DETERMINATION OF THE NUMBER OF ITEMS IN A MULTIPLE CHOICE TEST USING SUBSET SELECTION TECHNIQUES

BY

JEAN D. GIBBONS, INGRAM OLKIN, and MILTON SOBEL

TECHNICAL REPORT NO. 134
OCTOBER 1978

PREPARED UNDER THE AUSPICES OF
NATIONAL SCIENCE FOUNDATION GRANT
MPS 75-09450

Ingram Olkin, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
DETERMINATION OF THE NUMBER OF ITEMS IN A MULTIPLE CHOICE TEST
USING SUBSET SELECTION TECHNIQUES

by

Jean D. Gibbons, Ingram Olkin, and Milton Sobel

TECHNICAL REPORT NO. 134

OCTOBER 1978

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT
MPS 75-09450
Ingram Olkin, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
1. Introduction

Consider a multiple choice test consisting of $n$ questions, where each question has exactly $k$ alternatives (possible responses) and only one alternative is the correct answer (the other alternatives are called distractors). Many models have been suggested for the methods of response and scoring of such tests.

The following survey of possible response methods is based on de Finetti (1972) and encompasses most of the methods to date. The procedures in category A require a fixed number of responses, whereas those in B permit a flexible number of responses.

A. Fixed number of responses

A-1. Order completely the $k$ alternatives.

A-2. Order only the best $m$ alternatives, $m \leq k - 2$ where $m$ is a fixed number (the traditional method if $m = 1$).

A-3. Select $m$ alternatives and cross out $m'$, $m + m' < k$ where both $m$ and $m'$ are fixed numbers.
B. Flexible number of responses

B-1. Select one alternative or none.

B-2. Cross out any number of alternatives, from none to all.

B-3. Select one alternative and cross out any number.

B-4. Select one alternative or none, and cross out any number.

B-5. Select (with double or single emphasis) one alternative or none and cross out any number.

B-6. Select and cross out freely.

Another general type of response method requires the specification of certain distances. These are discussed in some detail in de Finetti (1972).

For any designated response method, there must be a corresponding scoring procedure. Willey (1950) suggested and analyzed a scoring procedure for response method A-3 with \( m = 1 \), \( m' = 2 \), and \( k = 5 \) alternatives. The flexible response methods generally require more complicated scoring procedures. Coombs (1953) suggested a scoring procedure for method B-6 based on \( k = 4 \) alternatives. This same procedure was tested experimentally in Coombs, Milholland, and Womer (1956) and compared with the traditional method (A-2 with \( m = 1 \)) and with a variant of method B-3. Dressel and Schmid (1953) proposed several response and corresponding scoring methods, including a scoring formula for method B-2 (with \( k = 5 \) alternatives) which they termed a "free-choice" test; they also reported data from an experiment designed to evaluate student performance on these methods as well as on the traditional method (A-2 with \( m = 1 \)).
It should be noted that methods B-2 and B-6 allow equivalent sets of responses, and that in each case the subject is permitted to cross out all answers as well as to select all answers and hence not to select any answer as correct. De Finetti (1972, pp. 35-56) claims that crossing out none (and therefore selecting all) is not equivalent to crossing out all the alternatives because the subject can consider all answers plausible but not wrong when he is told that exactly one is correct. Further, it may be more difficult to recognize an answer as correct than to eliminate one as being wrong. Coombs (1953, p. 308) comments that, "It seems to be a common experience of individuals taking objective tests to feel confident about eliminating some of the wrong alternatives and then guess from among the remaining ones." Thus there may be some psychological difference between the two sets of instructions even though they are mathematically equivalent.

The previous literature has investigated the performance and reliability of some of these response and scoring procedures by administering such tests with a predetermined number of questions to experimental subjects. An aspect of multiple choice tests that has not previously been discussed is the determination of how many questions should be on the test. If one response method provides more information about the subject's knowledge than another, then that method should require fewer questions to elicit the same total amount of information. The purpose of using response methods other than the traditional one is to take into account any partial knowledge of the subject and the certainty of his knowledge and also to discourage guessing.
In order to investigate this problem we first place it in the context of subset selection. Subset selection methods are designed to partition a group of \( k \) populations (these may be applications for a job or admission to college, or a collection of drugs) into two groups in such a way that the experimenter has confidence that one group contains the \( t \) "best" populations. A partition (or selection) is correct if the selected subset contains the \( t \) best populations. Further, the method of partitioning should be such that there is some guaranteed probability of making a correct selection.

