a practical restriction on the extent to which the items could be made heterogeneous with respect to their quality. Three sample items are shown in Appendix A.

In summary, there were eighteen items designed to be "fault-free," eighteen designed to be moderately faulty, and eighteen designed to be seriously faulty. The items were matched in terms of the type and extent of fault, and one member of each matched pair of items was randomly selected for inclusion in form A of the test; the remaining item was included in form B. In constructing the parallel forms, an attempt was not made to match items in terms of the specific arithmetic reasoning skills that they were designed to measure.

## Administration of the Special Items to the Validation Group

The two parallel forms of the arithmetic-reasoning test, consisting of twenty-seven items each, were administered to undergraduates who were applying for admission to the teacher credential program during the fall quarter of 1971, at the California State College, Los Angeles. As part of the application procedure, students are administered a series of tests in various academic areas. They were informed that the test used in this study was experimental and was being administered in order to determine how well the test worked. They were told, furthermore, that the experimental test would provide them with practice in some of the skills that they would need to use on an arithmetic skills-and-concepts test that would be administered to them about a month later. The latter test is considerably easier than the one used in this study and is used to determine eligibility for the credential program. Observations of the examinees while they were taking the experimental test indicated that they were well motivated.

The two forms were administered separately with one week between administrations. Ninety-nine of the examinees were present for the administration of only one of the forms, and their responses were excluded from all analyses. Since conventional item-analysis may not be meaningful if the data are obtained under speeded conditions, the responses of the forty-two examinees who did not mark at least one of the last three items on both forms were also excluded. Consequently, the results reported in this study are based upon the responses of 364 examinees who marked an answer to at least one of the last three items on both forms.

## Computation of the Measures of the Quality of the Special Items

In order to compute the conventional discrimination index and three of the four new measures of item quality, a criterion measure of the examinees' over all ability in arithmetic reasoning was needed. In this study, scores on the nine "fault-free" items in one parallel form were used as the criterion in computing the indices for each item in the other parallel form. These scores were corrected for chance success. The parallel-forms reliability coefficients for the two nine-item forms were found to be .562 and .579 in two non-overlapping random halves of the examinees. Since only two hours of testing time were available, it was not possible to include a larger number of items intended to be "fault-free".

Conventional discrimination indices, which estimate the degree of relationship between marking or not marking the keyed choice and scores on the criterion variable, were obtained for each item by means of an item-analysis computer program. This program expresses the discrimination index in terms of a point-biserial correlation coefficient. Since an "external" criterion was used (i.e., scores on the nine items designed to be "fault-free" in a separately administered parallel form), the spurious inflation of these coefficients that would have occurred if part-whole correlations had been used was precluded. It is widely recognized, however, that the values of the point-biserial coefficient are related to item difficulty. In order to obtain a measure of item discrimination that is less related to item difficulty, the point-biserial coefficients were converted to biserial correlation coefficients. The biserial coefficients subsequently were converted to Davis Discrimination Indices, which are described in detail elsewhere (Davis, 1949). The essential characteristics of these indices are that their values constitute an interval scale and range from 0 to 100.

The four new indices of item quality were computed. These are:

Index 1 ($I_1$)

$$I_1 = \sum_{(i=2)}^{k} (C_1 - C_i)$$

where

$I_1$ is Item Quality Index 1;

$C_1$ is the choice weight for the keyed choice;

$C_i$ is the choice weight for choice i, where "omits" are treated as choices and $i \neq 1$;

k is the number of choices.

If the choice weights are on a 7-point scale from +3 to -3, for five-choice items, the maximum value of Index 1 is +24 (where $C_1$=3 and $C_i$=-3 for all values of i); the minimum value is -24. The higher the value of $I_1$, the more likely it is that those who are high on the criterion variable are attracted to the keyed choice and that those who are low on the criterion variable are attracted to the distracters. Thus, the value of $I_1$ for any given item indicated that the extent to which it differentiates between those who know the answer and those who do not. It has long been an accepted principle of item writing that items should make this differentiation, and the extent to which they do is shown by Index $I_1$. From this basic principle are derived many specific rules for item writing.

Index $1_m$ ($I_{1m}$)

$$I_{1m} = C_1 - C_a$$

where

$C_a$ is the choice weight for the most attractive distracter.

If the choice weights are on a 7-point scale from +3 to -3, the maximum value of Index 1m is +6 and the minimum value is -6. The higher the value of this index, the more likely it is that those who are high in ability are attracted to the keyed choice and that those who are low in ability are attracted to the most attractive distracter. This index is a modification of Index 1 and was devised after an inspection of the choice weights and frequencies for the choices in several of the experimental items in Form B. This inspection revealed that in some items several distracters were selected by very few examinees. In computing Index 1, the choice weights for such ineffective distracters were given equal weight with highly effective distracters. Index 1m is less subject to this problem.

Index 2 ($I_2$)

$$I_2 = \sum_{(i=2)}^{k} \sum_{(j=3)}^{k} \left| C_i - C_j \right| \qquad (i \neq j)$$

If the choice weights are on a 7-point scale from +3 to -3 for five-choice items, the maximum value of $I_2$ is +24 (where $C_2$ and $C_3 = 3$ and $C_4$ and $C_5 = -3$). The maximum value is zero (when all values of C are the same). Therefore, the higher the value of $I_2$, the more likely it is that the distracters are attracting groups of subjects who differ with respect to their mean criterion scores. The basic assumption underlying the formulation of this index is that in an item of high quality, the distracters should discriminate among those who don't have sufficient knowledge to select the correct response but have varying amounts of information or misinformation. That is, each distracter should attract examinees at a different average level of ability than the other distracters. It is assumed that items with this characteristic will be especially effective in terms of providing plausible distracters for examinees who do not thoroughly know the point in question. It should be noted, however, that when the value of this index is at a maximum, the value of Index 1 cannot be at a maximum. This restriction does not apply to Index 1m.

