AN INVESTIGATION OF THE RELATIONSHIP BETWEEN THE
RELEVANCE CATEGORY OF ACHIEVEMENT TEST ITEMS
AND THEIR INDICES OF DISCRIMINATION

by

THOMAS DOUGLAS MUIR MC KIE

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS

in the

COLLEGE OF EDUCATION

We accept this thesis as conforming to the
required standard

THE UNIVERSITY OF BRITISH COLUMBIA

September, 1962
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Education

The University of British Columbia, Vancouver 8, Canada.

Date Sept. 26th 1967
ABSTRACT

It was hypothesized that on an achievement test, items measuring complex cognitive objectives would exhibit a higher mean discrimination index -- based on the whole test as criterion -- than would an equal number of items measuring less complex cognitive objectives; and that the mean discrimination index of these items would in turn be higher than that of the same number of still less complex items. The proviso was made that the difficulty indices of the items be similarly distributed within the several categories of items, hereafter called "relevance-categories," since discriminating power is related to difficulty. The categories selected were, from simplest to most complex, the Knowledge, Comprehension, and Application categories of Bloom's "Taxonomy of Educational Objectives."

An achievement test was constructed, consisting of items in all three categories, and covering the content of two units of the British Columbia university-programme grade nine science course. A try-out of this test, on 200 students in two schools, permitted negatively discriminating items to be rejected and, in addition, provided difficulty indices for the remaining items. It was possible to match forty Knowledge items and forty Comprehension items very closely for difficulty; however, the mean difficulty of the Application items was so high that they could not be used in a test of the hypothesis without reducing numbers too drastically in all categories. Two "equivalent forms," matched for content, relevance-category, and difficulty were constructed from these eighty items and administered to 530 students in three schools.
The reliability coefficient of the total test, estimated by correlating the sub-test scores and applying the Spearman-Brown formula, was .84; those of the Knowledge and Comprehension categories were similarly found to be .69 and .77, respectively. Revised difficulty indices, based on the new and larger sample, were calculated. Their distribution within the two relevance categories were found to be very similar, though not as closely matched as on the basis of the try-out test.

For each item, the point-biserial coefficient of correlation between item and total score was computed — this being the selected index of discrimination — and Fisher's z-transformation was applied to produce measures with more nearly an equal-unit scale, in the hope that the parametric t-test could be used. However, the shapes of the resulting distributions were such that they could not be claimed to be samples from a normal population or populations. Accordingly, the t-test was rejected in favour of the non-parametric Mann-Whitney test of "no difference in median discrimination indices." The respective medians were .27 and .30, in terms of Fisher's z-values, but the difference proved to be non-significant at the pre-selected 1%-level of significance.

It was concluded that this experiment provided no grounds for accepting the hypothesis of the study. However, the actual probability of obtaining, in random sampling from a single population, a difference as large as that observed was only about .10; in addition, the results consistently favoured the Comprehension items, whose discrimination indices exceeded those of the Knowledge items at the extremes as well as at the mean. It was therefore suggested that if adequate testing time could be obtained, the use of larger numbers of items in all categories
might increase test-reliability and possibly produce a significant result.

Suggestions were advanced, based upon observations from the data, for refining the experiment and for further research.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. THE PROBLEM AND HYPOTHESIS</td>
<td>1</td>
</tr>
<tr>
<td>The Problem</td>
<td>2</td>
</tr>
<tr>
<td>Statement of the problem</td>
<td>2</td>
</tr>
<tr>
<td>Statement of the hypothesis</td>
<td>3</td>
</tr>
<tr>
<td>Definition of Terms</td>
<td>3</td>
</tr>
<tr>
<td>Discussion of the Hypothesis</td>
<td>5</td>
</tr>
<tr>
<td>Importance of the Study</td>
<td>9</td>
</tr>
<tr>
<td>Organization of Remainder of the Thesis</td>
<td>11</td>
</tr>
<tr>
<td>II. REVIEW OF THE LITERATURE</td>
<td>12</td>
</tr>
<tr>
<td>Literature on Categorizing Educational Objectives</td>
<td>12</td>
</tr>
<tr>
<td>Literature Specifically Related to the Present Hypothesis</td>
<td>15</td>
</tr>
<tr>
<td>III. DELIMITATION OF THE PROBLEM</td>
<td>19</td>
</tr>
<tr>
<td>Grade-Level and Content of Test</td>
<td>19</td>
</tr>
<tr>
<td>Degree of Speededness of the Test</td>
<td>21</td>
</tr>
<tr>
<td>Type and Number of Items</td>
<td>22</td>
</tr>
<tr>
<td>Nature and Size of Sample of Students</td>
<td>24</td>
</tr>
<tr>
<td>The Tryout Problem</td>
<td>26</td>
</tr>
<tr>
<td>IV. METHODOLOGY</td>
<td>30</td>
</tr>
<tr>
<td>Preparation of the Tryout Tests</td>
<td>30</td>
</tr>
<tr>
<td>Writing and classifying the items</td>
<td>30</td>
</tr>
<tr>
<td>Assembly of tryout tests</td>
<td>32</td>
</tr>
<tr>
<td>Administration of the Tryout Tests</td>
<td>33</td>
</tr>
<tr>
<td>The problem of guessing</td>
<td>33</td>
</tr>
<tr>
<td>The test-directions</td>
<td>34</td>
</tr>
</tbody>
</table>
CHAPTER PAGE

Administration and results of the tryout tests .................. 35
The Preparation and Administration of the Final Test ........... 38
Selection of items for the final test ............................... 38
Equivalence of relevance categories in respect of
difficulty ................................................................. 41
The final test .............................................................. 44
Administration of the final test ........................................ 47
The test results ............................................................ 48

V. THE TEST RESULTS AND THEIR STATISTICAL ANALYSIS ........ 50
The Test Scores and Reliabilities ...................................... 50
Knowledge, comprehension, and total scores ...................... 50
Reliability of the test and relevance categories ................. 51
The relationship between the knowledge and
comprehension scores ................................................. 56
Distributions of Difficulty Indices .................................... 58
Distributions of Discrimination Indices .............................. 58
The Statistical Test of the Hypothesis .............................. 61

VI. CONCLUSIONS AND SUMMARY ..................................... 67
Weaknesses ................................................................. 67
Conclusions ............................................................... 68
Possibilities of Further Research ...................................... 69
Summary of the Investigation .......................................... 71

BIBLIOGRAPHY .......................................................... 74
APPENDIX A. Content-Specifications for Test (Condensed) ....... 77
APPENDIX B. Test-Directions (Excerpts) .............................. 80
APPENDIX C. Reproduction of the Final Test . . . . . . . 84
APPENDIX D. Table XI: Taxonomy Category, Difficulty Index, and Discrimination Indices (Point-Biserial and Biserial) for all Items . . . 107
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Distribution of Scholastic Aptitude Scores in Two Groups of Students Used for Tryout Tests</td>
<td>29</td>
</tr>
<tr>
<td>II. Distribution of Difficulty Indices Computed from Tryout Tests, Grouped by Relevance Category</td>
<td>37</td>
</tr>
<tr>
<td>III. Distribution of Difficulty Indices in the Knowledge and Comprehension Categories of the Final Test (Frequency and Cumulative Frequency)</td>
<td>40</td>
</tr>
<tr>
<td>IV. Values and Ranks of the Difficulty Indices in the Knowledge and Comprehension Categories</td>
<td>45</td>
</tr>
<tr>
<td>V. (a) Distribution of Unit I - Items by Content Category in the Two Final Subtests</td>
<td>46</td>
</tr>
<tr>
<td>(b) Distribution of Unit II - Items by Content Category in the Two Final Subtests</td>
<td>46</td>
</tr>
<tr>
<td>VI. Distribution of Knowledge, Comprehension, and Total Scores in Final Tests</td>
<td>51</td>
</tr>
<tr>
<td>VII. Distribution of Transformed Difficulty Indices in the Knowledge and Comprehension Categories</td>
<td>59</td>
</tr>
<tr>
<td>VIII. Distribution of Discrimination Indices of Knowledge and Comprehension Items</td>
<td>62</td>
</tr>
<tr>
<td>IX. Value and Ranks of Discrimination Indices of Items in Knowledge and Comprehension Categories</td>
<td>65</td>
</tr>
<tr>
<td>X. Content-Specifications for Test (Condensed)</td>
<td>78</td>
</tr>
<tr>
<td>XI. Taxonomy Category, Difficulty Index, and Discrimination Indices (Point-Biserial and Biserial) for all Items</td>
<td>108</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cumulative Distribution of Transformed Difficulty&lt;br&gt;Indices of Items Selected for the Final Knowledge and Comprehension Categories</td>
</tr>
<tr>
<td>2</td>
<td>Distribution of Final Scores of 530 Students on the Final Test</td>
</tr>
<tr>
<td>3(a)</td>
<td>Distribution of Knowledge-Scores of 530 Students on the Final Test</td>
</tr>
<tr>
<td>3(b)</td>
<td>Distribution of Comprehension-Scores of 530 Students on the Final Test</td>
</tr>
<tr>
<td>4</td>
<td>Cumulative Distributions of Transformed Difficulty&lt;br&gt;Indices of Items in the Knowledge and Comprehension Categories Obtained from the Administration of the Final Test</td>
</tr>
<tr>
<td>5(a)</td>
<td>Distribution of Discrimination Indices for Knowledge Items</td>
</tr>
<tr>
<td>5(b)</td>
<td>Distribution of Discrimination Indices for Comprehension Items</td>
</tr>
</tbody>
</table>
CHAPTER I

THE PROBLEM AND HYPOTHESIS

In recent years, there have been several attempts, for example, those of Bloom and his associates, Ebel, and Gerberich, to classify the objectives of cognitive learning in what have been called "relevance categories." These categories are arranged in a kind of hierarchy intended to reflect increasing complexity of behaviours. To illustrate, the simplest of these categories usually deals with simple recall of information, while the more complex ones involve behaviours such as the ability to apply principles in unfamiliar situations. Of these systems of classification, that proposed by Bloom and his co-workers, in the Taxonomy of Educational Objectives, is the most detailed; it is the one in terms of which the problem of this study is stated, and which was used throughout. For simplicity of expression, categories involving simpler behaviours will be referred to as "lower categories" and those involving more complex behaviours as "higher categories."

A major purpose of such classifications and of the specifications which accompany them is to aid the test constructor in the preparation of items which will serve to sample behaviours of the degrees of complexity.

---


4Bloom, et al., op. cit.
represented by the various categories.

I. THE PROBLEM

Statement of the problem. It was assumed that the categories of the Taxonomy do represent the different degrees of complexity of behaviour which they purport to represent; and, in general terms, it was the purpose of this investigation to discover whether, under restrictions to be specified, the magnitudes of the discrimination indices of test items sampling behaviours in different relevance categories reflect differences in complexity assumed to exist among the categories.

Reasons for supposing that such differences in discrimination indices might be found lay in the following considerations, which formed the basis of the hypothesis presented below. On a course whose objectives in the cognitive field may be classified in terms of relevance categories of increasing complexity, it is generally assumed that the best student is the one who obtains the highest score on a test. The items of the test are constructed to probe for all of the stated objectives, in so far as these are measurable. On such a test, it would appear that while the top students are likely to obtain comparatively high scores within all of the relevance categories involved, the performance of the poorer students may be expected to be less uniform. That is to say, they are likely to perform even more poorly on items sampling the less complex ones. It would then seem to follow that the characteristic which distinguishes the poor student from the good student should be not only his general inability to perform well in any of the categories, but also the fact that he will be likely to do even more poorly on higher-category items than on lower-category ones. Thus, it was anticipated that items in the more complex relevance categories would exhibit
higher indices of discrimination than would those in the less complex
categories, under restrictions specified below.

Statement of the hypothesis. Given a self-defining achievement test
with equal numbers of items sampling behaviours in the "Knowledge,"
"Comprehension," and "Application" categories of the Taxonomy, it was
hypothesized that the magnitudes of the discrimination indices of the
items, based on the whole test as criterion, would tend to increase from
the first-to the last-named category, provided that mean item-difficulty
were kept constant from category to category, and the difficulty indices
were similarly distributed within each category.

II. DEFINITION OF TERMS

Self-defining achievement test. This term was interpreted as
meaning a test of achievement in a given subject field, whose claim to
validity rests upon the care with which it was constructed to reflect the
judgment of authorities regarding the content and measurable objectives of
the course, and which thus constitutes its own criterion-measure.\(^5\)

Relevance. The definition of this term contained in the Dictionary
of Education was adopted, namely, "... the quality on the part of a test
of measuring a function that it is used to measure; the more nearly the
behaviour elicited by the test approximates the criterion behaviour, the more
relevant the test."\(^6\)

\(^5\) Adapted from Frederick B. Davis, "Item Selection Techniques,"
Educational Measurement, E.F. Lindquist, editor (Washington: Amer. Coun. on
Educ., 1951).

\(^6\) Carter V. Good (ed.), Dictionary of Education (second edition;
Relevance category. In extension of the foregoing definition of "relevance," any system of categorizing by degree of complexity the intended behaviours constituting the measurable objectives of a course was considered to result in a set of "relevance categories." In particular, the categories employed in this study were the "Knowledge," "Comprehension," and "Application" categories of the Taxonomy. Fuller discussion is provided in Chapter II, where the relevant literature is surveyed. For the present, it will suffice to note that the authors of the Taxonomy consider these categories to constitute a hierarchy of learning-outcomes, arranged from simplest to most complex in terms of the behaviours involved. 7

Discrimination index. Following Davis, this term was taken to mean a measure of "... the value of each test item for making the test in which it is included rank examinees accurately with respect to a defined criterion variable." 8 In this study, the interest was in a measure of "internal consistency discrimination," that is, one in which the total test scores are used as the criterion. Of the many indices in common use for this purpose, the point-biserial correlation coefficient, $r_{pb}$, was selected. Further discussion of its choice for use in this investigation is deferred until the hypothesis is discussed in Section III of this chapter.

Item difficulty. The definition adopted was that of Davis, namely, "... the per cent of the tryout group that marks it correctly." 9 However, as will be seen in Chapter IV, a transformation, suggested by Davis, was

7 Bloom, et al., op. cit., p.18.
9 Ibid., p.267.
applied to the difficulty indices to try to produce a scale approximating an interval scale.

III. DISCUSSION OF THE HYPOTHESIS

It will have been remarked that the statement of the hypothesis required that mean difficulty be kept constant from category to category, and that item difficulty be similarly distributed within each category. This means that within all categories, the distributions of difficulty indices must have equal means and variances, and also have the same form. The reason for including these restrictions was the familiar one of attempting to control an irrelevant variable which would otherwise exert its own influence on the dependent variable and thus lead to a confounding of influences.

The way in which differences in the mean level of difficulty from category to category could affect the size of the discrimination indices of the items in those categories may be explained in the following way. If it be assumed, for illustrative purposes, that of all the items in a test the application-items were much more difficult than the others, then the scores obtained on those items would be lower than those obtained on the lower-category items. They would thus constitute a smaller fraction of the total or criterion score. The items most nearly measuring what the test as a whole was measuring would be those most heavily represented in the criterion scores. Thus, the lower-category items and not the application-items would account for most of the variance of the total scores. In other words, as a result of their greater difficulty, the application-items might be expected to exhibit smaller discrimination indices. From this argument it follows that if item difficulty were effectively to be
prevented from introducing an extraneous influence on the mean discrimination indices of the three relevance categories, the items in all categories must have about the same distribution of difficulty indices.

In the early stages of the planning of this study, before the point made in the foregoing paragraph became evident, a different manifestation of the effect of difficulty on discrimination indices was noted. This concerned the fact, illustrated graphically by Thorndike, that the maximum size of most kinds of discrimination indices in common use rises from zero to unity and then decreases again to zero as item difficulty changes from zero through fifty to one hundred per cent. To minimize this effect, it had been decided to use biserial-\(r\) as a measure of discriminative power. This was because, as noted by Adams, biserial-\(r\) is theoretically independent of difficulty level when the criterion variable is normally distributed, and nearly so even when moderate degrees of non-normality are present.

However, once the principle had been adopted of controlling difficulty by ensuring that the three categories of items had about the same distribution of difficulty, the need for using a measure of discrimination which is independent of difficulty vanished. The question of which index of discrimination to use was therefore reconsidered.

What was required was an index which would use all of the information provided by the test scores, rather than only that given by an upper and a lower group. The two correlation coefficients, the biserial and the point-biserial, both share this property. It was also considered desirable

\(^{10}\) Robert L. Thorndike, Personnel Selection (New York: John Wiley and Sons, Inc., 1949, 229.

to select an index to which parametric statistical tests of significance could be applied, since non-parametric tests lack power; but even the basic operation of averaging cannot be applied directly to correlation coefficients. Davis notes:

It is well known that it becomes more and more difficult to raise a correlation coefficient by a certain number of hundredths as perfect correlation is approached. A difference of .05 between correlation coefficients of .90 and .95 represents a far greater difference in relationship than a difference of .05 between coefficients of .05 and .10. This is not a serious limitation for item analysis purposes, but in the case of product-moment coefficients it is so easily removed that there is no reason to tolerate it if it works any inconvenience.\[12\]

The usual procedure, for product-moment correlation coefficients, is to apply to them Fisher's z-transformation, or a linear function thereof, and to proceed thereafter on the assumption that something approaching an interval scale has been attained. Peters and Van Voorhis state that unit increments of Fisher's-z have nearly the same meaning, all along their range, in terms of difficulty of attaining them. This fact makes adding, subtracting, or averaging Fisher's-z's more legitimate processes than are like processes with correlation coefficients.\[13\]

However, biserial-r is not a product-moment coefficient. The only transformation which has been developed for use with it is that of Tate, which is applicable only when the proportion of students responding correctly to the item is near .50, and when the population-value, $\rho$, is not near plus or minus one.\[14\]

---

\[12\] Davis, op. cit., p.299.