The original ideas of subset selection procedures are due primarily to Gupta (1956), to Gupta and Sobel (1960), and to Gupta, Huyett, and Sobel (1957). These principles are an extension of the development by Bechhofer (1954) on methods of ordering and selecting populations. For a general exposition of the methods of ordering and selecting populations, see Gibbons, Olkin, and Sobel (1977).

In the context of a multiple choice test, for each question the subject is instructed to choose a subset of the \( k \) given alternatives that he believes contains the one correct answer; the subset may be of any size (including 0 and \( k \)). (As a result, this response method is mathematically equivalent to both B-2 and B-6.) In this framework we can make use of the theory of selecting a subset containing the one best population, which here means the correct answer. We first derive the mean and variance of the score per question as a function of \( k \), the probability \( p \) of screening out a wrong answer and the probability \( p' \) of including the correct answer on any question. We further give methods of determining the total number \( n \) of questions that should be included on the test such that the average score on all questions can be regarded as a good measure of the subject's knowledge.
2. Application of Subset Selection to the Scoring of Tests

Consider a test consisting of \( n \) questions; each question has \( k \) possible answers, of which only one answer is correct. The subject is instructed to select a subset of answers that contains the unique correct answer and he can choose a subset of any size (including \( k \)). If the subset to be selected must contain only one element or choice, the test format is the traditional one. However, if the subset size is not restricted to one, what is to prevent the subject from including all \( k \) possible answers in his subset, so that the subset will always contain the correct answer? The problem is to devise a scoring procedure such that the score decreases with the size of the selected subset and is weighted in such a manner that any simple strategy of pure guessing has an expected score of zero. (Two simple strategies of guessing are: (i) select one or two answers at random, or (ii) toss a fair coin for each of the \( k \) choices to decide whether it should be included or not.)

To be more specific, suppose that \( k = 5 \) so that for each question there are five possible choices, only one of which is the correct answer. The subject is instructed to select any subset of the five possible choices that he thinks contains the one correct answer. The method of scoring is not shown to be optimal and it would be interesting to determine a criterion and theory for optimal scoring. Table 1 gives the set of weights that provides an expected score of zero on each question if the subject resorts to pure guessing, regardless of how the guessing is carried out. For the present case of exactly one correct answer, this scoring system is equivalent to the scoring method given by Coombs (1953), and tested experimentally by Dressel and Schmid (1953) and by Coombs, Milholland, and Womer (1956). Note that the extension of this scoring system to general \( k \) is immediate; the discussion starting in Section 3 is for general \( k \).
### Table 1

<table>
<thead>
<tr>
<th>Size of Subset</th>
<th>Score if Subset Includes the Correct Answer</th>
<th>Score if Subset Does Not Include the Correct Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>-2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>-3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>-4</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>

To illustrate the main ideas, suppose a subject guesses the answer using a strategy of selecting two choices at random. Then the probability that the two choices selected include the correct answer is $2/5$ and the probability they do not include the correct answer is $3/5$. Consequently, the expected score $S$ per question (using Table 1) is

$$E(S) = 3(2/5) - 2(3/5) = 0.$$  

It is readily checked that the same result holds if any number of choices is selected at random. If the subject tosses a fair coin for each of the five choices, the expectation is +2 that the correct answer is included and -2 that it is not included, and hence the net result is again zero. The same expectation results for any strategy based on pure guessing.

Note that the scoring method of Table 1 on a single question can be obtained by grading each of the five possible choices separately as
follows. If the one correct answer is included in the subset, four points are scored; if the correct answer is not included, zero points are scored. From the appropriate number four or zero, one point is subtracted for each incorrect choice that is also in the subset in order to obtain the net score for that question. Note that this scoring method permits substantial negative scores and, as noted above, gives an expected score of zero for pure guessing. These properties have a tendency to keep the subject honest; they also offer a subject with partial information a method of response in which partial credit can be obtained.

The property that the expected score is zero for pure guessing determines the negative scores given any set of fixed values for the positive scores. For example, if the positive score for a subset of size two in Table 1 were, say, $a$, then the negative score must be $-2a/3$ so that $(a)(2/5) - (2a/3)(3/5)$ is equal to 0. We have merely chosen the smallest constants that force the score always to be an integer.

This scoring method differs significantly from the system suggested by de Finetti (1972), in which the subject is required to provide his or her own a priori probability distribution of correctness on all the choices. In contrast, the present scoring appears to measure the ability of the subject to properly screen out the wrong answers and include the correct one without explicitly bringing in his prior (or posterior) probability of a choice being correct.