Index 3 ($I_3$)

$$I_3 = -\sum_{(i=2)}^{k} \left| \frac{\sum_{(1=2)}^{k} f_i}{k-1} - f_i \right|$$

where

$f_1$ is the frequency for the keyed choice;

$f_i$ is the frequency for choice i, where "omits" are not treated as choices;

k is the number of choices.

Whatever the percent of examinees who choose the keyed choice, $I_3$ will equal zero when equal percents mark all distracters. Its value will be larger in the negative direction when this condition does not exist. Index 3, therefore, is a measure of the extent to which the distracters in a multiple-choice item are equally attractive. Horst (1933) has shown that, other things being equal, item scores will tend to be more reliable for items where the distracters are more equally attractive than in items where they are less equally attractive. This occurs regardless of the level of difficulty. Furthermore, it is widely recommended that distracters that attract very few or no examinees probably should be regarded as implausible and be replaced. Items with such distracters will tend to have larger negative values on Index 3 than items that do not.

## Judgments of the Quality of the Special Items

In order to obtain a criterion to use in determining the validity of the objective measures of item quality, three judges were asked to rate independently each of the fifty-four special items for quality by using a special check list of the nine characteristics of items discussed earlier (See Appendix B).* Specifically, the judges were asked to indicate which, if any faults were present in each item and the extent to which each fault would be likely to affect adversely a given item's ability to discriminate between those who know and those who do not know the point in question. The extent to which each item's ability to discriminate was impaired by each fault was indicated on a three-point scale consisting of these categories: "not detrimental," "moderately detrimental," and "seriously detrimental." Furthermore, the judges were asked to explain the nature of each fault that they found.

Originally, it was planned to give each item a score of one point for each moderately detrimental fault and a score of two points for each seriously detrimental fault. In 20 of the 171 ratings of individual items, however, a given judge gave the same explanation for marking two or more faults for a given item. This occurred most often in response to scales five and six even though these scales were worded in a manner designed to preclude this occurrence. This raised the problem of whether an item should accumulate points under various headings on the check list for a single characteristic. It finally was decided that whenever two or more faults were marked for a given item by a single judge and the same explanation was given for the various faults, the multiple faults would be counted only once. It is interesting to note, in

*Charlotte Croon Davis, Test Research Service; Gordon Fifer, Hunter College, City University of New York; and Mary B. Willis, American Institutes for Research, Palo Alto, served as the judges of the quality of the items.

retrospect, that this problem could have been avoided by having the judges check off the faults that they found in each item, but give only one over all rating of the likely effects of all faults on each item's ability to discriminate.

. The scores obtained by each item were averaged in order to obtain a single criterion measure of item quality. To make higher average values indicate higher quality than lower average values, the average scores for each item were subtracted from a constant positive number that was larger than any of the average ratings.

Despite the relatively minor problem that arose in obtaining scores from the ratings, inspection of them reveals that the judges possess considerable insight into the desirable characteristics of multiple-choice items and the probable reactions of examinees to them. Furthermore, the reliability coefficients for the average of the judges' ratings were computed to be .667 and .857 for forms A and B, respectively, which are high considering the types of judgments involved.

## CHAPTER 4

## THE FINDINGS

For each of the specially prepared items, six scores were available: Indices $I_1$, $I_{1m}$, $I_2$, $I_3$; the Davis Discrimination Index (DDISC); and the Item-Quality Rating (IQR) obtained by averaging the ratings of the three judges. The sample of examinees was divided in half at random for subsequent cross validation, and the indices were computed separately on the basis of the responses of random halves (I and II) of the examinees on each form (A and B) of the test.

The criterion used in computing Indices $I_1$, $I_{1m}$, $I_2$, and the Davis Discrimination Index for each item consisted of the scores on the nine items intended to be "fault-free" on the parallel form of the test. The mean scores corrected for chance success on the nine items on Form A of the test were 3.30 and 2.89 for halves I and II of the examinees; the associated standard deviations were 2.37 and 2.32, respectively. On Form B, the mean corrected scores on the nine items were 2.25 and 1.77, and the standard deviations were 2.45 and 2.41 in halves I and II, respectively.

The mean corrected scores on all twenty-seven items on Form A of the test were 8.92 and 9.31 for the two halves of the examinees; the associated standard deviations were found to be 4.49 and 4.35, respectively. On Form B, the mean corrected scores on all items were 8.99 and 9.78, and the standard deviations were 4.79 and 4.49.

The first step in the analysis of the data was to obtain the intercorrelations of the six indices of item quality separately for each random half of the examinees on each form of the test. Tables 1 through 4 present the intercorrelations along with the means and standard deviations of the variables. The columns labeled "IQR" show the validity coefficients for the indices. Inspection of the scatter plots for these relationships indicate that they are not curvilinear.

### The Validity of the Conventional Index of Item Quality

The conventional discrimination index was expressed in terms of the Davis Discrimination Index (DDISC). For Form A of the test, the validity coefficients for this index were .540 and .418 for the two random halves of the examinees. Using the appropriate $\underline{z}$ transformation, the average of these coefficients was .485. This value is significantly different from zero at the .01 level.