The point-biserial coefficient, on the other hand, "is a product-moment coefficient of correlation and can be used with, and under the same conditions as, the usual product-moment coefficient." This coefficient is a legitimate and frequently used index of item-discrimination, with which the use of the Fisher transformation would seem justified. Unfortunately, its purity as an index of discriminating power is impaired by the fact that its maximum size is restricted for very easy and very hard items. As noted earlier, given a similar distribution of item-difficulty within each relevance category, this restrictive effect should apply equally to all categories. However, the possibility had to be contemplated that the final difficulty-indices might differ somewhat from those obtained on the tryout group. Under such a possibility, the "purer" biserial-$r$ would be the preferred index, despite the fact that non-parametric methods would have to be used for the test of the hypothesis. The choice of which index to use thus presented a problem, complicated by the thought that the final distributions of discrimination indices might so grossly violate the assumptions of normality and homogeneity of variance as to forbid, in any event, the use of parametric methods!

The decision finally taken was to employ the point-biserial coefficient and $z$-transformation, in the hope that control of difficulty would be adequate and that the greater efficiency of the parametric tests could be employed. Results using biserial-$r$ would be reported for comparison.

Finally, in this discussion of the hypothesis, it is to be noted that if the number of items were permitted to vary from category to category, this inequality of numbers might be expected to have a direct effect upon

---

the magnitudes of the discrimination indices. Items belonging to the relevance category with the greatest number of items in the test would be likely to have higher discrimination indices than if all categories had been equally represented. It was decided, therefore, to make the number of items in each category the same.

IV. IMPORTANCE OF THE STUDY

This study is essentially an attempt to add to knowledge within a restricted area of test-theory, namely, the behaviour of item-discrimination indices under specified conditions. Much is already known, for example, regarding the effect upon indices of item-discrimination of such variables as the difficulty level of the item, the heterogeneity of content of the test, and the distribution of the criterion variable — that is, whether it is normally distributed, leptokurtic, platykurtic or skewed. Such knowledge is of value to those who are concerned with the construction of tests, and to the professional test-builder in particular. The present study is an attempt to extend what is known in this area by clarifying the relationship which may exist between relevance category and indices of item-discrimination.

All of the practical values attached to a discovery in the area of theory do not generally become immediately apparent; but if the present hypothesis is confirmed, at least one possibility of application becomes evident. It is frequently desired to select from a group of examinees the outstanding candidates. This may be because one wishes to determine scholarship winners or members of some such special group. A test used for such a purpose must discriminate with maximum accuracy at such a level of difficulty that the desired percentage of candidates can be selected. For a test to possess such discrimination, its items should all be at the level
of difficulty corresponding to the cut-off point, as has been shown by Richardson;\(^{16}\) in addition, of course, of the available items at that level of difficulty, those with the highest discrimination indices would be chosen. If it were known in advance that higher-category items tend to have higher discrimination indices than do lower-category items, the test constructor could concentrate on the production of these more complex items, with the knowledge that they are the ones most likely to be effective for the purpose stated. He should, however, avoid overloading the test with such items, since if he did this his test would lose validity through its failure to sample adequately all of the objectives of the course.

It would seem preferable to administer a battery of tests, each designed to measure achievement at a different level of complexity. If, however, administrative time is limited -- and it usually is -- so that several sub-tests, each long enough to possess adequate reliability, cannot be given, then it is well to know how best to build a single composite test. For such a purpose it is useful to know what kind of items are likely to be most effective.

Further justification for this study arises from the following consideration: if items sampling more complex behaviours do discriminate better than those sampling less complex behaviours, they will increasingly be used by test constructors. The cause of education would benefit from such a trend, since teachers often teach only toward objectives which they know will be tested.

\(^{16}\) Davis, op. cit., p.309.
Chapter II surveys the literature pertaining to the categorization of educational objectives by degree of complexity of intended behaviours, and also provides a critical discussion of the one study which was conducted to test an hypothesis similar to that of the present study.

In Chapter III, the problem is delimited and the general plan of procedure outlined. The preparation of items, their classification into relevance categories and subsequent assembly into tryout tests are discussed in Chapter IV. This chapter also treats the administration of the tryout tests and the preparation of the final tests on the basis of information provided by the tryout tests. In particular, the statistical equivalence of the relevance categories, in terms of the difficulty indices of the items, is demonstrated.

The data obtained from the tests are presented in Chapter V. Test reliability is discussed here, together with the correlations obtained between pairs of categories. The test of the hypothesis is also described in this chapter.

The results and conclusions are summarized in Chapter VI, which also discusses weaknesses that must be considered in evaluating the conclusions.
CHAPTER II

REVIEW OF THE LITERATURE

I. LITERATURE ON CATEGORIZING EDUCATIONAL OBJECTIVES

Reference was made in the last chapter to the Taxonomy of Educational Objectives, developed by Bloom and his associates, and to its selection as the system used in this study for classifying test items into relevance categories. The Taxonomy, which was the product of the deliberations of over thirty workers in the behavioural sciences, was based on the belief that educational objectives, stated in behavioural terms, have their counterparts in the behaviour of individuals, that such behaviours can be observed and described, and that these descriptive statements can be classified.\(^1\) While educational considerations were given priority in deciding how behaviours should be categorized, logical and psychological factors were kept constantly in mind. With regard to this latter point, the authors note:

One may take the Gestalt point of view that complex behavior is more than the sum of the simpler behaviors, or one may view the complex behavior as being completely analyzable into simpler components. But either way, so long as the simpler behaviors may be viewed as components of the more complex behaviors, we can view the educational process as one of building on the simpler behavior. \(\ldots\) In order to find a single place for each type of behavior, the taxonomy must be organized from simple to complex classes of behavior.\(^2\)

---


\(^2\)Ibid., p.16.
In empirical justification of their claim that their system of classification does have this simple-to-complex order, the authors note that evidence has been found suggesting that test items sampling the more complex categories are more difficult than those sampling the simpler categories. They further note that it is more common for students to have high scores on lower-category items and low scores on higher-category items than the reverse. They conclude the discussion with the statement: "Our evidence on this is not entirely satisfactory, but there is an unmistakable trend pointing toward a hierarchy of classes of behavior which is in accordance with our present tentative classification of these behaviors."4

Mention has been made of the care taken by the authors of the Taxonomy to justify their classification upon psychological, logical, educational, and empirical grounds. Apart from this, the major argument for its selection for use in this study was its detailed description and exemplification of the behaviours classified in each category. Such detail helped to ensure maximum accuracy of classification of the test items into the relevance categories which were basic to this investigation.

It would be difficult to communicate fairly, in small compass, the essential distinctions made by the authors of the Taxonomy among the selected categories, but it was felt that some attempt should be made to provide a brief description of each. The behaviours included in the Knowledge-category involve "little more than bringing to mind the appropriate material,"5 though in a knowledge-test situation the subject must find in

3Ibid., pp.18-19.
4Ibid., p.19.
5Ibid., p.201.
the problem the clues and cues needed to stimulate recall of the relevant material. The Comprehension category is divided into three parts labelled "Translation," "Interpretation," and "Extrapolation," themselves forming a hierarchy. Given an unfamiliar communication, which may be in verbal, symbolic or pictorial form, the examinee displays translation-behaviours when he demonstrates understanding of its literal meaning by rendering it at a different level of abstraction. If he goes beyond this to draw inferences from it or to make generalizations, then he is "interpreting." If he displays sufficient understanding of trends and conditions described in the communication to enable him to make accurate predictions, he is "extrapolating." The essence of the definition supplied for application-behaviours seems to be the following:

A problem in the comprehension category requires the student to know an abstraction well enough so that he can correctly demonstrate its use when specifically asked to do so. "Application," however, requires a step beyond this. Given a problem new to the student, he will apply the appropriate abstraction without having to be prompted as to which abstraction is correct or without having to be shown how to use it in that situation.

This brief outline summarizes the definitions and distinctions which, together with many illustrative test items, formed the criteria used to classify according to relevance category the items of the test used in this investigation.

---

6 Ibid.
7 Ibid., p.91.
8 Ibid., p.90
9 Ibid.
10 Ibid., p.120.
II. LITERATURE SPECIFICALLY RELATED TO THE PRESENT HYPOTHESIS

With the implicit purpose of attempting to validate the concept of relevance category, Cook conducted an investigation. One of his hypotheses is similar to that of the present study. Specifically, he hypothesized that item-difficulty and item-discrimination indices would show an increase from the simplest to the most complex of several relevance categories. He was, however, unable to accept either of these hypotheses. Indeed, the mean discrimination indices showed a slight tendency to decrease with increasing complexity of relevance category. In view of Cook's failure to substantiate the hypothesis relating to the discrimination indices, it is proposed to discuss his methodology in some detail and to point out certain weaknesses which constitute more than adequate grounds for taking a fresh look at the problem.

Cook selected from the files of the Iowa University Examinations service a sample of tests each of which had already been item-analyzed on the basis of total test score, and whose items had previously been classified as to relevance category. The item statistics obtained were based upon Johnson's Upper-Lower method. The tests were of the objective type, and had been built as classroom tests by instructors in ten different subject fields. Cook grouped together items belonging to each of six relevance categories, using the system of classification proposed


by Ebel. He hypothesized, as already stated, that the mean discrimination
and difficulty indices for the several categories would show an
increase from the simplest to the most complex category, and failed to
establish either hypothesis using analysis of variance procedures.

Cook points to several weaknesses which might have affected his
findings. He notes: "The reliability of the item analysis data, the
intra-individual reliability of item classification, and the variation
in number of items in each category operate to limit the results." The
small number of cases, sixteen, in his simplest category, the possible
inadequacy of the instructors as test makers, and the existence of "good"
students who study details and "poor" students who don't, are advanced as
further reasons for cautious interpretation of the results. Certain
additional weaknesses are apparent, however.

First, the tests had been constructed by different instructors and
covered ten subjects ranging from Air Science through Elementary Nutrition
to Western Civilization. A finding of McConnell, later confirmed by
Furst, bears directly upon this point. Their investigations involved
correlating scores obtained by students on simple-recall items with those
obtained by them on items relating to application of principles. The
investigations were conducted both within and across several subject fields.

In the words of McConnell, "... the differences between the outcomes

13 Robert L. Ebel, "How an Examination Service Helps College Teachers
to Give Better Tests," *Proceedings of the Invitational Conference on


Objectives of Instruction," in *An Appraisal of Techniques of Evaluation:

16Edward J. Furst, "Effect of the Organization of Learning Experiences
upon the Organization of Learning Outcomes," *Journal of Experimental
within the same subject seem to be smaller than those between the subjects within the same outcome." This would imply that, for example, application-items covering a variety of subject fields should exhibit less similarity than should items sampling a variety of objectives within a single subject field. It would seem, therefore, that the items in each of Cook's relevance categories must be heterogeneous rather than homogeneous with respect to the behaviours for which he was probing.

The second of these additional weaknesses is that the tests were item-analyzed on different groups whose ability levels are unlikely to have been the same, so that Cook's indices of difficulty and discrimination can scarcely be considered comparable from test to test.

A third weakness lies in Cook's implicit assumption that his difficulty and discrimination indices are mutually independent, which is not the case. As previously stated, Cook used as his index of discrimination the U-L Index of Johnson, which is defined by the formula

\[ \text{U-L Index} = \frac{R_U - R_L}{.27N} \]

where \( R_U \) and \( R_L \) represent the number of correct responses made to an item by the upper and lower 27\% of the students, respectively, and \( N \) represents the total number of students. The effect of difficulty level upon the maximum values which can be assumed by the U-L Index may readily be demonstrated algebraically. In particular, if it is assumed that the index of difficulty of an item is represented by the symbol, \( x \), where \( x \) is less than or equal to .27, then \( xN \) of the students must have responded correctly;

---

18 Johnson, *loc. cit.*
for maximum discrimination, all of these must be in the upper 27%, so that
\[ R_U = xN \quad \text{and} \quad R_L = 0. \]

Hence,
\[
U-L \text{ Index } (\text{max}) = \frac{xN - 0}{.27N} = \frac{1}{.27} x
\]

It is therefore clear that for items with difficulty indices between 0 and .27, the maximum possible value of the discrimination index is a linear function of the difficulty index, decreasing from unity to zero as the difficulty index varies from .27 to zero, that is, as the items become more difficult.

It thus appears that it is impossible for U-L indices to continue to increase as item difficulty approaches maximum. Cook's hypotheses are thus evidently mutually incompatible; if his hypothesis of increasing difficulty with increasing complexity of relevance category were true, then his hypothesis of increasing discrimination could not be expected to hold. To test the latter hypothesis, it becomes necessary to hold difficulty level constant from category to category. It is perhaps worth repeating what was demonstrated in Chapter I on this point; namely, that it is insufficient merely to employ a discrimination index independent of difficulty.

This critical discussion of the study most pertinent to the present investigation concludes the survey of the relevant literature.
CHAPTER III

DELIMITATION OF THE PROBLEM

Reference to the hypothesis indicates that a self-defining achievement test was required, and that its items had to be classifiable, with respect to educational objectives, into one or another of the Knowledge, Comprehension, and Application categories of the Taxonomy. It was further required that the number of items should be the same from category to category and that the items should have approximately the same distribution of difficulty within each category. It was evident that no standardized, commercial test could be obtained which would meet all of these specifications, so that a test had to be constructed.

A number of decisions had to be taken prior to constructing the test, however. These concerned the grade-level and content of the test, the number and type of items to be used, and whether it was to be a power test or one which was somewhat speeded. In addition, since arrangements had to be made before the beginning of the school year with those schools in which the test was to be administered, some consideration had to be given to the size and nature of the sample of students to be used. Such preliminary decisions are discussed in this chapter.

I. GRADE-LEVEL AND CONTENT OF TEST

It was decided to restrict the study to two teaching-units in the field of science at the grade nine level. In passing, it may be noted that the objectives of science-teaching, in the cognitive field, as
prescribed for the public schools of British Columbia, include those quoted below from the relevant curriculum bulletin:

1. To acquire useful facts and information concerning the environment and to develop functional concepts and an understanding of scientific principles.

2. To acquire an appreciation of the scientific or problem-solving method and to develop the ability to use it.

3. To acquire instrumental skills such as the following:
   (a) Reading science content with understanding.

It thus appears that a valid test of achievement in this area should contain items probing for behaviours classifiable as Knowledge, Comprehension, and Application, as previously defined.

A single grade-level was chosen for a variety of reasons. Among these was the fact that it is not known whether the relationship among the behaviours represented by the various relevance categories is constant from one educational level to another; it is conceivable that they may show relatively slight relationship at one level but be more closely integrated at another. Another reason was the difficulty if not impossibility of constructing a test which would have content-validity and relevance for a variety of grades.

A single subject was chosen in view of the conclusion of McConnell and of Furst, already noted, to the effect that, for example, the ability to apply principles does not seem to be a general ability, but is rather specific to a subject field. The restriction of the study to the two units of the course dealing with the properties of matter and energy, from the points of view of chemist and physicist, narrowed the field still further.

It has been noted by Bloom that "We can only distinguish between

---

behaviors as we analyze the relation between the problem and each student's background of experience." This provides further justification for the choice of grade nine science, designated "Science 10", inasmuch as the writer's familiarity with the course and with the preceding courses in science could be expected to aid him in making valid judgments when categorizing the items. While no one can know every student's background of experience, an observation of Dressel and Nelson, who classified a very large number of items, is pertinent:

The classification of any item under a particular objective is a fallible judgment, but nevertheless one upon which a fair degree of agreement has been achieved by individuals who have independently undertaken to classify the same group of items.

To endeavour to reduce the number of possible errors in classifying items, the co-operation of two science specialists, one of whom was familiar with the Taxonomy, was enlisted, as will be described in Chapter IV.

II. DEGREE OF SPEEDEDNESS OF THE TEST

It was decided that the test should be essentially a power test since, to whatever extent a test is speeded, both difficulty and discrimination indices will be affected. Speed thus becomes an extraneous variable to be controlled. This effect of speeding is discussed by Guilford. He notes that toward the end of a test in which speed is a significant factor, the number of items omitted increases, so that the proportion of students "passing" the later items is lowered. Thus, there is an apparent increase in

---


difficulty, not necessarily because these items are inherently much more difficult, but because fewer students have had the opportunity to try them. And this situation would not be improved by defining the difficulty index as "the ratio of those passing to those attempting an item," since in this case the later items would appear easier in virtue of the fact that it is only the faster, and in general more able, students who would attempt them.

Again, if a test is timed so as to permit only the more able students to finish it, the discrimination indices of the later items are likely to be higher than if all students were permitted to finish. This is because most of the students who reach these items would do quite well while the poorer students, who fail to reach them, would be the ones counted wrong. This argument would not, of course, apply to a pure speed-test. It does, however, apply to a test in which speed and the kind of abilities being tested are positively correlated; and this is the usual situation in testing achievement in schools.

All things considered, then, the decision was taken to time the subtests so that not more than three per cent of the examinees failed to finish.