The main idea in our analysis below is that we are trying to treat the responses to the $k$ parts of a single item as $k$ bits of information, rather than as only one bit of information. As we shall see, this has some surprising results when we compare the efficiency of our procedure with that of the conventional one.
3. **Mean and Variance of Scores**

Now suppose that the subject is not necessarily guessing but has probability $p$ of screening out (i.e., not including) a wrong answer and probability $p'$ of including (i.e., not screening out) the correct answer on any question. This assumption requires some clarification. In any realistic context the probabilities $p$ and $p'$ will differ for each answer. The $p$ and $p'$ do not change from answer to answer in our analysis and hence should reflect an average over answers. It will later become evident that these average values provide a first approximation to the number of test questions needed.

With $k$ possible answers and the scoring method of Table 1, the expected score $S$ per question is

$$E(S) = (k-1)p'p - (k-1)(1-p')(1-p) = (k-1)(p + p' - 1)$$

out of a maximum of $k - 1$ points. Thus the expected relative score per question in percent is $100(p + p' - 1)$.

We now sketch the derivation of (1); the variance of $S$ is also derived below. If there are $k$ choices per question and exactly one is correct, $k - 1$ choices are incorrect. There are two kinds of choices, (a) the correct one and (b) the incorrect ones. For (a), the scoring method gives $k - 1$ points if the correct choice is kept and zero otherwise, so that the expected gain is $(k-1)p$. For (b), the scoring method simply takes away one point for each wrong choice included in the subset (and makes no change otherwise); thus the contribution to the expected gain is $-(k-1)(1-p')$, i.e., a loss of $(k-1)(1-p')$. The sum $(k-1)p - (k-1)(1-p')$ then yields (1).
To illustrate the use of (1), take $k = 5$ so that each question has five possible choices, and suppose that a subject has probability .9 of screening out a wrong choice and .8 of including the correct choice. Then the subject has an expected score of $4(.9 + .8 - 1) = 2.8$ per question out of a maximum of four points (or an expected score of 70%).

If $p = p' = 1$, then the expected score is 100%; if $p = p' = .5$, then the expected score is 0. The lowest possible score of -100% occurs when $p = p' = 0$. A simple transformation can be used to change the scores into positive values that range between 0 and 1, or 0 and 100%. For example, for any score $X$ that ranges between -1 and +1, the transformed value $Y = (X + 1)/2$ ranges between 0 and 1, and the expected score for $Y$ is $(p + p')/2$; if $X$ is a percentage that ranges between -100% and +100%, then $Y = [(X + 100)/2] %$ ranges between 0 and 100%.

In order to derive the mean $E(S)$ and variance $V(S)$ of the score $S$ when there are $k$ multiple choices per question and the scoring method is similar to that of Table 1, let $X_0, X_1, X_2, \ldots, X_{k-1}$ denote the $k$ Bernoulli random variables associated with the decisions made on the $k$ choices per question. Here $X_0$ denotes the zero-one random variable associated with the single correct choice (1 for including it and 0 for omitting it), and the other $(k-1)$ variables $X_1, \ldots, X_{k-1}$ are zero-one random variables associated with the $k-1$ remaining (wrong) choices (taking the value plus one if screened out and zero otherwise). Thus each $X$ is equal to 1 for a correct action on any single choice and 0 for an incorrect choice.
Various assumptions can be made concerning the dependence between the random variables $X_0, X_1, \ldots, X_{k-1}$, the simplest is that they are all independent. Such an assumption is unrealistic in that we expect the correlations between $X_0$ (the response to the correct alternative) and each of $X_1, \ldots, X_{k-1}$ (the responses to the distractors) to be positive and the correlations among the distractors to be positive. (Note that the correlations between $X_0$ and $1-X_i$, $i=1, \ldots, k-1$, are negative.) As a first approximation we assume that

$$\text{Corr} (X_0, X_i) = \tau, \quad i=1, \ldots, k-1,$$
$$\text{Corr} (X_i, X_j) = \rho, \quad i,j=1, \ldots, k-1.$$ 

The condition that the correlation matrix be positive definite requires that $[(k-1)e-1]/(k-2) < \rho < 1$. In general, $\tau$ will be negative and small relative to $\rho$. On any single trial, $X_0$ has probability $p'$ of success; each of $X_1, \ldots, X_{k-1}$ has probability $p$ of success. The score per question, $S$, can be written as

$$S = (k-1)X_0 - (1-X_1) - (1-X_2) - \ldots - (1-X_{k-1})$$
$$= (k-1)(X_0-1) + X_1 + \ldots + X_{k-1}. \quad (2)$$