For Form B of the test, the validity coefficients for the Davis Discrimination Index were .488 and .572 for the two halves

TABLE 1

INTERCORRELATIONS, MEANS, AND STANDARD DEVIATIONS FOR THE SIX VARIABLES FOR RANDOM HALF I ON FORM A OF THE TEST.

| | $I_1$ | $I_{1m}$ | $I_2$ | $I_3$ | DDISC | IQR | M | SD |
|---|---|---|---|---|---|---|---|---|
| $I_1$ | | .675 | -.285 | .370 | .824 | .482 | 53.185 | 32.550 |
| $I_{1m}$ | | | -.084 | .545 | .901 | .489 | 6.926 | 7.817 |
| $I_2$ | | | | -.082 | -.100 | .267 | 77.111 | 46.416 |
| $I_3$ | | | | | .623 | .126 | -54.815 | 49.062 |
| DDISC | | | | | | .540 | 15.852 | 11.430 |
| IQR | | | | | | | 3.444 | 1.207 |

## TABLE 2

INTERCORRELATIONS, MEANS, AND STANDARD DEVIATIONS FOR THE SIX VARIABLES FOR RANDOM HALF II ON FORM A OF THE TEST.

| | $I_1$ | $I_{1m}$ | $I_2$ | $I_3$ | DDISC | IQR | M | SD |
|---|---|---|---|---|---|---|---|---|
| $I_1$ | | .620 | .154 | -.048 | .771 | .456 | 62.630 | 38.458 |
| $I_{1m}$ | | | -.191 | .456 | .876 | .456 | 9.296 | 9.633 |
| $I_2$ | | | | -.431 | -.226 | .121 | 91.593 | 41.284 |
| $I_3$ | | | | | .456 | .071 | -52.148 | 50.430 |
| DDISC | | | | | | .418 | 16.889 | 12.673 |
| IQR | | | | | | | 3.444 | 1.207 |

## TABLE 3

INTERCORRELATIONS, MEANS, AND STANDARD DEVIATIONS FOR THE SIX VARIABLES FOR RANDOM HALF I ON FORM B OF THE TEST.

| | $I_1$ | $I_{1m}$ | $I_2$ | $I_3$ | DDISC | IQR | M | SD |
|---|---|---|---|---|---|---|---|---|
| $I_1$ | | .827 | .010 | .192 | .897 | .417 | 58.444 | 38.539 |
| $I_{1m}$ | | | -.298 | .343 | .939 | .593 | 9.518 | 9.658 |
| $I_2$ | | | | -.297 | -.188 | -.473 | 94.593 | 41.139 |
| $I_3$ | | | | | .361 | .423 | -50.963 | 39.693 |
| DDISC | | | | | | .572 | 15.333 | 11.829 |
| IQR | | | | | | | 3.790 | .987 |

TABLE 4

INTERCORRELATIONS, MEANS, AND STANDARD DEVIATIONS FOR THE SIX VARIABLES FOR RANDOM HALF II ON FORM B OF THE TEST.

| | $I_1$ | $I_{1m}$ | $I_2$ | $I_3$ | DDISC | IQR | M | SD |
|---|---|---|---|---|---|---|---|---|
| $I_1$ | | .736 | .011 | .157 | .870 | .360 | 66.296 | 32.592 |
| $I_{1m}$ | | | -.330 | .305 | .922 | .414 | 10.778 | 9.378 |
| $I_2$ | | | | -.232 | -.331 | -.287 | 85.037 | 35.546 |
| $I_3$ | | | | | .182 | .356 | -47.518 | 41.517 |
| DDISC | | | | | | .488 | 18.222 | 11.789 |
| IQR | | | | | | | 3.790 | .987 |

of the examinees. The average of these coefficients was .530, which is significantly different from zero at the .01 level.

In summary, the conventional discrimination index was positively correlated with the criterion variable at the .01 level. It is important to note, however, that approximately three-quarters of the variation of item quality, as determined by judges' ratings, remained unexplained by this index.

## The Validity of the New Indices of Item Quality

Tables 1 and 2 show the intercorrelations, means, and standard deviations of the indices for the two random halves of the examinees on Form A of the test. With respect to this form, all of the new indices have positive validity coefficients. Only the coefficients for Indices $I_1$ and $I_{1m}$, however, are of appreciable size.

Tables 3 and 4 show the intercorrelations, means, and standard deviations of the indices for the two random halves of the examinees on Form B of the test. With respect to this form, $I_2$ has negative validity coefficients. Possible reasons for this unexpected finding and suggestions for a future study of this index are discussed in the next chapter.

The multiple-correlation coefficients between the best-weighted combination of $I_1$, $I_{1m}$, $I_2$, and $I_3$ and the Item-Quality Rating for the two halves of the examinees on Form A were .693 and .632, respectively. On Form B the multiple-correlation coefficients for the two halves of the examinees were .689 and .544, respectively.

To eliminate the capitalization on chance elements that causes spurious inflation of multiple-correlation coefficients, cross-validated correlation coefficients were obtained by using the beta weights obtained in half I of the sample with the intercorrelations and validity coefficients of the variables in half II of the sample. Likewise the beta weights obtained in half II of the sample were used with the intercorrelations and validity coefficients of the variables in half I of the sample.