III. TYPE AND NUMBER OF ITEMS

In constructing objective test items, it is often found that a desired response can more readily be elicited by casting the item in one form rather than another, for example, in "matching" rather than "multiple-choice" form. Consequently, some tests employ a variety of item-types. In the present situation, however, it was believed that if more than one type were permitted, then the variation in the proportion of each type which would be likely to occur from one relevance category to another might introduce
an unknown factor into the problem. For a similar reason it was decided to have the same number of options for each item. Accordingly, only one type, the multiple-choice item with four options, was permitted. This choice was dictated mainly by the fact that of all item-types the multiple-choice item appears to be the one which can serve adequately in the widest variety of situations. Ebel notes that it can be used with great skill and effectiveness to measure complex abilities and fundamental understandings."^5 Furst stresses its relative freedom from response-sets. 6 Regarding the number of options, it was felt that while five would be preferable, as tending to minimize the effect of guessing, four would be a more practical aim, since in many cases it is difficult to produce additional distracters which are really functional.

The decision on the number of items to use was influenced by both theoretical and practical considerations. It is well known that, other things being equal, increasing the number of items tends to increase the reliability of a test. In addition, while the concept of "degrees of freedom" makes possible the use of small samples in statistical tests of significance, studies like those of Norton7 and Boneau8 indicate that the effect on the F and t distributions of any violations of the underlying assumptions is less for large than for small samples. Further, as a result of the fact that each item is part of the criterion with which it is to be correlated, an element of spurious correlation is introduced;

---

8 C. Alan Boneau, "The Effects of Violations of Assumptions Underlying the t Test," Psychological Bulletin, 57 (No.1, 1960), 49-64.
this will be reduced as the number of items is increased. For these three reasons, therefore, it was decided to use as many items as maintenance of item quality and other practical considerations would permit. In an effort to set a lower limit to this number, it was recalled that constructors of achievement test batteries commonly produce reliability coefficients of of around .85 to .95 for their subtests by using from about forty to eighty items, the majority containing fifty or more. Accepting these figures as a rough guide, it was felt that an attempt should be made to obtain at least fifty items for each relevance category.

However, limitations imposed by the practical situation had also to be considered. Of these, the most serious was the time available for testing, a factor which was aggravated by the fact that test items sampling the more complex categories of behaviours require more testing time per item than do simple-recall items. On the optimistic assumption that an average time of one minute per item would suffice, it was clear that a test containing one-hundred fifty items could not be administered in fewer than three fifty-minute periods. Since more testing time than this could not reasonably be hoped for, it became apparent that fifty items in each of the three categories was the maximum that might be achieved.

IV. NATURE AND SIZE OF SAMPLE OF STUDENTS

Careful scrutiny of the selected portion of the Science 10 course indicated that variations in students' experiential background resulting from such factors as geographical location would be unlikely to exert much direct influence on their performance. The portion which was tested dealt with the properties of matter, physical and chemical changes, energy-transformations and various forms of energy, liberally illustrated with laboratory demonstrations. Had a later part of the course been chosen,
for example, that dealing with elementary geology and mineralogy, or with adaptations among wild animals, it would have been considered necessary to allow for factors such as urban-rural proportions in the sample. As it was, it was considered likely that the two major factors influencing results would be (a) the ability of the students in the field of science, and (b) the quality of teaching.

With regard to the second of these, little could be done in view of the well known difficulty of measuring teacher-effectiveness; not only is there a severe criterion-problem, but there also exists an ethical difficulty which restricts attempts to locate the "best teachers" for experimental purposes. Uniformity of teaching-quality could, therefore, only be hoped for.

An adequate measure of science-ability, particularly with regard to the more complex kinds of behaviours was not available. It was felt, therefore, that a measure of scholastic aptitude was the next best criterion that could be used for this purpose. Efforts were made to secure from the provincial Division of Tests, Research and Standards, information regarding the distribution of scholastic aptitude in the grade nine university programme population. These proved unavailing. The assumption was then made that a fairly large sample of students from several schools in the southern part of the province could be considered to be a random sample from the population of students in the province.

No evidence was available on the size of sample desirable, on either theoretical or practical grounds, to ensure a high degree of stability for internal-consistency discrimination indices, though Conrad regards four to five hundred students as appropriate. Accordingly, it was arbitrarily

---

decided that a minimum of five hundred students should be used. Six schools were approached, with a total of nearly one thousand grade nine university-programme students. Of these schools, no reply was received from one, arrangements broke down with another over the timing of the test, and one small school failed to send in its results sufficiently early for them to be of use. The remaining three schools provided the sample of five hundred thirty cases on which the final results were based. These schools were the junior high schools at Trail, Kelowna and Sardis in British Columbia.

V. THE TRYOUT PROBLEM

The decisions reported to this point concern the test in its final form, that is to say, the test with equal numbers of items and equivalence of difficulty-level in all categories, and with defective items removed. To obtain indices of difficulty for all items, and to detect faulty items, it was necessary to try them out on a sample of students different from those to be used in the final administration. Such a tryout test was also required to obtain reasonably precise estimates of the time limits required for the final test. Naturally, most of the decisions about the final test applied equally to the tryout test, though the number of items in the tryout test had to be considerably greater.

At an estimated one minute per item, the tryout test of 180 items or more would require about four fifty-minute periods. Since for any one group of students and teachers, this would have involved almost all the science periods for one complete week, it was decided to try out half of the items on one group of students and the other half on a second group, both groups being reasonably similar to the group chosen to write the final test, with
respect to scholastic aptitude. While this was by no means an ideal arrangement, it appeared to be the best solution to the problem.

The most practical division of the items was in terms of content. One group of students would have to study Unit I while the other group studied Unit II. This was possible since the two units were nearly independent of each other. At least one practical advantage followed from this mode of division: the tryout tests for both units could be administered in late October and the final test could be assembled ready for administration by the time the selected schools had completed both Units I and II in December. Thus, it would be possible to do the final testing while the work of both units was fresh in the minds of the students, instead of at Easter when the students would also be loaded with additional tests covering the units studied between Christmas and Easter.

Since factors other than scholastic aptitude — such as the educational climate of a community and the general discipline of a school — affect student performance, it was believed that each of the two tryout groups should include classes from more than one school. Unfortunately, this could not be achieved; most school authorities appear reluctant to permit disruption of the status quo and to devote class-time to tryout out unproved and perhaps worthless tests. The best that could be achieved was to have the students of the Creston junior high school act as the tryout group for the Unit I items while the students of the Salmon Arm junior high school handled the Unit II items. The distributions of the scholastic aptitude scores of the grade nine university-programme students in these two schools, together with the distribution for the final group, are shown in Table I. The measures employed were Otis "Gamma-I.Q.'s". No means were available since the data reported for scholastic aptitude were in the form of
frequencies in seven letter-grade categories as shown in the table. Using the proportions in the final group as the "expected proportions" for each category, the chi-square technique described by Walker and Lev\textsuperscript{10} was employed to test the hypothesis that the discrepancies between the tryout and the final proportions were such as might have occurred rather frequently in random sampling. The obtained value of $\chi^2$ for the Creston group was 5.326, and for the Salmon Arm group was 6.474. Since, for six degrees of freedom, the probability of obtaining values greater than these exceeds .50 and .30, respectively, the hypothesis was accepted.

As noted in the foregoing section, this arrangement of using two tryout groups in two different schools was far from ideal, and was adopted only out of practical necessity. Since the final test would consist, within each relevance category, of about equal numbers of items on each content-unit, it was hoped that any discrepancies in the difficulty indices resulting from the use of different groups of students might manifest themselves more or less equally within each category.

This discussion of the preliminary decisions and arrangements being concluded, it is now possible to proceed to a description of the methodology of the investigation.

TABLE I

DISTRIBUTION OF SCHOLASTIC APTITUDE SCORES
IN TWO GROUPS OF STUDENTS
USED FOR TRYOUT TESTS

<table>
<thead>
<tr>
<th>Scholastic aptitude (letter-grade*)</th>
<th>I.Q.-range</th>
<th>Expected frequencies (from final group)</th>
<th>Creston (Unit I)</th>
<th>Salmon Arm (Unit II)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>f %</td>
<td>f %</td>
<td>f %</td>
</tr>
<tr>
<td>A</td>
<td>126 -</td>
<td>49 9.0</td>
<td>15 12.2</td>
<td>10 8.9</td>
</tr>
<tr>
<td>B</td>
<td>116 - 125</td>
<td>131 24.2</td>
<td>27 22.3</td>
<td>20 17.9</td>
</tr>
<tr>
<td>C+</td>
<td>110 - 115</td>
<td>124 22.8</td>
<td>22 18.2</td>
<td>30 26.8</td>
</tr>
<tr>
<td>C</td>
<td>104 - 109</td>
<td>133 24.4</td>
<td>27 22.3</td>
<td>31 27.7</td>
</tr>
<tr>
<td>C-</td>
<td>98 - 103</td>
<td>78 14.3</td>
<td>23 19.0</td>
<td>13 11.6</td>
</tr>
<tr>
<td>D</td>
<td>85 - 97</td>
<td>26 4.7</td>
<td>7 5.8</td>
<td>6 5.3</td>
</tr>
<tr>
<td>E</td>
<td>84 -</td>
<td>3 0.6</td>
<td></td>
<td>2 1.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>544 100.0</td>
<td>121 100.0</td>
<td>112 100.0</td>
</tr>
</tbody>
</table>

1The Department of Education of British Columbia divides the grade nine population into seven groups designated by letters, "A" representing the most able and "E" the least able group. In the entire grade nine population the proportions represented by the letters are: A: 5%, B: 20% C+: 15%, C: 20%, C-: 15%, D: 20%, E: 5%.
CHAPTER IV

METHODOLOGY

The preparation, administration, and results of the tryout test are described in this chapter, which also treats the selection of the items for the final test. The statistical equivalence of the relevance categories of the final test, in terms of their difficulty level, is also demonstrated. The chapter concludes with a discussion of the composition and administration of the final test, consideration of results being deferred until Chapter V.

I. PREPARATION OF THE TRYOUT TESTS

Writing and classifying the items. Guidance was sought from the appropriate curriculum bulletin\(^1\) regarding sections of the course considered particularly important and topics to be omitted or treated in outline only. A detailed table of specifications was then prepared to cover the content of the chosen units. This table was used constantly as a guide during the preparation of the items, and copies were mailed to the teachers of the participating classes to try to ensure uniformity of content-coverage. Valuable as it was for these purposes, this table was far too detailed for use in classifying items by content. This was a necessary preliminary to assembling the final forms of the test, in which items were matched on content as well as on relevance-category and difficulty.

---

\(^1\)British Columbia Department of Education, The Sciences (Victoria, B.C.: The Queen's Printer, 1956), 39-42.
Accordingly, this table was condensed and, in its abbreviated form, is reproduced in Appendix A.

A total of 192 items was produced, each item being typed on a 5" x 8" card and classified by content according to the condensed table of specifications. As the items were written they were also classified by relevance category, constant reference being made to the Taxonomy in an effort to secure maximum accuracy of classification. In addition, the hundreds of items classified according to the Taxonomy by Dressel and Nelson were consulted frequently.² There were, altogether, 60 knowledge-items, 67 comprehension-items, and 65 application-items.

The aid of two colleagues was enlisted to check the accuracy of the classification and to provide estimates of difficulty. Both were science-graduates with experience in teaching the Science 10 course and previous courses. In addition, one had taken graduate work in educational measurement and was familiar with the Taxonomy. Both consulted the Taxonomy to resolve doubts that arose. The criterion for acceptance of an item was set at 100 per cent agreement among the classifiers. Twelve items were rejected since they failed to satisfy the 100 per cent criterion, another eleven as being too abstruse or otherwise difficult, and six as possessing defects unperceived by the writer. The remaining 163 items comprised 49, 58, and 56 in the knowledge, comprehension, and application categories, respectively.

This number was less than had been hoped for. It left little room for the rejection of further items during the process of equating the

relevance categories for difficulty. However, this discovery was made too close to the time scheduled for the tryout tests to make it possible for additional items to be constructed and appraised.

**Assembly of tryout tests.** Six subtests were prepared, three for each unit. The number of items in Unit I was 82. These were divided on the basis of content, into subtests containing 31, 31 and 20 items, respectively. For Unit II, the total number was 81, similarly divided. Each subtest contained items in all three relevance categories, and within each subtest the items were arranged in order of increasing difficulty, as estimated by the three teachers involved in the checking of the items.

The arrangement of the items in order of difficulty is generally accepted in testing-practice. The purpose of mixing items of different relevance categories was two-fold. First, it was intended to eliminate what Lindquist calls "Type G errors," that is, errors which would systematically affect one relevance category rather than another, with respect to difficulty indices. Such errors could result from changes taking place within students in the interval between subtests, and should be allowed to manifest themselves equally in all relevance categories. And secondly, the items were mixed to avoid the possibility of giving the students all at once a large number of items whose difficulty level might prove to be fairly high if Bloom's contention \(^1\) regarding increasing difficulty with increasing complexity of behaviours should turn out to be true — as indeed it did!

---


The first two subtests were timed to take fifty minutes each, and the third to take thirty minutes, these times being sufficient for almost everyone to consider every item.

Care was taken to ensure that the correct responses were divided about equally among the alternatives a, b, c, and d. In addition, the order in which the correct responses appeared was without pattern so as to provide no extraneous clues for the students.

II. ADMINISTRATION OF THE TRYOUT TESTS

The problem of guessing. It was clearly desirable to curb blind guessing as far as possible since it would distort the distribution of criterion scores and also affect the item-statistics; but it was felt that it would be unwise for a number of reasons to apply a "correction for guessing" to the scores in this investigation. First, the derivation of the usual correction-formula assumes that all wrong answers are derived from pure guessing, so that when the correct answer is not known all possible responses to an item are equally likely to be chosen. It is, in fact, extremely improbable that all incorrect responses are the result of pure guessing. Many wrong responses are likely to result from partial information or misinformation. In the former case, the correction-formula would undercorrect, since the subject is actually guessing among fewer alternatives and thus increasing the possibility of chance-success. In the latter case, the formula would overcorrect since the subject is not guessing at all; he is marking what he genuinely believes to be the correct response. Second, the occurrence of pure or blind guessing is likely to be greatest in speeded tests, where subjects have insufficient time to finish, and in the last few seconds blindly mark the remaining questions; but the test considered here
was constructed as a power test, so that this situation would not be likely to occur. Further, a strongly worded penalty for guessing would be likely to influence the cautious, conscientious, sensitive student, while the one with the "gamblers" personality would be relatively unaffected. Lastly, and not least in importance, it was believed that correction for guessing might introduce a factor which would interact in a totally unknown manner with the relevance categories.

Thus, despite the existence of arguments in the opposite direction, the decision was taken to apply no correction for guessing, but to phrase the directions in such a way as to discourage blind guessing. The instruction finally selected was: "You may answer questions even when you are not completely sure that your answers are correct. In such cases, intelligent consideration of the choices provided may help you to gain marks. HOWEVER, you should AVOID WILD GUESSING as this may result in a reduction of your score." It was hoped that the suggestion regarding the intelligent weighing of alternatives might encourage the cautious, and at the same time appeal to the "gamblers" as a reasonable alternative to blind guessing.

The test-directions. The directions for the final test are reproduced in Appendix B. Since the directions for the tryout tests were essentially the same it is not proposed to detail them here. However, brief reference may be made to their content. Besides the instructions concerning guessing, they included certain preliminary directions. It was impossible to schedule all students in a school to write a given sub-test during the same period. Consequently it was possible that students who had written a sub-test might communicate information to those who had not. Directions were included to minimize this. The preliminary directions also ensured that students were given advance preparation on sample items of the types they were likely to
meet. Further, they dealt with the physical arrangements which had to be
made in advance of the tests. In addition, they included the instructions,
previously reported, regarding time-limits, and the necessity for strict
adherence to these, together with instructions to the students to draw a
line under the highest-numbered item which they had time to consider, whether
they could answer it or not. The purpose of this latter instruction was, of
course, to determine the extent to which the tryout tests were speeded, and
hence to estimate the reliability of the difficulty indices of the later
items.

**Administration and results of the tryout tests.** The three subtests
dealing with Unit I were written by 107 students taught by four teachers,
while those dealing with Unit II were written by 100 students under three
teachers. No student indicated that he was unable to consider every item.

For each group, the papers were arranged in order of total score,
and divided into three groups consisting of the top 27%, middle 46%, and
bottom 27%. The performance of every student on every item was recorded.
A note was also made of omissions, but only for the upper and lower groups.

Four of the Unit I-items and two of the Unit II-items were answered
correctly by more students in the lower than in the upper portion of the
group writing them. These items were rejected as being unsuitable for
further consideration, on the ground that -- perhaps as a result of some
unperceived ambiguity of wording or other defect -- they were getting at
something essentially different from what the total test was measuring.
In addition, the presence of such items would have increased heterogeneity
within the test, and have resulted in a reduction of the reliability of
the criterion scores.

Difficulty indices were computed for all other items; these
indicated the per cent of the tryout group responding correctly to the item, according to the definition adopted in Chapter I. Davis has pointed out that the use of per cents as difficulty indices is open to the criticism that per cents "do not even approach an interval scale of difficulty." He suggests, on the assumption of approximate normality in the trait being measured, a procedure to transform the percentage-indices into numerical indices which can reasonably legitimately be averaged and otherwise statistically manipulated. His suggested method is to use the table of the normal probability integral to transform the per cents to standard-deviation units, to multiply these figures by 21.063, and to add 50. This amounts to setting the mean at 50 and the standard deviation at 21.063, and results in a set of indices for which the values 1, 50 and 99 correspond to the 1%, 50%, and 99% levels respectively. This procedure was adopted. In the following discussion, the term "difficulty index" is to be regarded as referring to the transformed indices, unless otherwise specified.