Strictly speaking, the resulting cross-validated coefficients are product-moment correlation coefficients between standard measures in the criterion variable (denoted in the following equations as c) and a weighted sum of standard measures in each of the predictor variables (the four Indices, denoted in the following equations as variables 1, 2, 3, and 4) where the weights are the partial regression coefficients in standard-measure form (beta weights denoted in the following equations as $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$). The equation for obtaining the cross-validated coefficients, written for sample I intercorrelations and validity coefficients and sample II beta weights in Form A of the test, follows:

$$
{}_A r_{({}_I z_c)({}_{II}\beta_1\,{}_I z_1 + {}_{II}\beta_2\,{}_I z_2 + {}_{II}\beta_3\,{}_I z_3 + {}_{II}\beta_4\,{}_I z_4)} = \frac{{}_{II}\beta_1\,{}_I r_{c1} + {}_{II}\beta_2\,{}_I r_{c2} + {}_{II}\beta_3\,{}_I r_{c3} + {}_{II}\beta_4\,{}_I r_{c4}}{\sqrt{\begin{array}{l} {}_{II}\beta_1^{\,2} + {}_{II}\beta_2^{\,2} + {}_{II}\beta_3^{\,2} + {}_{II}\beta_4^{\,2} \\ +2\,({}_{II}\beta_1\,{}_{II}\beta_2\,{}_I r_{12} + {}_{II}\beta_1\,{}_{II}\beta_3\,{}_I r_{13} \\ + {}_{II}\beta_1\,{}_{II}\beta_4\,{}_I r_{12} + {}_{II}\beta_2\,{}_{II}\beta_3\,{}_I r_{23} \\ + {}_{II}\beta_2\,{}_{II}\beta_4\,{}_I r_{24} + {}_{II}\beta_3\,{}_{II}\beta_4\,{}_I r_{34}) \end{array}}} \tag{1}
$$

(dof = $n_I$ - 2)

Analogous equations provide similar data for sample II intercorrelations and validity coefficients used with sample I beta weights for the 27 items in Form A of the test; sample I intercorrelations and validity coefficients used with sample II beta weights for the 27 items in Form B of the test; sample II intercorrelations and validity coefficients used with sample I beta weights for the 27 items in Form B of the test.

The results of these computations are as follows:

$$
{}_A r_{({}_I z_c)({}_{II}\beta_1\,{}_I z_1 + {}_{II}\beta_2\,{}_I z_2 + {}_{II}\beta_3\,{}_I z_3 + {}_{II}\beta_4\,{}_I z_4)} = .615 \tag{2}
$$

(dof = $n_I$ - 2)

$$
{}_A r_{({}_{II} z_c)({}_I\beta_1\,{}_{II} z_1 + {}_I\beta_2\,{}_{II} z_2 + {}_I\beta_3\,{}_{II} z_3 + {}_I\beta_4\,{}_{II} z_4)} = .459 \tag{3}
$$

(dof = $n_{II}$ - 2)

$$
{}_B r_{({}_I z_c)({}_{II}\beta_1\,{}_I z_1 + {}_{II}\beta_2\,{}_I z_2 + {}_{II}\beta_3\,{}_I z_3 + {}_{II}\beta_4\,{}_I z_4)} = .667 \tag{4}
$$

(dof = $n_I$ - 2)

$$
{}_B r_{({}_{II} z_c)({}_I\beta_1\,{}_{II} z_1 + {}_I\beta_2\,{}_{II} z_2 + {}_I\beta_3\,{}_{II} z_3 + {}_I\beta_4\,{}_{II} z_4)} = .496 \tag{5}
$$

(dof = $n_{II}$ - 2)

Before obtaining the average cross-validated coefficients for Form A and Form B, it should be determined whether the coefficients yielded by equations 2 and 3 are significantly different and whether those yielded by 4 and 5 are significantly different. The .05 level of significance was used in making this decision.

All four coefficients of interest are product-moment correlation coefficients; consequently, they may legitmately be converted to Fisher's z statistics (the hyperbolic arc tangent). The appropriate _t_ test, expressed in notation appropriate for testing the significance of the difference between the two coefficients based on nonoverlapping samples who took Form A is:

(6)
$$t\left[A^{z}(I^{z_c})(II^{\beta_1}\,I^{z_1} + \text{etc.}) - A^{z}(II^{z_c})(I^{\beta_1}\,II^{z_1} + \text{etc.})\right] = \frac{A^{z}(I^{z_c})(II^{\beta_1}\,I^{z_1} + \text{etc.}) - A^{z}(II^{z_c})(I^{\beta_1}\,II^{z_1} + \text{etc.})}{\sqrt{\dfrac{1}{n_I-3} + \dfrac{1}{n_{II}-3}}}$$

$(\text{dof} = n_I + n_{II} - 6)$

where, in this case, n equals the number of items in Form A.

Use of equation 6 indicates that the cross-validated correlations of the best-weighted combinations of the four indices and the criterion variable (the judges' ratings) for Form A were not significantly different at the .05 level. Similarly, use of the appropriate analogue of equation 6 shows that the two validity coefficients for Form B are not significantly different at the .05 level. Consequently, it is legitimate to combine the data for the two coefficients pertaining to Form A and to Form B to obtain one product-moment coefficient showing, for each form, the extent to which the best-weighted combination of standard measures corresponding to the four item indices correlate with the criterion variable.

The required equation for the within-sequences correlation coefficient, expressed in terms of Fisher's z statistic and written in notation appropriate for Form A is as follows:

(7) within-group (samples I and II in Form A)
$$\bar{z}\,({}_{A}z_c)(\beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4) = \frac{(n_I-3)(z_1) + (n_{II}-3)(z_2)}{(n_I-3) + (n_{II}-3)}$$

$(\text{dof} = n_I + n_{II} - 3 - 3)$

When equation 7 and its analogue for use with Form B are used, the weighted combination of standard measures of the four indices yield correlations with the criterion of .544 for Form A and of .592 for Form B.

Since these are product-moment correlation coefficients, each with 48 degrees of freedom, the difference between each of them and a true coefficient of zero may be tested with the usual equation:

(8)
$$t = r\left[\frac{n-2}{1-r^2}\right]$$

Use of equation (8) indicates that the correlations for both Form A and Form B are significantly different from zero at the .01 level.

The beta weights for obtaining the best-weighted combinations of the new indices for Random Halves I and II on Form A of the test computed to be:

| $I_1$ | $I_{1m}$ | $I_2$ | $I_3$ |
|---|---|---|---|
| .448 | .080 | .421 | .258 |
| .204 | .375 | .147 | -.028 |

In terms of the relative importance of the four new indices, these weights must be interpreted with caution because the size of the weight for each index is dependent upon the relationships among the particular predictors that were used in this study as well as its validity. Thus, in a strict sense, these weights indicate the relative contribution of each of the new indices only when all these and only these predictors are used.

A measure of the importance of each predictor that is independent of the relationships among a given set of predictors is the squared validity coefficient. This indicates the amount of variation in the criterion scores that each predictor independently explains. For the two halves of the examinees on Form A of the test these are:

| $I_1$ | $I_{1m}$ | $I_2$ | $I_3$ |
|---|---|---|---|
| .232 | .239 | .071 | .016 |
| .208 | .208 | .015 | .005 |

These indicate that in absolute terms both Indices $I_1$ and $I_{1m}$ are relatively good predictors of item quality and are about equally effective. The beta weights examined previously show, however, that when these two are used in the set of four new predictors, they are differentially effective because of the nature of the relationships among the predictors. These relationships are shown in Tables 1 and 2. It is also interesting to note that for the first half of the examinees, $I_2$ received a relatively large beta weight even though its squared validity coefficient is low. For the second half, $I_2$ is not particularly important by either measure. $I_3$, furthermore, does not appear to be particularly effective in terms of its squared validity coefficients.

The beta weights for obtaining the best-weighted combination of the new index for Random Halves I and II on Form B of the test are:

| $I_1$ | $I_{1m}$ | $I_2$ | $I_3$ |
|---|---|---|---|
| .086 | .362 | -.305 | .189 |
| .094 | .300 | -.127 | .284 |

These indicate the relative importance of the new predictors when all these and no additional predictors are used to obtain the best-weighted combination of the predictors.

The squared validity coefficients for these predictors with respect to Form B are:

| $I_1$ | $I_{1m}$ | $I_2$ | $I_3$ |
|---|---|---|---|
| .174 | .352 | .224 | .179 |
| .130 | .171 | .082 | .127 |

These indicate the relative importance of the predictors if each is to be used alone.

## Comparison of the Validity of the New Indices with the Conventional Index

For Form A of the test, the average of the validity coefficients obtained for the two halves of examinees for the conventional discrimination index was found to be .485. The average cross-validated multiple-correlation coefficient that indicates the validity of the weighted combination of the new indices for this form was found to be .544. The coefficients of determination indicate that, on the average, the weighted combination of the new indices explain thirty percent of the criterion variance while the conventional index explains twenty-one percent.

With respect to Form B, the average of the validity coefficients obtained for the two halves of examinees was found to be .530. The average cross-validated multiple-correlation coefficient between the weighted combination of the new indices and the criterion was found to be .592. For this form, the new indices, on the average, explain thirty-five percent of the criterion variance while the conventional index explains twenty-eight percent.

In conclusion, the validity of the weighted combination of the new indices appears to be somewhat better than the validity of the conventional discrimination index. However, Index 2 had negative validity coefficients with respect to Form B. Consequently, the nature of this index's contribution to the weighted combination of new indices is not clear from a logical or theoretical point of view. In light of this fact, it was decided not to statistically determine the significance of the differences in validities of the weighted combinations and the conventional index for each form. That is, even if the differences were shown to be statistically significant, they would not be of any practical significance without a rather thorough understanding of the nature of the superior indices. This point is discussed in greater detail in the next chapter.

## CHAPTER 5

## SUMMARY, DISCUSSION, AND CONCLUSIONS

### Summary

Despite the widespread emphasis on avoiding faults in multiple-choice items, faulty items continue to appear in standardized tests. Consequently, it seemed desirable to investigate systematically the validity of the conventional discrimination index as well as the validity of four new measures of item quality. Three of the new indices were based upon analyses of the average criterion scores of examinees who select each choice and the fourth was based upon the relative attractiveness of the distracters in items.

The first step in this study was to construct two parallel forms of a 27-item arithmetic test. In each form, nine items were designed to be free of faults, nine were designed to be moderately and nine were designed to be seriously faulty with respect to nine specific item-writing principles.

The quality of each of the items was rated independently by three experts on item-writing with the aid of a special check list of critical points to be considered in evaluating multiple-choice items. Specifically, the judges were asked to describe each fault that they found in each item and to indicate the extent to which each fault probably would affect the item's ability to discriminate between high and low ability examinees. The average of these ratings for each item served as the criterion for determining the validity of the indices.