The distribution of these indices within each category is shown in Table II below. Their means, for the knowledge, comprehension, and application categories were 49.5, 44.1 and 39.1, respectively, thus displaying the trend anticipated by the authors of the Taxonomy toward increasing difficulty of items with increasing complexity of behaviours sampled. Indeed, as will be seen, this trend was sufficiently pronounced to force the elimination of the application category from further consideration, and to reduce the test of the hypothesis to a comparison of the knowledge and comprehension categories only.

6Ibid.
**TABLE II**

DISTRIBUTION OF DIFFICULTY INDICES COMPUTED FROM TRYOUT TESTS GROUPED BY RELEVANCE CATEGORY

<table>
<thead>
<tr>
<th>Transformed difficulty indices (real intervals)</th>
<th>Original difficulty indices (per cents)</th>
<th>Number of items</th>
<th>Knowledge</th>
<th>Comprehension</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>82.25 - 85.75</td>
<td>95 - 96</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>78.75 - 82.25</td>
<td>92 - 94</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75.25 - 78.75</td>
<td>89 - 91</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>71.75 - 75.25</td>
<td>85 - 88</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>68.25 - 71.75</td>
<td>81 - 84</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64.75 - 68.25</td>
<td>76 - 80</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>61.25 - 64.75</td>
<td>71 - 75</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>57.75 - 61.25</td>
<td>65 - 70</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>54.25 - 57.75</td>
<td>59 - 64</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>50.75 - 54.25</td>
<td>52 - 58</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>47.25 - 50.75</td>
<td>45 - 51</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>43.75 - 47.25</td>
<td>39 - 44</td>
<td>8</td>
<td>7</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>40.25 - 43.75</td>
<td>33 - 38</td>
<td>5</td>
<td>7</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>36.75 - 40.25</td>
<td>27 - 32</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>33.25 - 36.75</td>
<td>22 - 26</td>
<td>3</td>
<td>5</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>29.75 - 33.25</td>
<td>17 - 21</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>26.25 - 29.75</td>
<td>13 - 16</td>
<td>1</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>22.75 - 26.25</td>
<td>10 - 12</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>19.25 - 22.75</td>
<td>8 - 9</td>
<td>2</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>15.75 - 19.25</td>
<td>6 - 7</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12.25 - 15.75</td>
<td>4 - 5</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>8.75 - 12.25</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.25 - 8.75</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

\[ N = 48 \quad 56 \quad 53 \]

Mean difficulty index = 49.5 \quad 44.1 \quad 39.1

---

1Six items, rejected as discriminating negatively, have been omitted from this table.

2The transformed difficulty indices were obtained from the original per cents by using Davis' procedure, described on page 36.

3The dotted lines indicate the boundaries of the range of difficulty indices from which items were selected for the final test.
Before passing on to a discussion of the preparation of the final tests, it may be of interest to note that of the students who wrote the Unit I test, twenty-nine, or 50%, of the students in the upper and lower groups omitted one or more items, as did thirty, or 55%, of those who wrote the Unit II test. It would appear that these students, at any rate, believed that they were not guessing blindly. However, the fact that the remaining 45 to 50% of the upper and lower groups responded to all items suggests that a certain amount of guessing may have been taking place and that the results must be assessed with this in mind.

III. THE PREPARATION AND ADMINISTRATION OF THE FINAL TEST

Selection of items for the final test. As indicated in the discussion of the hypothesis in Chapter I, it was necessary that the difficulty indices of the items within each relevance category of the final test should have as nearly as possible the same distribution. The distribution aimed at was a symmetrical one with a mean of about 50. A mean much lower than this was considered to be undesirable since it might lead to increased guessing and consequent distortion of both criterion scores and discrimination indices.

As may be seen from Table II, the four easiest knowledge-items had to be excluded since there were no items of a corresponding level of difficulty in either of the other categories. At the other end of the scale, the knowledge items having a difficulty index of less than 31.5 were rejected, leading forty items with a mean of 47.8. To obtain a mean of 50.0, it would have been necessary to reject six additional items, reducing the total number in that category -- and hence in all
categories -- to thirty-four, which was considered too low to offer hope of reasonably reliable results.

At this point, it became clear that the application-items could not be retained, for the mean difficulty of the top forty of these items was only 43.8; to raise the mean to that of the forty knowledge-items would have meant using only the top twenty-eight items, which was again too low. It was with reluctance that this decision was taken. It had, throughout, been foreseen as a possibility but, as was pointed out earlier in this chapter, the construction of more items to provide a greater choice would probably have been futile, owing to the impossibility of securing significantly more testing time.

The next problem was to select from the available comprehension-items a group of forty with a mean difficulty index of about 47.8 and roughly the same distribution of these indices as in the knowledge category. First, the four items whose difficulty indices fell outside the range, covered by those of the knowledge-items were rejected. Then, twelve more items were discarded. The final distributions are shown in Table III in which smaller class-intervals have been used than were employed in Table II, in order to show more detail.

A consideration in determining which items to discard was that since only the knowledge- and comprehension-items remained, the most time-consuming items having been rejected, it should be possible to assemble the final test in the form of two subtests. Since these would have to be administered in two separate periods, it was considered that from the point of view of determining test-reliability it would be better to try to assemble the tests as "equivalent forms" in which difficulty level, relevance category, and content were as nearly alike in both forms as possible. In this way, the component of variance in the final test
### TABLE III

DISTRIBUTION OF DIFFICULTY INDICES IN THE KNOWLEDGE AND COMPREHENSION CATEGORIES OF THE FINAL TEST
(FREQUENCY AND CUMULATIVE FREQUENCY)

<table>
<thead>
<tr>
<th>Difficulty indices (real intervals)</th>
<th>No. of items (frequency)</th>
<th>Cumulative frequency (^1)</th>
<th>Max. difference (^1) in cumulative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K</td>
<td>C</td>
<td>K</td>
</tr>
<tr>
<td>63.99 - 65.99</td>
<td>2</td>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>61.99 - 63.99</td>
<td>2</td>
<td>3</td>
<td>38</td>
</tr>
<tr>
<td>59.99 - 61.99</td>
<td>1</td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td>57.99 - 59.99</td>
<td>2</td>
<td>1</td>
<td>35</td>
</tr>
<tr>
<td>55.99 - 57.99</td>
<td>1</td>
<td>1</td>
<td>33</td>
</tr>
<tr>
<td>53.99 - 55.99</td>
<td>3</td>
<td>3</td>
<td>33</td>
</tr>
<tr>
<td>51.99 - 53.99</td>
<td>6</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>49.99 - 51.99</td>
<td>2</td>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>47.99 - 49.99</td>
<td>2</td>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td>45.99 - 47.99</td>
<td>2</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>43.99 - 45.99</td>
<td>7</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>41.99 - 43.99</td>
<td>3</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>39.99 - 41.99</td>
<td>2</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>37.99 - 39.99</td>
<td>2</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>35.99 - 37.99</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>33.99 - 35.99</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>31.99 - 33.99</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>29.99 - 31.99</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

\(^1\) These figures are provided for reference in connection with a statistical test used on page 43 of this chapter.
scores attributable to testing on two different occasions would be properly treated as error variance rather than as systematic variance.

In view of these considerations, the problem of discarding the excess comprehension-items became one of card-shuffling with two aims in view: (1) to produce similar distributions of difficulty for both relevance categories, together with near-equivalence of content, and (2) to produce, from the same items, two subtests characterized by near-equivalence in terms of difficulty, relevance category and content.

**Equivalence of relevance categories in respect of difficulty.** It is evident from Table III that the final distributions of difficulty indices for items in the knowledge and comprehension categories do not approach the normal form. Despite this, the distributions appeared very similar, as may be seen from Figure 1, below, which depicts the distributions as ogives. For the knowledge- and comprehension-items, respectively, the means were 1.7.8 and 1.7.9, the medians 1.6.0 and 1.6.5, and the standard deviations 9.03 and 9.55. Though the close correspondence of these statistics and of the ogives indicated that the two distributions could reasonably be said to satisfy the requirements of the hypothesis, it was thought desirable to supplement such a subjective claim by statistical tests. The non-normal form of the distributions rendered parametric tests of doubtful validity, so the Kolmogorov-Smirnov non-parametric test was applied. This test, which Walker and Lev describe as "sensitive to any kind of difference in the distributions from which the samples are drawn," tests the null hypothesis that both samples were drawn from populations with the same continuous distribution. With reference to the assumption

---

FIGURE 1
CUMULATIVE DISTRIBUTION OF TRANSFORMED DIFFICULTY INDICES OF ITEMS SELECTED FOR THE FINAL KNOWLEDGE AND COMPREHENSION CATEGORIES

KEY: __________________________ Knowledge items

______________________ Comprehension items
of continuity, it is not unreasonable to believe that item-difficulty is continuously distributed, even though the numerical indices used to characterize it may exhibit discreteness. In any event, the parametric tests of significance in common use are frequently applied to discrete or only quasi-continuous data, though the theoretical requirements call for random samples from a normal, and therefore continuous, population.

For equal samples, each with forty cases or fewer, the Kolmogorov-Smirnov test involves referring the statistic $D_N$ to a specially devised significance table. Since $D$ represents the maximum distance between the cumulative percentage distributions, and $N$ is the number of cases in each of the two samples, $D_N$ represents the maximum difference, in any interval, between the actual cumulative frequencies. The last column of Table III shows this difference to be three. Since the size of this statistic required to reject the null hypothesis at the .05-level for sample-size forty is twelve, the null hypothesis must be accepted, and the conclusion drawn that it is highly reasonable to accept the view that the two samples have essentially the same distribution.

In corroboration, the Mann-Whitney test was employed to test the null hypothesis of "no difference in medians." This involved assigning ranks to the eighty difficulty indices, finding the sum of the ranks of the indices of one sample, and referring to the table of the normal probability integral the statistic

$$z = \frac{2R_K - N(N+1)}{\sqrt{NKN_C(N+1)}}$$

\[Ibid., p.427.\]  \[Ibid., pp.434-35.\]
where \( R_K = \text{sum of ranks of difficulty indices of knowledge-items.} \)
\( N_K = \text{number of cases in the knowledge category.} \)
\( N_C = \text{number of cases in the comprehension category.} \)
\( N = N_K + N_C \)

The difficulty indices, together with their ranks, are supplied in Table IV, below, for reference. Where ties occurred between indices, the rank assigned to each was the mean of the ranks which would have been assigned to them had they been slightly different.

Here, the obtained value of \( z \) was .072. Since the probability of obtaining an absolute value of \( z \) greater than .072 is about .94, this figure being obtained from the table of the normal probability integral, for a two-tailed test, it was concluded that there was no basis for believing that the two samples differed significantly in their medians.

As a result of these two tests, it was felt safe to assume that the knowledge- and comprehension-items had been adequately matched on the variable, difficulty, as required by the hypothesis.

The final test. The decision to assemble the final test in the form of two subtests closely matched in terms of difficulty level, content, and relevance category, has already been reported. The result of the matching process was two subtests, I and II, with mean difficulty indices of 47.9 and 47.8, respectively, and each containing equal numbers of knowledge- and comprehension-items. Table V indicates that the subtests were also quite closely matched on the basis of the content-groupings shown in the table of specifications in Appendix A. Care was taken, however, to ensure that items were not matched to the point of near-identity of content for, as Thorndike points out, this could result in a spurious inflation of the
### Table IV

Values and Ranks of the Difficulty Indices in the Knowledge and Comprehension Categories

<table>
<thead>
<tr>
<th>Difficulty indices</th>
<th>Ranks of indices</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>C</td>
<td>K</td>
<td>C</td>
</tr>
<tr>
<td>65.5</td>
<td>61.9</td>
<td>1</td>
<td>2.5</td>
</tr>
<tr>
<td>61.1</td>
<td>61.9</td>
<td>4</td>
<td>2.5</td>
</tr>
<tr>
<td>62.8</td>
<td>62.8</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>62.2</td>
<td>62.8</td>
<td>8.5</td>
<td>6</td>
</tr>
<tr>
<td>59.9</td>
<td>62.2</td>
<td>11</td>
<td>8.5</td>
</tr>
<tr>
<td>58.6</td>
<td>60.9</td>
<td>12.5</td>
<td>10</td>
</tr>
<tr>
<td>58.2</td>
<td>58.6</td>
<td>14</td>
<td>12.5</td>
</tr>
<tr>
<td>58.1</td>
<td>56.9</td>
<td>19</td>
<td>15</td>
</tr>
<tr>
<td>54.2</td>
<td>55.9</td>
<td>20.5</td>
<td>16</td>
</tr>
<tr>
<td>54.2</td>
<td>55.9</td>
<td>20.5</td>
<td>17.5</td>
</tr>
<tr>
<td>53.8</td>
<td>55.3</td>
<td>22.5</td>
<td>17.5</td>
</tr>
<tr>
<td>53.1</td>
<td>53.8</td>
<td>24.5</td>
<td>22.5</td>
</tr>
<tr>
<td>52.7</td>
<td>53.1</td>
<td>26.5</td>
<td>24.5</td>
</tr>
<tr>
<td>52.7</td>
<td>52.1</td>
<td>26.5</td>
<td>30</td>
</tr>
<tr>
<td>52.1</td>
<td>52.1</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>52.1</td>
<td>52.1</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>49.4</td>
<td>51.1</td>
<td>34</td>
<td>33</td>
</tr>
<tr>
<td>48.3</td>
<td>48.3</td>
<td>35.5</td>
<td>35.5</td>
</tr>
<tr>
<td>47.3</td>
<td>47.3</td>
<td>37.5</td>
<td>37.5</td>
</tr>
<tr>
<td>46.2</td>
<td>46.9</td>
<td>40.5</td>
<td>39</td>
</tr>
<tr>
<td>45.8</td>
<td>46.2</td>
<td>44</td>
<td>40.5</td>
</tr>
<tr>
<td>45.8</td>
<td>45.8</td>
<td>44</td>
<td>44</td>
</tr>
<tr>
<td>45.2</td>
<td>45.8</td>
<td>47</td>
<td>44</td>
</tr>
<tr>
<td>44.7</td>
<td>45.8</td>
<td>48.5</td>
<td>44</td>
</tr>
<tr>
<td>44.7</td>
<td>44.1</td>
<td>48.5</td>
<td>51.5</td>
</tr>
<tr>
<td>44.1</td>
<td>44.1</td>
<td>51.5</td>
<td>51.5</td>
</tr>
<tr>
<td>43.1</td>
<td>43.1</td>
<td>51.5</td>
<td>57</td>
</tr>
<tr>
<td>43.5</td>
<td>43.1</td>
<td>54.5</td>
<td>57</td>
</tr>
<tr>
<td>43.5</td>
<td>40.8</td>
<td>54.5</td>
<td>61</td>
</tr>
<tr>
<td>43.1</td>
<td>39.5</td>
<td>57</td>
<td>62.5</td>
</tr>
<tr>
<td>41.8</td>
<td>39.1</td>
<td>59</td>
<td>65</td>
</tr>
<tr>
<td>41.4</td>
<td>39.1</td>
<td>60</td>
<td>65</td>
</tr>
<tr>
<td>39.5</td>
<td>37.8</td>
<td>62.5</td>
<td>67</td>
</tr>
<tr>
<td>39.1</td>
<td>37.2</td>
<td>65</td>
<td>68</td>
</tr>
<tr>
<td>35.9</td>
<td>36.5</td>
<td>70.5</td>
<td>69</td>
</tr>
<tr>
<td>35.9</td>
<td>35.1</td>
<td>70.5</td>
<td>72.5</td>
</tr>
<tr>
<td>34.4</td>
<td>35.1</td>
<td>74.5</td>
<td>72.5</td>
</tr>
<tr>
<td>33.0</td>
<td>34.4</td>
<td>77.5</td>
<td>74.5</td>
</tr>
<tr>
<td>33.0</td>
<td>33.0</td>
<td>77.5</td>
<td>77.5</td>
</tr>
<tr>
<td>31.5</td>
<td>33.0</td>
<td>80</td>
<td>77.5</td>
</tr>
</tbody>
</table>

Sum of ranks: 1622.5  
N = 40
### TABLE V (a)

**DISTRIBUTION OF UNIT I-ITEMS BY CONTENT CATEGORY IN THE TWO FINAL SUBTESTS**

<table>
<thead>
<tr>
<th>Unit I content-category</th>
<th>Number of items in Subtest I</th>
<th>Subtest II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td><strong>21</strong></td>
<td><strong>22</strong></td>
</tr>
</tbody>
</table>

**No. of Unit I items in final test** 43

### TABLE V (b)

**DISTRIBUTION OF UNIT II-ITEMS BY CONTENT CATEGORY IN THE TWO FINAL SUBTESTS**

<table>
<thead>
<tr>
<th>Unit II content-category</th>
<th>Number of items in Subtest I</th>
<th>Subtest II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td><strong>19</strong></td>
<td><strong>18</strong></td>
</tr>
</tbody>
</table>

**No. of Unit II items in final test** 37
reliability coefficient of the test.  

As in the tryout test, the items in each final subtest were arranged in order of increasing difficulty; also, the correct responses were about equally distributed among the four options for the items, and were arranged in a quasi-random fashion.

The final test if reproduced in Appendix C.

Administration of the final test. It was estimated from the tryout tests that, with the time consuming application-items deleted, the eighty items of the final test could be administered in two fifty-minute periods, without introducing a speed-factor, provided that the students were familiarized with the directions before the test began.