Next, the conventional discrimination index was computed for each item. It was expressed in terms of the Davis Discrimination Index, and the criterion of examinee ability used to compute it for each item in a given form consisted of the scores on the nine items in the parallel form that were designed to be faultless.

The four new indices of item quality were computed for each item, with the scores on the nine faultless items on the parallel form as the criterion of examinee ability. Indices $I_1$ and $I_{1m}$ were designed to indicate the extent to which each item discriminates between those who know the point in question and those who do not. Index 2 was designed to indicate the extent to which the various distracters in a given item attract examinees at different levels of ability. That is, this index was designed to be a measure of the distracters' ability to discriminate among the examinees who do not thoroughly know the point in question. Index 3 was designed as a measure of the relative attractiveness of the distracters.

Thus, for each of the specially prepared items, six scores were available: the Item-Quality Rating obtained by averaging the ratings of the three judges; the Davis Discrimination Index; and the four new indices. The examinees were divided in half at random for subsequent cross validation, and the indices were computed separately on the basis of the responses of each random half of the examinees (I and II) on each form (A and B).

The first step of the analysis was to obtain the intercorrelations of the variables named above. Inspection of the correlations between each of the indices and the Item-Quality Rating indicates that the Davis Discrimination Index and Indices $I_1$ and $I_{1m}$ are moderately valid measures of item quality. Furthermore, the intercorrelations among these three indices are strong and positive.

Index 2, on the other hand appears to be operating differently from the way expected. In fact, with respect to Form B, the values of this index were negatively related to the Item-Quality Rating. Possible reasons for this result are discussed in the next section of this chapter as well as suggestions for future studies of this index.

The validity coefficients for Index 3 are in the expected direction but are disappointingly small. This finding is discussed in detail in a later section of this chapter.

The second step in the analysis was to determine the multiple-correlation coefficients of the new indices with the Item-Quality Rating for each half of the examinees on each form. The cross-validated multiple-correlation coefficients indicate that the weighted combination of the new indices is a moderately valid predictor of item quality. The usefulness of this result is severely limited by the fact that the way Index 2 operated in this study is not fully understood. That is, given the results of this study, there is no theoretical or logical basis for using this index as a measure of item quality.

In conclusion, the conventional discrimination index appears to be a moderately valid measure of item quality. A substantial amount of the variance in the judges' ratings, however, remains unexplained by the index. This suggests that further research on other indices of item quality is desirable. Furthermore, the new indices investigated in this sutdy, in general, appear to be promising measures. Further research will be needed, especially on Index 2, before recommendations can be made regarding the use of the new indices in operational settings.

## Discussion of Factors to be Considered in Future Studies of Index 2

With respect to Form A of the test, the correlation coefficients between Index 2 and the average of the judges' ratings for the two halves of the examinees were positive but weak. On Form B, the relationships between $I_2$ and the criterion were negative, and for one random half of the examinees, the negative relationship was substantial in size. These findings were disappointing since strong positive relationships were expected. Consequently, it is desirable to reexamine the assumptions used in the formulation of this index and the methods used in this study to determine its validity.

The basic assumption used in the formulation of Index 2 is that each distracter should attract examinees at a different average level of ability than the other distracters. It is interesting to note that this assumption is compatible with the widespread assumption that the right-or-wrong distinction for a given item is an arbitrary dichotomy and that the ability of examinees with respect to the point in question is in reality normally distributed. If this latter assumption is true, it should be possible to write an item to test a given point in which the distracters attract examinees at different levels along a continuum of ability. Logically, such an item should be especially effective in providing plausible alternatives for examinees who do not have adequate information to select the correct choice. In retrospect, therefore, the basic assumption underlying Index 2 still seems reasonable.

Inspection of data for individual items, however, indicates that large differences in the choice weights for the distracters may occur as a result of several types of faults in items. With respect to this possibility, consider the choice weights for the second item shown in Appendix A. Distracter C, on the average, attracted examinees at a higher level of ability than the keyed choice. Close inspection of the item indicates that there is an ambiguity in the stem that makes choice C defensible as the correct answer. Consequently, the large weight for distracter C appears to be the result of a fault in the item, and when the weight for this distracter is subtracted from the weights for the other distracters, large remainders are obtained, which increase the value of $I_2$. The undesirable influence of certain kinds of faults such as that discussed above could be controlled, to some extent, by ignoring the choice weight for any distracter that has a larger weight than the keyed choice in the computation of Index 2. The assumption underlying this provision for a modification of Index 2 is that the weight for the keyed choice in a given item should be larger than the weights for any of the distracters, which is the basic assumption for Indices 1 and 1m.

In retrospect, it seems possible that the difficulty levels of the items may have had an undue influence on the rank order of the items on Index 2. Specifically, it is unlikely that the value of $I_2$ will be large for a very easy item, regardless of its quality, since those that do not know the point in such an item probably represent a narrow range of ability. Although the items in this study were, on the average, rather difficult, there was considerable variation in the difficulty of the items, and this variation may have accounted for a substantial amount of the variation in the values of $I_2$.

Finally, weaknesses in the criterion used to determine the validity of Index 2 may have contributed to the negative results. The judges were asked to determine whether nine types of faults were present in the items and the extent to which each fault probably would effect the item's ability to discriminate between those who do and those who do not know the points in question. While this seems to be a reasonable criterion of the over all quality of test items, it does not deal directly with the characteristics of items that are likely to influence the values of $I_2$. Specifically, the directions to the judges emphasized the quality of the keyed choice and its relationship to the distracters and stem of a given item, rather than emphasizing the relationships among the distracters in terms of their relative effects on examinees. A rating scale that emphasized the latter relationships may have led to different average ratings for at least some of the items. Consider, for example, ratings that dealt with the plausibility of distracters on scales five and six in the present study. An implausible distracter in a given item was likely to lead to a low quality rating for that item. Yet, in terms of the considerations underlying the formulation of Index 2, a single implausible distracter does not necessarily reduce the quality of an item as long as the distracter is effective in attracting some of the examinees and as long as other distracters are present that are effective in attracting examinees at higher levels of ability. It is interesting to note that one implausible distracter was incorporated into each item in an early study of choice-weight scoring (Nedelsky, 1954). Such distracters were included in order to identify examinees at very low levels of ability.

In summary, despite the disappointing results regarding Index 2 in this study, the index still appears to be reasonable from a subjective point of view and probably deserves further investigation. In future studies, it is suggested that in computing Index 2, choice weights for distracters in a given item that are larger than the choice weight for the keyed choice should not be used. Furthermore, it is suggested that the items in such a study should be relatively homogeneous with respect to difficulty and be of medium or greater difficulty. Finally, the criterion of item quality should be redefined in terms of the basic assumptions underlying the index.

## Discussion of Factors to be Considered in Future Studies of Index 3

The validity coefficients for Index 3, which is a measure of the relative attractiveness of the distracters in a given item, were not as large as expected. In the present study, at least two factors may have led to the poor results. First, the values of this index may have been unduly influenced by the variation in item difficulty. That is, the fact that frequencies were used in computing this index make its values dependent, to some extent, upon the number of people who mark the item incorrectly. Specifically, it is not possible for an easy item to assume a large negative value on $I_3$, regardless of its quality, since the average number of examinees that mark distracters in the item will be low, and the deviations from this value must be small. The importance of this restriction was not recognized when this study was planned.

Furthermore, the criterion used to determine the validity of the index did not deal directly with the relative quality of the distracters in terms of their attractiveness, but rather with the quality of the keyed choice and its relationships to the distracters and the stem. Consequently, it is suggested that in future studies of the validity of Index 3, items be used that are relatively homogeneous with respect to difficulty and that a criterion be employed that deals more directly with the quality of the distracters.

## Conclusions

Several general conclusions seem appropriate as a result of this study. First, the conventional discrimination index appears to be a reasonably effective measure of item quality. In this study, much of the variation in item quality, however, remained unexplained by this index. Secondly, the new indices appear to be promising as measures of item quality. Additional research, however, is needed in order to fully understand them and to determine their value in regular test-construction projects.

# REFERENCES

Adams, G. S. *Measurement and evaluation in education, psychology, and guidance*. New York: Holt, Rinehart and Winston, 1964.

Ahmann, J. S. and Glock, M. D. *Evaluating pupil growth: Principles of tests and measurements*. (4th ed.) Boston: Allyn and Bacon, 1971.

Anastasi, A. *Psychological testing*. (3rd ed.) Toronto: Macmillan, 1968.

Board, C. and Whitney, D. R. *The effect of selected poor item-writing practices on test difficulty, reliability, and validity*. (Research report no. 55) Iowa City: University Evaluation and Examination Service, University of Iowa, 1972.

Chase, C. I. Relative length of option and response set in multiple-choice items. *Educational and Psychological Measurement*. 1964, *24*, 861 - 866.

Davis, F. B. *Item-analysis data: Their computation and use in test construction*. Cambridge: Graduate School of Education, Harvard University, 1949.

Davis, F. B. Item selection techniques. In E. F. Lindquist (Ed.) *Educational measurement*. Washington, D. C.: American Council on Education, 1951.

Davis, F. B. Estimation and use of scoring weights for each choice in multiple-choice test items. *Educational and Psychological Measurement*, 1959, *19*, 291 - 298.

Dunn, F. and Goldstein, L. G. Test difficulty, validity, and reliability as functions of selected multiple-choice item construction principles. *Educational and Psychological Measurement*, 1959, *19*. 171 - 179.

Ebel, R. L. *Measuring educational achievement*. Englewood Cliffs: Prentice-Hall, 1965.

Gulliksen, H. *Theory of mental tests*. New York: John Wiley and Sons, 1950.

Henryssen, H. Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.) *Educational Measurement*. (2nd ed.) Washington, D. C.: American Council on Education, 1971

Hoffman, B. The tyranny of testing. New York: Crowell-Collier Press, 1962.

Horst, A. P. The difficulty of a multiple-choice test item. Journal of Educational Psychology, 1933, 24, 229 - 232.

Masonis, E. J. Comparing two patterns of instruction for teaching item writing theory and skills. Unpublished doctoral dissertation, University of Pennsylvania, 1970.

Millman, J. Test-wiseness in taking objective achievement and aptitude examinations: Its nature and importance. New York: College Entrance Examination Board, 1966.

Millman, J. and Setijadi A comparison of the performance of American and Indonesian students on three types of test items. The Journal of Educational Research, 1966, 59, 273 - 275.

Nedelsky, L. Ability to avoid gross error as a measure of achievement. Educational and Psychological Measurement, 1954, 14, 459 - 472.

Slakter, M. J. et al. Grade level, sex and selected aspects of test-wiseness. Journal of Educational Measurement, 1970a, 7, 119 - 122.