This decision was incorporated into the directions, which are reproduced in Appendix B. The teachers administering the test were requested to read the directions without embellishing them in any way. They were asked to give prior practice to their students in handling sample items chosen from among the rejects. In reading the directions they conveyed the information that on tests like this it is expected that the average student will score around fifty per cent, and that the pass-mark on this test would be considerably lower than that. The purpose of this was to help in reducing tension and -- it was hoped -- to decrease the ever-present tendency among some students to guess wildly, regarding which the instructions used in the tryout test were repeated. It should be remarked that the students were told that the test-results would be used to help determine their Christmas mark in science. This had the

merit of forcing them to take the test seriously and of providing a kind of motivation. It could also have had the effect of increasing the tendency toward guessing, but this was in any event inevitable since the tests would never have been administered at all if the teachers had been denied the opportunity to use the results and the norms.

Strict adherence to time-limits was stressed. In addition, students were instructed to draw a line under the highest-numbered item which they had a chance to consider, and -- if they had been able to consider all items -- to place a check-mark in a space provided on the cover-sheet. Thus, it was possible to confirm the expectation that the test was a power-test. For convenience, it may be reported at this point that only one student indicated that he had been unable to consider all of the items of one subtest. He had marked the last few items, but had omitted some in the body of the subtest, so that it appeared that he had deliberately left for later consideration some items which had, on cursory reading, appeared difficult. Essentially, then, the final test was a power-test.

The teachers were requested to make internal arrangements so that subtest I could be administered during the same period for all classes, or at least during the same morning or afternoon; the same request was made for subtest II. These requests were complied with so that leakage during change of periods could be considered minimal. The two subtests were run on successive days.

The test results. A few students were absent for one or the other subtest, but 530 students wrote both. Each student's response to every item was recorded. The results were sent to the Computing Centre at the University of British Columbia for the computation of the point-biserial
correlation coefficients which represented the discrimination indices of all items.

The test data and their interpretation are deferred to Chapter V.
CHAPTER V

THE TEST RESULTS AND THEIR STATISTICAL ANALYSIS

In this chapter the distributions of scores in each relevance category, and of total scores, are presented. Test reliability is discussed together with inter-category correlation. The distributions of final difficulty indices are depicted graphically, and the adequacy of the process of matching for difficulty is tested. Finally, after applying Fisher's z-transformation to the point-biserial correlation coefficients, or discrimination indices, a test of significance is made. This is supported by corroborative evidence using biserial- r.

I. THE TEST SCORES AND RELIABILITIES

Knowledge, comprehension, and total scores. Table VI shows the distributions of knowledge scores, comprehension scores, and total scores. The same information is presented visually in Figures 2 and 3 which follow. All three distributions are somewhat positively skewed, this effect being least for the knowledge items. All exhibit a certain "piling up" of scores near the lower end, an effect which is particularly noticeable with the comprehension items. While the means and medians of the knowledge and comprehension scores, shown in the table, are highly similar, their difference in variance is noteworthy. Before discussing this point further, data on reliability will be considered.
TABLE VI
DISTRIBUTION OF KNOWLEDGE, COMPREHENSION, AND TOTAL SCORES IN FINAL TESTS

<table>
<thead>
<tr>
<th>Scores (real limits)</th>
<th>Frequency</th>
<th>Scores (real limits)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38.5 - 40.5</td>
<td>1</td>
<td>67.5 - 70.5</td>
<td>2</td>
</tr>
<tr>
<td>36.5 - 38.5</td>
<td></td>
<td>64.5 - 67.5</td>
<td>2</td>
</tr>
<tr>
<td>34.5 - 36.5</td>
<td>1</td>
<td>61.5 - 64.5</td>
<td>4</td>
</tr>
<tr>
<td>32.5 - 34.5</td>
<td>3</td>
<td>58.5 - 61.5</td>
<td>5</td>
</tr>
<tr>
<td>30.5 - 32.5</td>
<td>2</td>
<td>55.5 - 58.5</td>
<td>13</td>
</tr>
<tr>
<td>28.5 - 30.5</td>
<td>17</td>
<td>52.5 - 55.5</td>
<td>23</td>
</tr>
<tr>
<td>26.5 - 28.5</td>
<td>19</td>
<td>49.5 - 52.5</td>
<td>29</td>
</tr>
<tr>
<td>24.5 - 26.5</td>
<td>47</td>
<td>46.5 - 49.5</td>
<td>38</td>
</tr>
<tr>
<td>22.5 - 24.5</td>
<td>38</td>
<td>43.5 - 46.5</td>
<td>46</td>
</tr>
<tr>
<td>20.5 - 22.5</td>
<td>70</td>
<td>40.5 - 43.5</td>
<td>49</td>
</tr>
<tr>
<td>18.5 - 20.5</td>
<td>83</td>
<td>37.5 - 40.5</td>
<td>62</td>
</tr>
<tr>
<td>16.5 - 18.5</td>
<td>75</td>
<td>34.5 - 37.5</td>
<td>50</td>
</tr>
<tr>
<td>14.5 - 16.5</td>
<td>74</td>
<td>31.5 - 34.5</td>
<td>64</td>
</tr>
<tr>
<td>12.5 - 14.5</td>
<td>50</td>
<td>28.5 - 31.5</td>
<td>52</td>
</tr>
<tr>
<td>10.5 - 12.5</td>
<td>29</td>
<td>25.5 - 28.5</td>
<td>43</td>
</tr>
<tr>
<td>8.5 - 10.5</td>
<td>18</td>
<td>22.5 - 25.5</td>
<td>24</td>
</tr>
<tr>
<td>6.5 - 8.5</td>
<td>3</td>
<td>19.5 - 22.5</td>
<td>17</td>
</tr>
<tr>
<td>4.5 - 6.5</td>
<td>1</td>
<td>16.5 - 19.5</td>
<td>7</td>
</tr>
</tbody>
</table>

| N       | 530 | 530 | 530 |
| M       | 19.1 | 19.3 | 38.4 |
| M'n     | 18.9 | 18.9 | 37.9 |
| s²      | 25.97 | 33.19 | 100.60 |
| s       | 5.10 | 5.76 | 10.03 |
FIGURE 2

DISTRIBUTION OF FINAL SCORE OF 530 STUDENTS ON THE FINAL TEST
FIGURE 3(a)
DISTRIBUTION OF KNOWLEDGE-SCORES
OF 530 STUDENTS ON THE FINAL TEST

FIGURE 3(b)
DISTRIBUTION OF COMPREHENSION-SCORES
OF 530 STUDENTS ON THE FINAL TEST
Reliability of the test and relevance categories. As noted in Chapter IV, the final test was constructed in two subtests, matched as closely as possible for content, relevance category, and difficulty. Test reliability was determined by correlating these "equivalent forms," and applying to the result the Spearman-Brown formula. The correlation coefficient between the two subtests was .72. The Spearman-Brown formula

\[ R_{11} = \frac{2r_{12}}{1 + r_{12}} \]
gave an estimated test-reliability coefficient of .84. Despite the fact that the procedure used took account of variation in the individual from day to day, as well as that due to sampling of behaviours, this figure is lower than had been hoped. This low figure may partly be explained by the fact, noted by Thorndike, that maximum reliability results when the test items are of 50% difficulty level. Other reasons may include the presence of guessing on the test, and the rather small number of items used. Whatever the reasons, the reliability of the test is not as high as could have been desired.

Reliability coefficients for the knowledge items and the comprehension items, similarly obtained, were .69 and .77 respectively. Again, these coefficients, based on only forty items each, are rather low. The correlation coefficients from which they were estimated, by the Spearman-Brown formula, were .53 and .63, respectively.

---


The question of whether these could be considered significantly different arose since, if they were, the true variances of the knowledge and comprehension distributions might also be different. Because these correlation coefficients were not independent, being based on tests administered to the same individuals, the usual statistical test was not applicable. The following alternative approach was therefore used.

The figure, .58, intermediate between the two correlation coefficients, was selected for testing as a possible population-value for each correlation, and the technique described by Walker and Lev was employed for this purpose.\(^3\) Given that the observed value of \(r\) is .53, and it is desired to test the hypothesis that the population-value, \(\rho\), is greater than or equal to .58: the quantity, \((z_r - \bar{Z})\sqrt{N - 3}\), may be treated as a normally distributed variable, where \(z_r\) and \(\bar{Z}\) are the values of Fisher's-z corresponding to \(r\) and \(\rho\), respectively. Here, \(z_r = .5901\) and \(\bar{Z} = .6625\). For \(N = 530\), the normal deviate for this case was calculated to be -1.66, a value which, for a one-tailed test, is significant at the 5% level. The hypothesis was therefore rejected.

For the comprehension items, \(r\) had the value .63, and it was desired to test the hypothesis that \(\rho\) is less than or equal to .58. By a similar procedure, the value of the normal deviate was found to be 1.73, significant at the 5% level. Again, the hypothesis was rejected.

In summary of this argument, it is unlikely, at the 5% level, that the population-value of the correlation coefficient for the knowledge items is as great as .58, and also that the population-value for the comprehension

items is as low as .58. It is even more unlikely that both of these improbable events should be true simultaneously. It is reasonable to conclude, therefore, that the correlation of the comprehension half-tests is significantly higher than that of the knowledge half-tests, and--by extension--that the reliability of the comprehension items is also significantly higher.

Returning to the distribution of scores, and accepting these reliability coefficients at their face value, it was possible to estimate approximately the true variance of the knowledge and comprehension scores, by using the formula \[ s^2 = s_x^2 \times r_{xx} \]. This formula, which embodies, for a sample, the definition of the reliability coefficient, is taken from Walker and Lev.\(^4\) Using the obtained variances, shown in Table VI, the formula gave the figures 17.92 and 25.56 as the approximate true variances of the knowledge and comprehension distributions, respectively. The corresponding error-variances may be shown, by subtraction, to be 8.05 and 7.63--approximately the same. Thus, it may be concluded that the individuals who wrote the test exhibited greater genuine differences on the comprehension items than they did on the knowledge items; and, in addition, their performance on those items was more reliably determined. In view of the small samples of items which could be used, it seems unwise to attempt to generalize from this finding, but it opens up interesting possibilities for further research.

The relationship between the knowledge and comprehension scores.

The correlation between the knowledge and comprehension scores was .71.

\(^4\)Ibid., p.298.
The purpose of calculating this was to gain some idea of the extent to which the knowledge and comprehension items are measuring different functions. Strictly, this figure should be corrected for attenuation, in view of the low reliability of the measurements on which the correlation was based. However, as noted by Walker and Lev, the usual correction for attenuation employs a formula obtained by assuming, among other things, that the error-components of the knowledge and comprehension scores are uncorrelated. This can scarcely hold here, since the two kinds of items appeared equally in each subtest of the final test; thus, temporary distractions, for example, would affect both kinds of items. The correction for attenuation would thus overestimate, by an unknown amount, the correlation which would be found to exist between the knowledge and comprehension categories if the measurements were perfectly reliable. The authors of the Taxonomy cite no correlation coefficients to support their claim that their categories deal with different levels of complexity of behaviour, but coefficients reported in the literature range from .45 to .87, between tests of recall and application, within a single subject field. The first was found by Furst, and the second by McConnell. This range gives little guidance for comparison in the present situation, so that reliance must continue to be placed on the care with which the items were constructed, and double-checked, in accordance with the specifications of the Taxonomy.

5Ibid., p.299.
II. DISTRIBUTIONS OF DIFFICULTY INDICES

As at the tryout stage, the difficulty indices obtained from the final test were transformed by Davis' procedure. These transformed indices are shown in Table VII, and their cumulative frequency curves in Figure 4 below. It is clear that the matching process was partially but not entirely successful. The means are still almost identical, though about one unit larger. However, the dispersion of the knowledge items has increased while that of the other has decreased; the comprehension items have turned out to be more homogeneous than the knowledge items, in respect of difficulty. This fact may be partly responsible for their greater reliability.

Despite the agreement in means and in general form of distribution, the presence of one very easy knowledge item and two or three quite difficult ones, not matched by corresponding comprehension items, gives rise to concern as to their effect on the discrimination indices. As noted in Chapter I, the value of the point-biserial correlation coefficient cannot rise very high for very easy or very hard items. The presence of these few more extreme difficulty indices in the knowledge category gives an advantage to the comprehension category in respect of any test of significance involving the discrimination indices.

III DISTRIBUTIONS OF DISCRIMINATION INDICES

Point-biserial correlation coefficients were obtained for all items. For reference purposes, these are listed, for each item, in Appendix D, together with the item's Taxonomy-category, its final percentage-difficulty index, and its biserial correlation coefficient.
**TABLE VII**

**DISTRIBUTION OF TRANSFORMED DIFFICULTY INDICES IN THE KNOWLEDGE AND COMPREHENSION CATEGORIES**

<table>
<thead>
<tr>
<th>Transformed difficulty indices (real limits)</th>
<th>Knowledge items $f$</th>
<th>Comprehension items $f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>78.75 - 81.25</td>
<td>1</td>
<td>40</td>
</tr>
<tr>
<td>76.25 - 78.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>73.75 - 76.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>71.25 - 73.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>68.75 - 71.25</td>
<td>1</td>
<td>39</td>
</tr>
<tr>
<td>66.25 - 68.75</td>
<td>1</td>
<td>38</td>
</tr>
<tr>
<td>63.75 - 66.25</td>
<td>3</td>
<td>37</td>
</tr>
<tr>
<td>61.25 - 63.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>58.75 - 61.25</td>
<td>4</td>
<td>34</td>
</tr>
<tr>
<td>56.25 - 58.75</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>53.75 - 56.25</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>51.25 - 53.75</td>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>48.75 - 51.25</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>46.25 - 48.75</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>43.75 - 46.25</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>41.25 - 43.75</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>38.75 - 41.25</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>36.25 - 38.75</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>33.75 - 36.25</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>31.25 - 33.75</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**N** 40  40

**M** 68.9(47.8)  49.0(47.9)

**Md** 66.3(46.0)  47.3(46.7)

**s** 10.76(9.03)  8.50(9.55)

\(^1\)The figures in parenthesis are the corresponding values from the tryout stage.
FIGURE 4

CUMULATIVE DISTRIBUTIONS OF TRANSFORMED DIFFICULTY INDICES OF ITEMS IN THE KNOWLEDGE AND COMPREHENSION CATEGORIES, OBTAINED FROM THE ADMINISTRATION OF THE FINAL TEST

KEY:

--- Comprehension Items
--- Knowledge Items
According to the procedure described in Chapter I, Fisher's $z$-transformation was applied to each point-biserial coefficient. The distributions of these $z$-values for both relevance categories are shown in Table VIII, and the corresponding histograms in Figure 5, below.

These distributions exhibit the irregularity of form which often accompanies the use of small samples. Although both are somewhat negatively skewed, they have a predominance of cases in the middle range and fewer at the extremes. Nevertheless, neither can be considered to have been drawn from a normal population. Among the factors responsible for this, three may be cited: (1) the use of only forty cases in each category does not provide a very typical sample of all possible items which might have been written; (2) at the tryout stage, the negatively discriminating items were rejected, thus disrupting the randomness of selection; and (3) while the point-biserial correlation coefficient is unrestrictedly used for item analysis by test constructors, and under certain conditions its sampling distribution is known for a given item, there is little that can be said about its expected distribution for a set of items.

It had been hoped that the obtained distributions of discrimination indices would, despite the considerations of the previous paragraph, be (a) reasonably symmetrical, (b) similar in form, and (c) about equally variable; granted these conditions, the empirical work of Boneau and Norton on using the $t$-test and the $F$-distribution under considerable violation of the assumptions, could have been invoked to justify a parametric significance test. However, the irregularity of the obtained

---

8C. Alan Boneau, "The Effects of Violations of Assumptions Underlying the $t$ Test," Psychological Bulletin, 57 (No. 1, 1960), 49-64.

**TABLE VIII**

DISTRIBUTION OF DISCRIMINATION INDICES OF KNOWLEDGE AND COMPREHENSION ITEMS

<table>
<thead>
<tr>
<th>z-values (real limits)</th>
<th>Knowledge items $f$</th>
<th>Comprehension items $f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.4725 - .4975</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>.4475 - .4725</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>.4225 - .4475</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>.3975 - .4225</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>.3725 - .3975</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>.3475 - .3725</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>.3225 - .3475</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>.2975 - .3225</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>.2725 - .2975</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>.2475 - .2725</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>.2225 - .2475</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>.1975 - .2225</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>.1725 - .1975</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>.1475 - .1725</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>.1225 - .1475</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>.0975 - .1225</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>.0725 - .0975</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>.0475 - .0725</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>.0225 - .0475</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>-.0025 - .0225</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-.0275 - -.0025</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-.0525 - -.0275</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>40</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{z}$</td>
<td>.2580</td>
<td>.2908</td>
</tr>
<tr>
<td>$\bar{r}_{pb}$</td>
<td>.252</td>
<td>.283</td>
</tr>
<tr>
<td>$\text{Min}(z)$</td>
<td>.27</td>
<td>.30</td>
</tr>
</tbody>
</table>

1 These are Fisher's z-values corresponding to the point-biserial correlation coefficients obtained for each item.
These are Fisher's z-values corresponding to point-biserial correlation coefficients.
distributions is such that they cannot be said to resemble any of the populations which Boneau and Norton sampled in their investigations, so that their conclusions cannot be considered applicable. Accordingly the decision was taken to use the non-parametric Mann-Whitney test already employed in Chapter IV.

**IV. THE STATISTICAL TEST OF THE HYPOTHESIS**

The statistical hypothesis to be tested was the null hypothesis,

\[ H_0: \text{Mdn}_K - \text{Mdn}_C = 0, \]

the admissible alternative hypothesis being \( H_1: \text{Mdn}_C > \text{Mdn}_K \). This, therefore, required a one-tailed test. The level of significance was pre-set at the rather severe level of 1%. It was recognized that this would increase the risk of committing a Type II error, but it was considered better that the status quo should be maintained, even if false, rather than that a possibly true (null) hypothesis should be incorrectly rejected.

Ranks were assigned to the discrimination indices, and the procedure described in Chapter IV followed. The discrimination indices -- still expressed in terms of Fisher's \( z \) -- together with their ranks, are listed in Table IX. The normal deviate

\[ z = \frac{2R_K - N_K(N + 1)}{\sqrt{N_KN_C(N + 1)} / 3} \]

was calculated, using the values shown in Table IX, and its value found to be 1.26. For the one-tailed test, the probability of obtaining a value of \( z \) greater than 1.26 was determined from the table of the normal probability integral to be .1038. There was thus no basis for rejecting the null hypothesis, which was therefore accepted.
### TABLE IX
VALUES AND RANKS OF DISCRIMINATION INDICES\(^1\) OF ITEMS IN KNOWLEDGE AND COMPREHENSION CATEGORIES

<table>
<thead>
<tr>
<th>Discrimination indices</th>
<th>K</th>
<th>C</th>
<th>Ranks of indices</th>
<th>K'</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.456</td>
<td>.476</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.421</td>
<td>.462</td>
<td>7.5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.417</td>
<td>.454</td>
<td>9</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.414</td>
<td>.443</td>
<td>10</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.382</td>
<td>.439</td>
<td>11</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.368</td>
<td>.421</td>
<td>13</td>
<td>7.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.363</td>
<td>.374</td>
<td>15</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.347</td>
<td>.365</td>
<td>20</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.341</td>
<td>.360</td>
<td>22</td>
<td>16.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.339</td>
<td>.360</td>
<td>23</td>
<td>16.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.328</td>
<td>.356</td>
<td>27</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.327</td>
<td>.349</td>
<td>28</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.325</td>
<td>.344</td>
<td>29</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.310</td>
<td>.336</td>
<td>33.5</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.304</td>
<td>.334</td>
<td>35</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.296</td>
<td>.332</td>
<td>37</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.289</td>
<td>.321</td>
<td>39</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.288</td>
<td>.316</td>
<td>40</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.287</td>
<td>.312</td>
<td>41</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.283</td>
<td>.310</td>
<td>42</td>
<td>33.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.267</td>
<td>.298</td>
<td>43</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.265</td>
<td>.292</td>
<td>45.5</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.265</td>
<td>.266</td>
<td>45.5</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.259</td>
<td>.264</td>
<td>49</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.251</td>
<td>.262</td>
<td>50</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.242</td>
<td>.243</td>
<td>52</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.220</td>
<td>.235</td>
<td>56</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.195</td>
<td>.234</td>
<td>60</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.187</td>
<td>.225</td>
<td>61</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.186</td>
<td>.211</td>
<td>62</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.171</td>
<td>.206</td>
<td>68</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.152</td>
<td>.200</td>
<td>70</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.149</td>
<td>.185</td>
<td>71</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.131</td>
<td>.184</td>
<td>73</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.127</td>
<td>.180</td>
<td>74</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.124</td>
<td>.176</td>
<td>75</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.121</td>
<td>.175</td>
<td>76</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.120</td>
<td>.170</td>
<td>77</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.051</td>
<td>.135</td>
<td>78</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-.049</td>
<td>.025</td>
<td>80</td>
<td>70</td>
<td></td>
</tr>
</tbody>
</table>

\(^{1\text{Discrimination indices reported are Fisher's z-values of point-biserial correlation coefficients.}}\)

\[ N_K = 40 \quad N_C = 40 \quad R_K = 1751 \quad R_C = 1489 \]

\[ N = N_K + N_C = 80 \]
For comparison, the same test applied to the biserial correlation coefficients is offered. These coefficients are listed in Appendix D, for reference. Here, \( R_K \) was found to be 1716; and using the values 40, 40, and 80 for \( N_K \), \( N_G \), and \( N \), respectively, the value of the normal deviate was calculated to be .9238. The probability of obtaining a value greater than this is .18, so that using the biserial-\( r \), the null hypothesis must still be accepted.

In view of the non-significance of the difference, it was concluded that the observed difference could be ascribed to sampling error, and that there is no statistical basis for maintaining that the comprehension items discriminate better than do the knowledge items.
CHAPTER VI

CONCLUSIONS AND SUMMARY

This investigation was designed to avoid the inadequacies of the earlier one reported by Cook. It is believed that this objective was achieved, but unfortunately the present study has weaknesses of its own. It is in the light of these that conclusions must be stated.

I. WEAKNESSES

The major weakness lay in the small number of items used. Among the effects of this were the following.

First, although an objective of the Science 10 course is to develop functional problem-solving ability, the application items generally proved too hard; too few of moderate difficulty could be found. A larger number of items would have provided more items of moderate to low difficulty, and avoided the loss of what might have proved to be a set of items of high average discrimination.

Second, the almost rectangular distribution of difficulty indices shown in Table VII makes it clear that though the mean difficulty-level in both categories was around 50%, there were too many rather difficult items. Again, a larger number of items would have made some control of the distribution of difficulty possible. In any event, it may be surmised that the presence of these more difficult items increased the tendency to guess, thus contributing to lower the reliability of the test and its categories.
This unduly low reliability constitutes the third effect of small numbers. The generalized Spearman-Brown formula\(^1\) indicates that twice as many items of the same quality should produce a reliability coefficient for the total test of over .90. It is likely that, with a better distribution of difficulty indices, the same value could be produced with fewer than twice as many items. In passing, however, it may be remarked that in an experiment of this kind, really high reliability may always remain elusive, since the design forbids the rejection of poorly discriminating items during construction of the final test. The final test, therefore, can never be thoroughly refined.

Finally, on this point, a larger number of items was required to smooth out the distributions of discrimination indices and to permit their form to become apparent.

Another source of weakness lay in the use of two tryout groups, though this was inevitable in view of the lack of testing time available -- also the root-cause of the use of so few items. However, its disadvantages could have been avoided to some extent by the use of larger numbers of students in each group. Greater stability of the difficulty indices might have resulted.

II. CONCLUSIONS

The weaknesses noted in the foregoing reduce the precision with which conclusions can be drawn. The statistical tests provide no basis for believing the observed discrepancy in average discrimination index to be anything but sampling error. Nevertheless, the results obtained are

in the right direction; and inspection of Table IX indicates that the superiority of the comprehension items over the knowledge items is highly consistent, being evident from top to bottom of the table. The possibility cannot be ignored that the use of a larger number of items -- had this been possible to time-table -- might have increased the reliability of the criterion scores, smoothed the distributions of discrimination indices, and generally improved the precision of the experiment to the point where a significant discrepancy might have been obtained. In addition, it is impossible to avoid speculating whether the apparent trend in the direction hypothesized would have been continued by the application items, had they survived.

This is not the definitive experiment on this problem.

III. POSSIBILITIES OF FURTHER RESEARCH

With the exception of one point, the general design of the investigation appears sound. This point concerns the distribution of difficulty indices. It was observed in Chapter V that we cannot predict the distribution of point-biserial correlation coefficients to be expected for a set of items. We can, however, say that for a given item, the sampling distribution of the point-biserial coefficient will be that of \( r \) under certain conditions; and under the same conditions the sampling distribution of the corresponding Fisher's \( z \)-values will be approximately normal. These conditions are that during sampling the number of students "passing" the item and the number "failing" it remain constant, and that the criterion scores within each category, "pass" or "fail," are normally distributed. For a set of items

these conditions are unlikely to be met. In practice they could best be approximated by choosing items such that their transformed difficulty-indices were roughly normally distributed about a selected value and had a fairly narrow range. In this way the number of students "passing" and "failing" each item would not be too far from constant. Making allowance for the rejection at the tryout stage of a small number of negatively discriminating items, the distribution of Fisher's $z$-values corresponding to point-biserial coefficients might not be too far from normal. The appropriate parametric test could then be applied.

It is therefore suggested that the present design be retained, but that greater control be exercised over the difficulty indices of the items selected for the final test. The use of items with transformed difficulty indices distributed approximately normally about 50, with a standard deviation of about 10, would have certain advantages. First, very few items would be exceptionally difficult; most would have difficulty indices between 40 and 60. Thus guessing should be reduced and test reliability improved. In addition, conclusions could be more precisely drawn since they would refer to items whose difficulty-distribution was specified.

Further, it will be necessary to increase substantially the number of items used at the tryout stage. The number used in the final test should also be increased by perhaps fifty per cent. It is therefore plain that the co-operation of the provincial educational authority would be necessary to implement the expanded design.

It will be recalled from Chapter V that the comprehension items manifested a much larger true-variance than did the knowledge items, while exhibiting about the same error-variance. It would be interesting to know whether it is generally characteristic of comprehension items -- and,
indeed, higher category items of all kinds — to "spread" the scores of students more effectively. If it were shown that higher category items did in fact possess superior discrimination, then the ability to "spread" more effectively the scores of students would be a useful adjunct to their discriminating power. This would be particularly important where, as in the case of a cut-off test, it is desired to discriminate among students with particular precision at a stipulated level of difficulty.

Unconnected directly with the present study, but related to it, is the question of the extent to which schools are actually working toward developing in students the ability to apply principles, scientific or otherwise, to the solution of unfamiliar problems. The evidence of this study suggests that there may be a weakness here.

IV. SUMMARY OF THE INVESTIGATION

It was hypothesized, on a priori grounds, that on a self-defining achievement test, equally weighted with items sampling behaviours classifi-
able in the Knowledge, Comprehension, and Application categories of Bloom's Taxonomy of Educational Objectives, the mean discrimination indices of the items in each category would show an increase from the first- to the last-named category. The distributions of difficulty indices within the three categories were to be similar, to avoid weighting the total or criterion scores unequally with items from any one category, and thus affecting the discrimination indices. The discrimination index chosen was the point-biserial correlation coefficient, since of those indices which utilize all information provided by the test, it is the only one which has properties permitting the employment of a parametric test of significance. The subject chosen was General Science, and the grade level was nine.
Items were written categorized according to Taxonomy-specifications, and checked by two colleagues. Assembled into two subtests, half of the items were written by one tryout group, and the other half by a second tryout group, limitations of testing time forbidding the use of a single group. Both groups were reasonably, though not highly, similar in scholastic aptitude to the final group. Items which discriminated negatively were discarded. Of the remainder, the application items were found to be so difficult that they had to be rejected.

The knowledge and comprehension items were then matched for difficulty, necessitating further rejections until only forty remained in each category. Two subtests were then assembled, their items matched for relevance category, content, and difficulty, forming essentially "equivalent forms."

Five hundred thirty students wrote both subtests. The reliability coefficient of the total test was .81, and those of the knowledge and comprehension "sub-tests" were .69 and .77, respectively.

Revised difficulty indices were obtained. The adequacy of the original matching was found to have been fair, but not completely satisfactory; the two categories had almost identical mean difficulty indices, but the knowledge category contained two or three very hard items and some easy ones, unmatched by corresponding comprehension items. This was unfortunate, since such items are restricted with respect to the size of point-biserial discrimination index which they can attain.

The point-biserial coefficients were transformed to Fisher's $z$-values, and the distributions graphed. These were not sufficiently close to the normal form to justify assuming normality in the population. Thus,
the non-parametric Mann-Whitney test was employed, rather than the $t$-test. At the 1% level, the null hypothesis of no difference in medians had to be accepted. This decision was confirmed by application of the same test to the biserial correlation coefficients, whose advantage is that their maximum size is relatively unaffected by difficulty level.

It was concluded that the hypothesis could not be sustained on the basis of this experiment. However, certain inadequacies and trends combined to suggest that the attack on the problem be not dropped. Suggestions were made for refining the experimental design, and for further research.
A. BOOKS


B. PUBLICATIONS OF GOVERNMENT AND OTHER ORGANIZATIONS


C. PERIODICALS


<table>
<thead>
<tr>
<th>Unit</th>
<th>Content Category</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1</td>
<td>Units, systems of measurement; characteristics of science.</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Forms of matter -- elements, compounds, mixtures, etc.; properties; physical and chemical change; chemical activity.</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Acids, bases, salts; electrolysis; types of reaction.</td>
</tr>
<tr>
<td>II</td>
<td>1</td>
<td>Force, energy, work; elementary problems involving work. Kinetic and potential energy. Conservation of energy. Forms of energy; energy-transformations.</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Magnetic properties of some materials; power of iron to &quot;concentrate lines of force;&quot; temporary, permanent magnets. Molecular theory of magnetism. Induction and disruption of magnetism. Variation; inclination.</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Magnetic effects of electric current; increasing magnetic field of &quot;live&quot; wire. Relative motion of circuit-wire and magnetic field produces current. Dynamo; A.C., D.C.</td>
</tr>
</tbody>
</table>
II 

Current as electron-flow; conductors, insulators. 
Potential difference. Resistance; factors affecting. 
Chemical generation of electricity; cells. Circuits. 

APPENDIX B
TEST-DIRECTIONS
(EXCERPTS)

I. PRELIMINARY ARRANGEMENTS

1. Every student should be warned in advance to provide himself with
   a) 2 pencils
   b) an eraser
   c) some scrap-paper, for rough calculations

2. Each supervisor should have a watch with a second-hand, so that
   he may be able to adhere strictly to the time-limits.

II. PRETEST DIRECTIONS

Since each test will last exactly 50 minutes, schools operating on a 50-minute period are requested to take a few minutes at the end of the science-period immediately preceding the test-series to read the following test-directions to the participating classes.

Schools operating on a 60-minute period will have time to read these directions to the classes at the beginning of the first test-period.

III. DIRECTIONS TO STUDENTS

Please read these without attempting to elaborate.

1. Many of the items on these tests will appear different from any that you have seen before. Most of them require a little bit of thinking as well as just remembering something that you have learned. However, ALL of them CAN be done by any student who has learned the work done to date WELL. So however strange or hard a question may appear at first sight, remember ---- YOU KNOW ENOUGH TO ANSWER IT; all it needs is your knowledge plus a little thought!

2. You may answer questions, even when you are not completely sure that your answers are correct. In such cases, intelligent consideration of the choices provided may help you to gain marks. HOWEVER, you should AVOID WILD GUESSING as this may result in a reduction of your score.
3. Give each question careful thought, but work as quickly as you can. If you find a question too difficult, do not linger over-long on it. Pass on to the next ones, and return later to any that you missed, if you have time.

4. To show you how to mark each item, here is a sample item, taken from the first unit of your text. (Teachers should demonstrate this at the board.)

ITEM: Which of the following formulae represents a substance that might be expected to form a salt if reacted with vinegar?

a) HCl  b) NaOH  c) NaCl  d) H2O

c)  d) 

The correct answer is b), NaOH, so we write a cross (x) in the parentheses to the right of b, as shown. Make sure that the cross is actually in the parentheses or it may be overlooked and the item marked wrong.

5. Mark only ONE of the 4 choices given for each item.

6. As soon as you receive your paper, complete the required information on the front page (i.e., NAME, LOCATION OF SCHOOL, AND U.P. or G.P.) Do not turn the page until you receive the signal to start.

7. On tests of this kind, it is not expected that any student will get all of the items correct; in fact, it is usual for most average students to score around 50%. The pass-mark will undoubtedly be much less than 50%, so if you have to leave quite a few questions unanswered, do not worry -- and DO NOT GUESS.

8. Are there any questions, since you may not ask questions during the test (except regarding misprints or missing pages or things like that.)?

IV. DIRECTIONS FOR ADMINISTRATION

If the directions above were read during the science-period preceding the test, please check to see if any student is present for the test who was absent for the directions. Such a student may write the test, but his paper should bear a notation explaining that he missed the directions, since his results will be of no use to me in my project.

1. Hand out papers; remind students, as you do so, to complete the blanks on the front page and wait for the signal to start.
2. Say: YOU HAVE 50 MINUTES. START NOW. (Note precise time of starting.)

3. After exactly 50 minutes, say: STOP WORK. DRAW A LINE UNDER THE HIGHEST-NUMBERED QUESTION YOU HAD TIME TO CONSIDER, WHETHER YOU WERE ABLE TO ANSWER IT OR NOT . . . . IF YOU HAD TIME TO CONSIDER ALL OF THE ITEMS, PUT A CHECK MARK IN THE CIRCLE AT THE BOTTOM OF THE FRONT PAGE, WHETHER YOU WERE ABLE TO ANSWER ALL OF THEM OR NOT.

4. Collect the papers.
APPENDIX C
This simplified diagram represents a "hydro-pole" from which electrical energy is being drawn off to a house.

CHECK HERE ➔

if you had time to consider all items.
SCIENCE 10 TEST

SUB-TEST I

1. Which of the following types of chemical reaction best describes that represented by the equation $H_2CO_3 \rightarrow H_2O + CO_2$? a) Combination b) Decomposition c) Replacement d) Neutralization

2. Which of the following comparisons of lime-mortar and concrete is true? a) Both harden under water. b) Both set to about the same degree of hardness. c) Only lime-mortar undergoes chemical change in hardening. d) Both use limestone at some stage of their preparation.

3. (THIS ITEM REFERS TO THE DIAGRAM ON THE FRONT PAGE.) In this diagram, box A contains a) an electric generator to produce current at 110 volts for the house-circuits. b) a circuit-breaker to protect the house-circuits from overloading. c) a metering-device to measure the number of kilowatt-hours used each month. d) a device to reduce to 110 volts the current supplied to the house.