Slakter, M. J. et al. Learning test-wiseness by programmed texts. Journal of Educational Measurement, 1970b, 7, 247 - 254.

Thorndike, R. L. and Hagen, E. Measurement and evaluation in psychology and education. (3rd ed.) New York: John Wiley and Sons, 1969.

Wahlstrom, M. and Boersma, F. J. The influence of test-wiseness upon achievement. Educational and Psychological Measurement, 1968, 28, 413 - 420.

Wesman, A. G. Writing the test item, In R. L. Thorndike (Ed.) Educational Measurement. (2nd ed.) Washington, D. C.: American Council on Education, 1971.

## APPENDIX A

## SAMPLE ITEMS DESIGNED TO VARY IN QUALITY AND THEIR CHOICE WEIGHTS

| Items | Weights | |
|---|---|---|
| | Sample I | Sample II |

A. "Fault-free"

1. White tile costs 11 cents per 9-inch square, while colored tile costs 13 cents for a square of the same size. How much more will it cost to cover the floor of a shower room 36 feet by 54 feet with colored instead of white tile?

| | | Sample I | Sample II |
|---|---|---|---|
| A. | $5.76 | 60 | 57 |
| B. | $11.52 | 46 | 58 |
| C. | $21.87 | 52 | 56 |
| D. | $34.56 | 50 | 54 |
| * E. | $69.12 | 75 | 80 |
| | Omit | 55 | 60 |

B. Moderately faulty (ambiguous stem)

2. Milk which sells for 20 cents a quart is on sale for 70 cents a gallon. How much money could you save if you bought 18 quarts of milk at the sale?

| | | Sample I | Sample II |
|---|---|---|---|
| A. | $1.10 | 42 | 61 |
| B. | $ .90 | 41 | 49 |
| C. | $ .45 | 63 | 65 |
| * D. | $ .40 | 60 | 64 |
| E. | $ .05 | 45 | 42 |
| | Omit | 43 | 50 |

C. Seriously faulty
(inadequate keyed choice)

3. In a certain state, $1,000 of a man's income is not taxed. All of his income over $1,000 is taxed at 26 percent, and all over $2,000 is taxed 4 percent additional. His state income tax is $500. If you let X equal the amount of his income over $2,000, which one of the following equations is true?

| Items | Weights | |
|---|---|---|
| | Sample I | Sample II |
| A. $.26 + .04(X + 1000) = 500$ | 50 | 52 |
| * B. $.26X + 1000 + .04X = 500$ | 58 | 54 |
| C. $.26X + .04(X - 1000) = 500$ | 57 | 63 |
| D. $.26(X - 1000) + .04X = 500$ | 62 | 70 |
| E. $.30(1000) + .04X = 500$ | 72 | 58 |
| Omit | 57 | 62 |

# APPENDIX B

## ITEM-QUALITY CHECK LIST

DIRECTIONS FOR JUDGES: On the following pages you will find arithmetic reasoning items intended for use with college sophomores. Below each item is a list of nine common faults in multiple-choice items.

If you think that a particular fault is present in an item, estimate how detrimental it will be to the item's ability to discriminate between those who know and those who do not know the point being tested. If you think the fault will not be detrimental, place a check beside "Not detrimental"; if you think that it will be moderately detrimental, place a check beside "Moderately detrimental"; and if you think it will be seriously detrimental, place a check beside "Seriously detrimental".

For each fault you find, specify in the space provided the part of the item that is faulty and why you think it is faulty.

An answer key for the items is enclosed on a separate sheet.

ITEM: A 21. Milk which sells for 20 cents a quart is on sale for 70 cents a gallon. How much money could you save it you bought 18 quarts of milk at the sale?

A $1.10
B $ .90
C $ .45
D $ .40
E $ .05

1. Inadequate keyed choice
______Not detrimental
______Moderately detrimental
______Seriously detrimental
Explanation: ______________________________

2. Distracters that can be defended as adequately correct due to ambiguity in expressing the meaning of the stem and choices.
______Not detrimental
______Moderately detrimental
______Seriously detrimental
Explanation: ______________________________

3. Distracters that can be defended as adequately correct even though stem and choices are unambiguous.
______Not detrimental
______Moderately detrimental
______Seriously detrimental
Explanation: ______________________________

4. Ambiguity caused by the use of a negative or double negatives.
______Not detrimental
______Moderately detrimental
______Seriously detrimental
Explanation: ______________________________

5. Implausible distracters due to a lack of homogeneity with each other and with keyed choice.
______Not detrimental
______Moderately detrimental
______Seriously detrimental
Explanation: ______________________________

ITEM: A 21. Milk which sells for 20 cents a quart is on sale for 70 cents a gallon. How much money could you save if you bought 18 quarts of milk at the sale?

A. $1.10
B. $ .90
C. $ .45
D. $ .40
E. $ .05

6. Implausible distracters, including absence of naturally attractive distracters, even though all choices are relatively homogeneous.
______Not detrimental
______Moderately detrimental
______Seriously detrimental
Explanation: ______________________________

7. Long or precisely worded keyed choice.
______Not detrimental
______Moderately detrimental
______Seriously detrimental
Explanation: ______________________________

8. Logically overlapping distracters.
______Not detrimental
______Moderately detrimental
______Seriously detrimental
Explanation: ______________________________

9. Lack of grammatical agreement of stem with choices.
______Not detrimental
______Moderately detrimental
______Seriously detrimental
Explanation: ______________________________