4. Which of these compounds is used in fire extinguishers and baking powders? a) magnesium hydroxide b) magnesium sulphate c) sodium bicarbonate d) sodium carbonate

5. The statement has been made: "Never replace the fuse in a houselightning circuit with one of a higher ampere-rating, except on the advice of a skilled electrician."
The main reason for this is that

a) you may allow dangerous amounts of heat to form.
b) you may cause your range or water-heater to burn out.  a( × )
c) you may receive a dangerous electric shock. b(   )
d) you may "blow" the other fuses and lose their protection. c(   )

d) you may "blow" the other fuses and lose their protection. d(   )

6. The coil and soft-iron core of a dynamo or magneto are together referred to as the

a) armature  a( × )
b) commutator  b(   )
c) field- or electro-magnet  c(   )
d) split-ring and brushes d(   )

7. Machine A can raise a 50-lb. weight to a height of 40 feet in 5 seconds, while machine B can raise a 100-lb. weight to a height of 20 feet in 10 secs.

If A and B perform these tasks, the work done by A will be

a) one-quarter that done by B  a(   )
b) one-half that done by B  b(   )
c) the same as that done by B  c( × )
d) twice that done by B  d(   )

8. When an ebonite rod is charged by rubbing with fur

a) positive charges are removed from the fur and transferred to the rod.  a(   )
b) negative charges are removed from the fur and transferred to the rod.  b( × )
c) positive charges are removed from the rod and transferred to the fur.  c(   )
d) negative charges are removed from the rod and transferred to the fur.  d(   )

9. The firing of a rifle is an illustration of the principle that before useful work can be done

a) potential energy must be converted to kinetic energy.  a( × )
b) mechanical energy must be converted to potential energy.  b(   )
c) potential energy must be converted to chemical energy.  c(   )
d) mechanical energy must be converted to kinetic energy.
10. Which of the following statements is true of a compound but not of a mixture?

The original ingredients

a) can be separated out by heating. a( )
b) may include both elements and compounds. b( )
c) no longer have their original characteristics. c( x )
d) have the same total energy as the product(s). d( )

11. Whether a material can be magnetized depends entirely upon whether or not

a) it is composed at least partially of iron. a( )
b) it is composed of either iron or nickel. b( )
c) it is a metal, possessing lines of force. c( )
d) it is composed of magnetic particles. d( x )

12. (THIS ITEM REFERS TO THE DIAGRAM ON THE FRONT PAGE.)

In this diagram, the letters, B, refer to a device which is intended to

a) reduce loss of electrons through contact with conducting materials. a( x )
b) prevent people who touch the lower part of the pole from receiving electric shocks. b( )
c) prevent damage to wires and pole during lightning storms. c( )
d) reduce danger of overheated wires by lowering resistance. d( )

13. The metals iron, sodium and gold, differ greatly in their chemical activity. In which of the following are these metals arranged in order of increasing activity (i.e., least active first)?

a) iron, sodium, gold c) gold, sodium, iron a( )
b) gold, iron, sodium d) sodium, gold, iron. b( x )

14. Which of the following materials is a mixture?

a) air b) water vapour c) carbon dioxide a( x )
d) iron sulphide b( )

15. In which of the following materials is the Law of Definite Proportions best illustrated?

a) air    b) iron    c) concrete    d) water

16. According to the Electron Theory of matter, most of the elements tend to combine with others if the number of electrons in the outer orbit of each of their atoms is different from 8.

Which of the following elements is most likely to have an outer orbit containing exactly 8 electrons?

a) chlorine    b) mercury    c) argon    d) calcium

17. A wire carrying a strong, direct current lies in a north-south direction. Over it is held a compass. If north is in the direction indicated by the arrow at the right, which of the following most nearly represents the new direction of the compass needle?

18. What is the main advantage of using soft iron rather than steel as a material for the core of a lifting magnet?

a) It has less electrical resistance and thus conducts the current more efficiently.

b) Its amount of magnetism changes when the current-strength is varied.

c) It obstructs the passage of lines of force, which are thus available for work in lifting objects.

d) It is less affected by the heat produced by the flow of electrical energy through the coils.
19. Consider the electrolysis of water.

Which of the following statements, taken from your text, contains the explanation of why the passage of an electric current produces chemical change?

In chemical change

a) the reacting materials always change to different materials.  

b) the new substances formed always have properties different from those of the reacting materials.

c) energy is always given off or absorbed.

d) the weights of the reacting substances always equal the weights of the new substances formed.

20. To detect (or discover) the presence of a static charge on a body, an appropriate instrument to use would be

a) a galvanometer  

b) an electroscope  

c) an induction coil  

d) an ammeter

21. A certain compound is burned. The products are analyzed and found to be water vapour and carbon dioxide.

Which of the formulae listed below might reasonably be considered to be that of the original compound?

a) $C_2H_2$  

b) CO  

c) Ca(OH)$_2$  

d) CaH$_2$

22. The resistance of a circuit carrying a direct electric current may be raised by

a) using a lower voltage than before.  

b) coiling a part of the wire.  

c) shortening the length of the circuit.  

d) using thinner wire of the same material and length.
A gas, containing no moisture, was passed through tube A, containing heated copper oxide (CuO), and then through tube B, containing calcium chloride, which readily absorbs moisture. At the end of the experiment, a considerable amount of the CuO was found to have been transformed to pure copper (Cu), and the calcium chloride in tube B had become damp.

Of the following four gases, the one used in this experiment could have been

- a) oxygen
- b) hydrogen
- c) chlorine
- d) carbon dioxide

24. The reason that a dry cell is able to produce a steady current for much longer than an ordinary voltaic cell is that

- a) it has a carbon electrode which does not react with the electrolyte.
- b) it contains particles of carbon which reduce the cell's internal resistance.
- c) it contains a substance which reduces the rate at which hydrogen collects.
- d) its electrolyte, ammonium chloride, reacts with zinc less vigorously than does sulphuric acid.

25. A form of slaked lime is sometimes spread on the soil of gardens. The reason for this is that slaked lime

- a) is acidic in nature and hence useful in correcting soils containing too much alkali.
- b) is a base and hence useful in neutralizing sour soil.
- c) reacts with water to form a substance which makes sandy soil less porous.
- d) contains chemical elements which make it a useful fertilizer.
26. The statement, "These exist only in sets or pairs", applies
   a) to both electric charges and magnetic poles.           a(  )
   b) to neither electric charges nor magnetic poles.      b(  )
   c) only to electric charges and not to magnetic poles.  c(  )
   d) only to magnetic poles and not to electric charges.  d(  )

27. Consider the following simplified "equations". (Pt is the
      chemical symbol for platinum.)

   \[ \text{SO}_2 + \text{O}_2 \rightarrow \text{little reaction at } 400^\circ\text{C}. \]
   \[ \text{SO}_2 + \text{O}_2 + \text{Pt} \rightarrow \text{SO}_3 + \text{Pt} \text{ at } 400^\circ\text{C}. \]

   These equations most strongly illustrate
   a) a typical replacement reaction.                       a(  )
   b) the chemical activity of platinum.                   b(  )
   c) the operation of a catalyst.                         c(  )
   d) the formation of a radical.                          d(  )

28. A voltaic cell transforms chemical energy to electrical energy.

   The chemical energy comes from
   a) the zinc and sulphuric acid.                          a(  )
   b) the copper and sulphuric acid.                       b(  )
   c) the copper and zinc plates.                         c(  )
   d) the zinc, copper and sulphuric acid.                 d(  )

29. During a thunderstorm, the safest place to stay (of those
      listed below) is

   a) in the middle of a large, open, level field.         a(  )
   b) under a tree in the middle of a large, open, level  b(  )
      field.
   c) in a building with a steel frame.                   c(  )
   d) in the bathtub.                                    d(  )

30. Nowadays helium is used, rather than hydrogen, in most

   balloons employed by the armed forces.

   This is accounted for, or explained, by the fact that helium
a) is somewhat less easily oxidized than hydrogen.  
a) 
b) is much more chemically inert than hydrogen.  
b( X )
c) has greater density than hydrogen.  
c( )
d) does not decompose as easily as hydrogen.  
d( )

31. Powdered graphite is frequently used as a lubricator because 

a) it becomes liquid easily as heat is produced by friction.  
a( )
b) its atoms are arranged in flat crystals which can slide 
easily.  
b( X )
c) its atoms are spherical and behave like tiny ball-bearings.  
c( )
d) its atoms break down under pressure to form an oily paste.  
d( )

32. When a ball swings back and forth on the 
end of a string, it has kinetic energy due 
to its motion, and also potential energy 
due to its height above its lowest point. However, when it 
stops swinging, it has neither K.E. nor P.E.

Why does this not contradict the Law of Conservation of Energy? 

Because, according to this Law, 

a) although energy cannot be produced out of nothing, it 
can be destroyed.  
a( )
b) energy cannot be created or destroyed.  
b( )
c) energy cannot be created, but may be transformed or 
destroyed.  
c( )
d) energy can be changed into other forms.  
d( X )

33. Each molecule of the compound whose formula is Na₃PO₄ contains 

a) 7 atoms  
b) 8 atoms  
c) 9 atoms  
d) 14 atoms  

34. When an unknown gas is bubbled through clear limewater, the 
lime water becomes milky.

The gas can be assumed to be carbon dioxide, provided 

a) the gas does not react chemically with the limewater.  
a( )
b) no other substance turns milky with CO₂.  
b( )
c) no other gas turns limewater milky.  
c( X )
d) all of the last three statements are true.  
d( )
35. A roll of steel wool was dipped in a liquid which was supposed to prevent oxidation from taking place. It was then sealed in a large container filled only with moist air (a little moisture being necessary for rusting to occur). The container was weighed immediately, and again after several weeks, still unsealed. No change of weight had occurred.

What conclusion could be drawn?

a) Since there was no increase in weight, oxidation could not have taken place.
b) Since rusted steel is lighter than pure steel, oxidation could not have taken place.
c) Since the total weight at the end of a chemical change equals that at the beginning, oxidation took place.
d) There is not enough evidence to tell if oxidation occurred.

36. Below is a set of equations.

The one which best displays the meaning of the term "radical" is

a) \( \text{AgNO}_3 + \text{KCl} \rightarrow \text{KNO}_3 + \text{AgCl} \)
b) \( \text{Ca} + \text{Cl}_2 \rightarrow \text{CaCl}_2 \)
c) \( \text{Ca} + \text{MgCl}_2 \rightarrow \text{CaCl}_2 + \text{Mg} \)
d) \( \text{H}_2\text{SO}_3 \rightarrow \text{H}_2\text{O} + \text{SO}_2 \)

37. Which of the following formulae does NOT represent an actual chemical substance?

a) \( \text{Fe} \)  b) \( \text{H} \)  c) \( \text{C} \)  d) \( \text{Na} \)
38. When a positively charged glass rod is held near a neutral pith-ball, the ball will be attracted to the rod because
   a) a magnetic field of force surrounds the rod and influences the ball.  
   b) electrons are transferred from the ball to the rod.  
   c) electrons are transferred from the rod to the ball.  
   d) electrons are moved over the surface of the ball towards the rod.

39. Of the following formulae, which one represents a substance whose water-solution is likely to feel slippery and to change the colour of red litmus paper?
   a) HClO₃  b) LiCl  c) HI  d) LiOH

40. A transformer may be used to
   a) convert an alternating current to a larger current at a higher voltage.  
   b) transform alternating current into direct current.  
   c) increase an alternating current and decrease its voltage.  
   d) step up the voltage of a battery.
THE DIAGRAMS BELOW REFER TO AN ITEM APPEARING IN THE TEST

1. Charged rods

2. Wire carrying current from dry-cell

3. Compass-needle

4. Pith ball on insulating stand

CHECK HERE if you had time to consider all items.
41. No one has yet been able to make cobalt by joining two
or more simpler substances, or to break it down into any
other substances, using chemical means.

This suggests that cobalt is

a) a metal  c) an element
b) a non-metal  d) a compound

42. (This item refers to the diagrams of the figure on the front page.)

Which of the following pairs of these devices could be used
in repeating Oersted's famous experiment, establishing that
a wire carrying an electric current has a field of force?

a) 1,3  b) 2,3  c) 3,4  d) 3,5

43. Electrical currents produced by electrostatic processes are
of limited value in our civilization because

a) they are brief and not properly controllable.

b) they are weaker than those produced by regular methods.

c) they are an essentially different kind of
electricity from that produced by regular methods.

d) they flow from negative to positive instead of
from positive to negative.
44. According to the definition of "work", which of the following represents a situation in which work is being done?

a) a wooden prop, 4 feet long, holds up a corner of a building weighing 1½ tons.
b) a man pushes against a lawnmower, stuck against a stone, with a force of 40 lb.
c) a piece of rock is thrown 100 feet into the air by a volcano.
d) a mountaineer is supporting his 180-lb. comrade who, having lost his foothold, is dangling from the other end of a 12-ft. rope.

45. "Reduction" is the name given to a chemical process which is just the reverse of oxidation. Select, from the following equation, the formula of the substance which is being reduced:

\[ \text{EQUATION: } \text{C} + \text{H}_2\text{O} \rightarrow \text{CO} + \text{H}_2 \]

The substance which is being reduced is

a) C   b) H\hspace{0.2cm}_2\hspace{0.2cm}O   c) CO   d) H\hspace{0.2cm}_2

46. If you wished to prepare a sample of carbon dioxide, you could do it by putting some pieces of calcium carbonate in a solution of

a) acid   b) base   c) salt   d) limewater

47. To distinguish with certainty and ease between aluminum and zinc, which of the following four properties should be investigated?

a) density   c) combustibility
b) state   d) lustre
48. A condition which is necessary for magnetic induction to take place is that the material to be magnetized be

a) placed within a magnetic field of force.  
b) tapped with a hammer while held in a north-south direction.  
c) stroked with a magnet, always in the same direction.  
d) placed within a coil of wire through which an electric current is flowing.

49. In Europe in the 15th Century, it was known that the needle of a compass did not point to True North, but somewhat to the west of True North. One of the early voyages of Columbus nearly ended in failure when his men discovered that the compass was pointing east of True North.

This may be considered evidence that

a) even in the 15th Century, the north magnetic pole was slowly changing its location.  
b) in those days, it was not known that the basic principles of magnetism are reversed in the southern hemisphere.  
c) the sailors of that time did not know that the earth's rotation changes the relative position of Magnetic and True North.  
d) the nature of magnetic variation was not understood in the 15th Century.

50. During the 16th Century, Gilbert repeatedly tried to produce an electrical charge upon materials like brass, but failed each time.

This was because brass is a

a) non-electrolyte  
b) non-magnetic material  
c) metallic insulator  
d) good conductor

51. The number of degrees between the freezing point and boiling point of water on a centigrade thermometer is

a) 212  b) 180  c) 100  d) 32
52. Recently, a fire started in a heap of fresh, damp sawdust, a material having low conductivity.

Of the following statements, which one provides the best explanation of this?

a) Fresh, damp sawdust "heats" easily, so it produced combustion.
b) Chemical combination produced energy that could not escape.
c) Damp sawdust has a higher kindling point than dry sawdust.
d) Damp sawdust has a lower kindling point than dry sawdust.

53. To give a positive charge to an uncharged conductor, we must

a) rub it with fur. c) remove some electrons.
b) rub it with silk. d) add some electrons.

54. The element which is present in the earth's crust in greatest quantity is

a) hydrogen b) oxygen c) nitrogen d) silicon

55. A method of altering the molecules of liquid oils so as to produce solid fats is known as

a) chlorination c) hydrogenation
b) oxidation d) fluoridation

56. Which of the following best illustrates "charging by conduction"?

a) Uncharged body comes in contact with charged body and acquires same charge as that of charged body.
b) Uncharged body comes in contact with charged body and acquires charge opposite to that of charged body.
c) Uncharged body approaches charged body, is grounded, and acquires same charge as that of charged body.
d) Uncharged body approaches charged body, is grounded and acquires charge opposite to that of charged body.
57. When an object has lost or gained electrons in any way, it is
in an unbalanced condition electrically.
Which of the following does NOT illustrate this statement?

a) A windowpane, newly cleaned with a silk cloth.
b) A thundercloud.
c) A lightning-rod.
d) The zinc plate of a voltaic cell, in operation.

58. One would expect to find a field of force around

a) A wire carrying an electric current.
b) A negatively charged insulator.
c) A planet in the solar system.
d) All of these.

59. For an electric cell to work, which of the following is either
unnecessary or incorrect?

a) Different materials must be used for the two plates.
b) One electrode must react with the electrolyte.
c) The liquid must be an acid, base or salt solution.
d) The electrodes must both be metals.

60. No reaction occurred when some dry chlorine gas and some small
pieces of iron were placed in a completely empty container.
However, when a tiny drop of water was introduced into the
container, the iron and chlorine reacted with great vigour.
The new substance formed showed no trace of either hydrogen
or oxygen, after it had been dried.
The part played by water in this reaction most resembles

a) a spontaneous combustion b) a catalysis c) combination
d) replacement.
61. A certain compound of iodine is added to table salt in small quantities
   a) to prevent people from acquiring a disease.  
   b) to act as a preservative and prevent decay.  
   c) to prevent the salt from absorbing moisture and thus caking.  
   d) to control the rate at which salt is absorbed into the blood.

62. Which of the following best illustrates the transformation of kinetic to potential energy?
   a) The starting of an automobile engine.  
   b) The use of coal to produce steam under pressure.  
   c) The charging of an electric battery by a windmill.  
   d) The starting of a rock-slide by a falling boulder.

63. The diagram shows a circuit containing a fuse (F) and an appliance whose electrical resistance is R. If an experimenter now connects points A and B by a piece of thick, copper wire (indicated by the dotted line), the fuse will "blow". Which of the following principles help(s) to explain this?

PRINCIPLES: 1. More current flows when resistance is greater.  
            2. More current flows when resistance is lower.  
            3. More heat is produced when resistance is greater.  
            4. More heat is produced when current is greater.

   a) 2 alone  b) 3 alone  c) 1 and 4  d) 2 and 4

64. If it were possible to divide a grain of glucose (C\textsubscript{6}H\textsubscript{12}O\textsubscript{6}) into smaller and smaller particles, the smallest particle which would still be glucose would be

   a) an element  b) an electron  c) a molecule  d) an atom.
65. Which one of the following household materials is a base?  
   a) lye  
   b) turpentine  
   c) vinegar  
   d) common salt  

66. Absorption (i.e., taking in) of energy is characteristic of which one of the following kinds of reaction?  
   a) Combustion  
   b) Slaking (of lime)  
   c) Electrolysis  
   d) Oxidation  

67. Which of the following would be the best substance to use to prevent loss of electrical energy through leakage of electrons?  
   a) silver  
   b) sulphur  
   c) lead  
   d) "green" wood.  

68. Lithium, which is a metallic element, is never found in its metallic form in the earth as are some metals - for example, gold; it exists only in compounds. Which of the following true statements best explains this fact?  
   a) Lithium is extremely active, chemically.  
   b) Lithium forms compounds with certain elements.  
   c) Compounds of lithium often occur in large deposits.  
   d) Lithium may be manufactured by extracting it from one of its compounds.  

69. The filament of an electric light bulb is a long piece of tungsten wound into a very fine coil. The purpose of coiling it is to  
   a) create a magnetic field which increases resistance and hence heat and light produced.  
   b) avoid allowing the hot tungsten to come into contact with the glass bulb and thus crack it.  
   c) concentrate the tungsten in a small space to reduce oxidation of the filament.  
   d) enable heat produced by each turn of the coil to be usefully employed in heating neighbouring ones.
70. The chemical energy of coal, and the potential energy of water stored behind a dam, are end-products of a series of energy transformations. Of the kinds and forms of energy listed below, the one which formed the beginning of both of these series is

a) kinetic  b) radiant  c) mechanical  d) synthetic

71. Insulated wire A, carrying an alternating current, was placed close to insulated coil B. A student claimed that an alternating current then flowed in coil B.

A reasonable explanation is that

a) the student must have made an error since every electrical circuit requires a source of current as well as a pathway and an appliance.

b) the student must have made an error since coil B was not in motion and thus could not be cutting lines of force.

c) free electrons from wire A are transferred to coil B by electrostatic induction, and continue their alternating motion.

d) A's alternating current produces a variable magnetic field whose lines of force cut coil B in one direction, then another.

72. Power losses during transmission of electrical energy are lowest when the current is

a) A.C. with low amperage and high voltage.  
b) D.C. with low amperage and high voltage.  
c) A.C. with high amperage and low voltage.  
d) D.C. with high amperage and low voltage.
73. Consider the equations: \[ \text{C} + \text{O}_2 \rightarrow \text{CO}_2 \]
\[ \text{CO}_2 + \text{H}_2\text{O} \rightarrow \text{H}_2\text{CO}_3 \]

Decide which is the most likely result of holding a piece of litmus paper in the fumes of a barbecue-pit (using charcoal or carbon for fuel.)

a) Moist blue litmus paper will turn red.
b) Moist red litmus paper will turn blue.
c) Dry blue litmus paper will turn red.
d) Dry red litmus paper will turn blue.

74. Which of the following represents an actual compound?

a) \text{CO}_3 \hspace{1cm} b) \text{NO}_3 \hspace{1cm} c) \text{NH}_3 \hspace{1cm} d) \text{SO}_4

75. Which of the following observations refers to a physical, rather than a chemical property?

a) Burns rapidly in oxygen, above 80° C.
b) Boils at 68°C., at standard pressure.
c) Imparts a milky appearance to lime-water.
d) Releases bubbles when some metal is added.

76. Newspapers sometimes carry reports of the use by scientists of Uranium-235 and Uranium-238. These are samples of the element uranium, whose atoms weigh either 235 units or 238 units.

Such reports appear to be in conflict with

a) the Electron Theory of matter.
b) Dalton's Atomic Theory.
c) the Law of Definite Proportions.
d) the Molecular Theory of matter.

77. Which of the following formulae represents a salt?

a) \text{Mg(OH)}_2 \hspace{1cm} b) \text{NH}_3 \hspace{1cm} c) \text{H}_2\text{S} \hspace{1cm} d) \text{FeSO}_4
78. Which of the following equations illustrates the type of chemical reaction known as "replacement"?

a) $\text{NH}_3 + 2\text{O}_2 \rightarrow \text{HNO}_3 + \text{H}_2\text{O}$

b) $\text{Ca} + \text{SnCl}_2 \rightarrow \text{CaCl}_2 + \text{Sn}$

c) $\text{Ca} + \text{Cl}_2 \rightarrow \text{CaCl}_2$

d) $\text{CaCO}_3 \rightarrow \text{CaO} + \text{CO}_2$

79. The combination of hydrogen (H) from one compound with hydroxyl (OH) from another compound, to produce water, is characteristic of all chemical reactions of a particular kind.

Which kind?

a) Replacement  

b) Combination  

c) Decomposition  

d) Neutralization

80. Which of the following compounds is manufactured from one or more of the others?

a) sodium carbonate  

b) calcium carbonate  

c) sodium chloride  

d) cellulose
### TABLE XI

**TAXONOMY CATEGORY, DIFFICULTY INDEX, AND DISCRIMINATION INDICES (POINT BISERIAL AND BISERIAL) FOR ALL ITEMS**

<table>
<thead>
<tr>
<th>Item no.</th>
<th>Taxonomy-no.</th>
<th>Difficulty index (%) (final)</th>
<th>$r_{pb}$</th>
<th>$r_{bis}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.10 (C)</td>
<td>80.0</td>
<td>.335</td>
<td>.179</td>
</tr>
<tr>
<td>2</td>
<td>1.12 (K)</td>
<td>71.5</td>
<td>.253</td>
<td>.336</td>
</tr>
<tr>
<td>3</td>
<td>2.10 (C)</td>
<td>64.3</td>
<td>.125</td>
<td>.516</td>
</tr>
<tr>
<td>4</td>
<td>1.11 (K)</td>
<td>72.6</td>
<td>.193</td>
<td>.258</td>
</tr>
<tr>
<td>5</td>
<td>2.10 (C)</td>
<td>54.7</td>
<td>.281</td>
<td>.356</td>
</tr>
<tr>
<td>6</td>
<td>1.11 (K)</td>
<td>53.2</td>
<td>.169</td>
<td>.213</td>
</tr>
<tr>
<td>7</td>
<td>2.20 (C)</td>
<td>64.3</td>
<td>.256</td>
<td>.329</td>
</tr>
<tr>
<td>8</td>
<td>1.31 (K)</td>
<td>53.2</td>
<td>.185</td>
<td>.232</td>
</tr>
<tr>
<td>9</td>
<td>2.10 (C)</td>
<td>64.7</td>
<td>.302</td>
<td>.390</td>
</tr>
<tr>
<td>10</td>
<td>1.31 (K)</td>
<td>77.4</td>
<td>.295</td>
<td>.111</td>
</tr>
<tr>
<td>11</td>
<td>1.32 (K)</td>
<td>47.6</td>
<td>.051</td>
<td>.065</td>
</tr>
<tr>
<td>12</td>
<td>2.10 (C)</td>
<td>31.5</td>
<td>.197</td>
<td>.258</td>
</tr>
<tr>
<td>13</td>
<td>1.31 (K)</td>
<td>64.3</td>
<td>.327</td>
<td>.120</td>
</tr>
<tr>
<td>14</td>
<td>1.11 (K)</td>
<td>65.5</td>
<td>.328</td>
<td>.124</td>
</tr>
<tr>
<td>15</td>
<td>2.10 (C)</td>
<td>65.3</td>
<td>.324</td>
<td>.118</td>
</tr>
<tr>
<td>16</td>
<td>2.20 (C)</td>
<td>60.4</td>
<td>.320</td>
<td>.106</td>
</tr>
<tr>
<td>17</td>
<td>1.12 (K)</td>
<td>32.3</td>
<td>.261</td>
<td>.341</td>
</tr>
<tr>
<td>18</td>
<td>2.10 (C)</td>
<td>41.9</td>
<td>.345</td>
<td>.137</td>
</tr>
<tr>
<td>19</td>
<td>2.30 (C)</td>
<td>43.6</td>
<td>.178</td>
<td>.225</td>
</tr>
<tr>
<td>20</td>
<td>1.25 (K)</td>
<td>55.3</td>
<td>.316</td>
<td>.397</td>
</tr>
<tr>
<td>21</td>
<td>2.20 (C)</td>
<td>40.2</td>
<td>.208</td>
<td>.263</td>
</tr>
<tr>
<td>22</td>
<td>1.31 (K)</td>
<td>50.6</td>
<td>.334</td>
<td>.118</td>
</tr>
<tr>
<td>23</td>
<td>2.20 (C)</td>
<td>42.3</td>
<td>.443</td>
<td>.560</td>
</tr>
<tr>
<td>24</td>
<td>1.31 (K)</td>
<td>42.3</td>
<td>.281</td>
<td>.356</td>
</tr>
<tr>
<td>25</td>
<td>1.31 (K)</td>
<td>40.4</td>
<td>.364</td>
<td>.162</td>
</tr>
<tr>
<td>26</td>
<td>1.31 (K)</td>
<td>32.6</td>
<td>.119</td>
<td>.156</td>
</tr>
<tr>
<td>27</td>
<td>2.20 (C)</td>
<td>58.1</td>
<td>.143</td>
<td>.522</td>
</tr>
<tr>
<td>28</td>
<td>1.12 (K)</td>
<td>34.2</td>
<td>.123</td>
<td>.159</td>
</tr>
<tr>
<td>29</td>
<td>1.12 (K)</td>
<td>43.0</td>
<td>.151</td>
<td>.190</td>
</tr>
<tr>
<td>30</td>
<td>2.10 (C)</td>
<td>53.6</td>
<td>.260</td>
<td>.328</td>
</tr>
<tr>
<td>31</td>
<td>1.31 (K)</td>
<td>33.4</td>
<td>.352</td>
<td>.157</td>
</tr>
<tr>
<td>32</td>
<td>2.10 (C)</td>
<td>39.1</td>
<td>.230</td>
<td>.293</td>
</tr>
<tr>
<td>33</td>
<td>2.10 (C)</td>
<td>51.1</td>
<td>.357</td>
<td>.147</td>
</tr>
<tr>
<td>34</td>
<td>2.20 (C)</td>
<td>41.5</td>
<td>.300</td>
<td>.379</td>
</tr>
<tr>
<td>35</td>
<td>2.20 (C)</td>
<td>39.8</td>
<td>.310</td>
<td>.393</td>
</tr>
<tr>
<td>36</td>
<td>2.20 (C)</td>
<td>22.8</td>
<td>.221</td>
<td>.306</td>
</tr>
<tr>
<td>37</td>
<td>1.11 (K)</td>
<td>28.1</td>
<td>.126</td>
<td>.168</td>
</tr>
<tr>
<td>38</td>
<td>1.31 (K)</td>
<td>39.4</td>
<td>.300</td>
<td>.381</td>
</tr>
<tr>
<td>39</td>
<td>2.10 (C)</td>
<td>32.8</td>
<td>.398</td>
<td>.515</td>
</tr>
<tr>
<td>40</td>
<td>1.31 (K)</td>
<td>24.7</td>
<td>-.049</td>
<td>-.066</td>
</tr>
<tr>
<td>Item no.</td>
<td>Taxonomy-category</td>
<td>Difficulty index (%)</td>
<td>$r_{pb}$</td>
<td>$r_{bis}$</td>
</tr>
<tr>
<td>---------</td>
<td>-------------------</td>
<td>----------------------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>41</td>
<td>1.11 (K)</td>
<td>93.2</td>
<td>.288</td>
<td>.555</td>
</tr>
<tr>
<td>42</td>
<td>2.10 (C)</td>
<td>80.0</td>
<td>.289</td>
<td>.113</td>
</tr>
<tr>
<td>43</td>
<td>1.31 (K)</td>
<td>66.0</td>
<td>.392</td>
<td>.507</td>
</tr>
<tr>
<td>44</td>
<td>2.10 (C)</td>
<td>54.5</td>
<td>.146</td>
<td>.522</td>
</tr>
<tr>
<td>45</td>
<td>2.20 (C)</td>
<td>74.5</td>
<td>.203</td>
<td>.276</td>
</tr>
<tr>
<td>46</td>
<td>1.31 (K)</td>
<td>49.8</td>
<td>.118</td>
<td>.186</td>
</tr>
<tr>
<td>47</td>
<td>2.10 (C)</td>
<td>61.9</td>
<td>.331</td>
<td>.126</td>
</tr>
<tr>
<td>48</td>
<td>2.10 (C)</td>
<td>48.3</td>
<td>.350</td>
<td>.140</td>
</tr>
<tr>
<td>49</td>
<td>2.10 (C)</td>
<td>59.8</td>
<td>.174</td>
<td>.221</td>
</tr>
<tr>
<td>50</td>
<td>1.31 (K)</td>
<td>27.9</td>
<td>.279</td>
<td>.373</td>
</tr>
<tr>
<td>51</td>
<td>1.12 (K)</td>
<td>81.9</td>
<td>.276</td>
<td>.105</td>
</tr>
<tr>
<td>52</td>
<td>2.20 (C)</td>
<td>64.3</td>
<td>.183</td>
<td>.236</td>
</tr>
<tr>
<td>53</td>
<td>1.25 (K)</td>
<td>57.6</td>
<td>.127</td>
<td>.538</td>
</tr>
<tr>
<td>54</td>
<td>1.11 (K)</td>
<td>61.7</td>
<td>.130</td>
<td>.167</td>
</tr>
<tr>
<td>55</td>
<td>1.11 (K)</td>
<td>72.3</td>
<td>.317</td>
<td>.124</td>
</tr>
<tr>
<td>56</td>
<td>2.10 (C)</td>
<td>46.0</td>
<td>.258</td>
<td>.324</td>
</tr>
<tr>
<td>57</td>
<td>2.10 (C)</td>
<td>31.1</td>
<td>.315</td>
<td>.152</td>
</tr>
<tr>
<td>58</td>
<td>1.31 (K)</td>
<td>42.1</td>
<td>.120</td>
<td>.152</td>
</tr>
<tr>
<td>59</td>
<td>1.31 (K)</td>
<td>31.9</td>
<td>.237</td>
<td>.309</td>
</tr>
<tr>
<td>60</td>
<td>2.20 (C)</td>
<td>56.4</td>
<td>.231</td>
<td>.292</td>
</tr>
<tr>
<td>61</td>
<td>1.31 (K)</td>
<td>53.2</td>
<td>.318</td>
<td>.135</td>
</tr>
<tr>
<td>62</td>
<td>2.10 (C)</td>
<td>33.4</td>
<td>.342</td>
<td>.144</td>
</tr>
<tr>
<td>63</td>
<td>2.20 (C)</td>
<td>40.6</td>
<td>.306</td>
<td>.389</td>
</tr>
<tr>
<td>64</td>
<td>1.11 (K)</td>
<td>61.3</td>
<td>.398</td>
<td>.506</td>
</tr>
<tr>
<td>65</td>
<td>1.11 (K)</td>
<td>51.1</td>
<td>.394</td>
<td>.493</td>
</tr>
<tr>
<td>66</td>
<td>2.10 (C)</td>
<td>25.9</td>
<td>.025</td>
<td>.034</td>
</tr>
<tr>
<td>67</td>
<td>1.25 (K)</td>
<td>31.5</td>
<td>.314</td>
<td>.110</td>
</tr>
<tr>
<td>68</td>
<td>2.20 (C)</td>
<td>50.8</td>
<td>.132</td>
<td>.543</td>
</tr>
<tr>
<td>69</td>
<td>1.31 (K)</td>
<td>27.9</td>
<td>.216</td>
<td>.329</td>
</tr>
<tr>
<td>70</td>
<td>1.31 (K)</td>
<td>27.2</td>
<td>.280</td>
<td>.375</td>
</tr>
<tr>
<td>71</td>
<td>2.20 (C)</td>
<td>35.9</td>
<td>.168</td>
<td>.215</td>
</tr>
<tr>
<td>72</td>
<td>1.31 (K)</td>
<td>20.4</td>
<td>.184</td>
<td>.261</td>
</tr>
<tr>
<td>73</td>
<td>2.20 (C)</td>
<td>24.0</td>
<td>.134</td>
<td>.184</td>
</tr>
<tr>
<td>74</td>
<td>1.23 (K)</td>
<td>29.1</td>
<td>.216</td>
<td>.286</td>
</tr>
<tr>
<td>75</td>
<td>1.23 (K)</td>
<td>38.5</td>
<td>.259</td>
<td>.331</td>
</tr>
<tr>
<td>76</td>
<td>2.20 (C)</td>
<td>27.6</td>
<td>.173</td>
<td>.231</td>
</tr>
<tr>
<td>77</td>
<td>2.10 (C)</td>
<td>41.5</td>
<td>.322</td>
<td>.407</td>
</tr>
<tr>
<td>78</td>
<td>2.10 (C)</td>
<td>43.0</td>
<td>.238</td>
<td>.301</td>
</tr>
<tr>
<td>79</td>
<td>2.10 (C)</td>
<td>33.4</td>
<td>.182</td>
<td>.236</td>
</tr>
<tr>
<td>80</td>
<td>1.22 (K)</td>
<td>22.6</td>
<td>.259</td>
<td>.361</td>
</tr>
</tbody>
</table>