Since the means and variances of the $X_i$'s are

$$E(X_0) = p', \quad E(X_i) = p, \quad i = 1, \ldots, k-1,$$
$$V(X_0) = p'(1-p'), \quad V(X_i) = p(1-p), \quad i = 1, \ldots, k-1,$$
we obtain

\[ E(S) = (k-1)(p'-1) + (k-1)p = (k-1)(p'+p-1). \]

\[ V(S) = (k-1)^2 p'(1-p') + (k-1)p(1-p) + 2(k-1)^2 \tau \sqrt{p'(1-p')p(1-p)} \]

\[ + (k-1)(k-2)p(1-p)p. \]

Note that \(-(k-1) \leq S \leq (k-1). Then \ Y = (S+k-1)/2(k-1) \ ranges between 0 and 1, and

\[ E(Y) = \frac{p'+p}{2}, \]

\[ V(Y) = \frac{p'(1-p')}{4} + \frac{p(1-p)}{4(k-1)} + \frac{\tau \sqrt{p'(1-p')p(1-p)}}{2} + \frac{(k-2)}{4(k-1)} p(1-p)p. \quad (3) \]

4. Determination of Number of Questions

Now consider a test of \ n \ multiple-choice questions, where each question has exactly \ k \ alternative choices. Let \ \bar{Y} \ be the total average score for all questions, where the scoring method for general \ k \ as in Table 1 is employed for each question and \ Y \ is the score per question defined as above to range between 0 and 1. Let \ p_0 = (p+p')/2 \ be a true measure of the subject's ability. We will determine how many questions should be on the test such that for given \ p_0 \ and for some specified \ \epsilon, \ 0 < \epsilon < 1, \ we satisfy the requirement that the probability that the total average score \ \bar{Y} \ lies between \ p_0 - \epsilon \ and \ p_0 + \epsilon \ is at least \ P^*(= .95 \ say). In other words, we seek the smallest number \ n \ of questions that is sufficiently large so that the average
score $\bar{Y}$ can be regarded as a true reflection of the subject's knowledge (as measured by the $p_0$ value).

To answer the question for any given $p_0$, we use the fact that for large $n$, $\bar{Y}$ can be approximated (in distribution) by the normal distribution. In symbols, we want to determine the value of $n$ such that

$$P(p_0 - \varepsilon < \bar{Y} < p_0 + \varepsilon) \geq p^*,$$  \hspace{1cm} (4)

or equivalently,

$$P\left(\frac{-\varepsilon}{\sqrt{V(\bar{Y})}} < \frac{\bar{Y} - E(\bar{Y})}{\sqrt{V(\bar{Y})}} < \frac{\varepsilon}{\sqrt{V(\bar{Y})}}\right) \geq p^*.$$  \hspace{1cm} (5)

Using the central limit theorem, the solution of (5) is given by

$$\phi\left(\frac{\sqrt{n}\frac{\varepsilon}{\sqrt{V(\bar{Y})}}}{\frac{1+p^*}{2}}\right) = \frac{1+p^*}{2},$$  \hspace{1cm} (6)

where $\phi$ is the cumulative standard normal distribution. The solution for $n$ from (6) is

$$n = V(\bar{Y}) \left[\phi^{-1}\left(\frac{1+p^*}{2}\right)/\varepsilon\right]^2,$$  \hspace{1cm} (7)

where $\phi^{-1}$ denotes the inverse standard cumulative distribution function.

Since $n$ must be an integer and we want a conservative result, we round the answer from (7) upward to the next integer.
To illustrate the calculation of \( n \), suppose that \( k = 5 \), \( p' = .9 \) and \( p = .7 \) so that \( p_0 = .8 \). For different values of \( \tau \) and \( \rho \), Table 2 gives values of \( V(Y) \) obtained from (3), and Table 3 gives values of \( n \) obtained from (7) for \( \varepsilon = .05 \), \( p^* = .95 \).

Table 2

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>( \tau )</th>
<th>( p_0 )</th>
<th>( .1 )</th>
<th>( .2 )</th>
<th>( .3 )</th>
<th>( .4 )</th>
<th>( .5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>.00</td>
<td>.035625</td>
<td>.042499</td>
<td>.049373</td>
<td>.056247</td>
<td>.063120</td>
<td>.069994</td>
<td></td>
</tr>
<tr>
<td>.25</td>
<td>.045469</td>
<td>.052343</td>
<td>.059216</td>
<td>.066090</td>
<td>.072964</td>
<td>.079838</td>
<td></td>
</tr>
<tr>
<td>.50</td>
<td>.055313</td>
<td>.062186</td>
<td>.069060</td>
<td>.075934</td>
<td>.082808</td>
<td>.089682</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>.075050</td>
<td>.081924</td>
<td>.088798</td>
<td>.095672</td>
<td>.102545</td>
<td>.109419</td>
<td></td>
</tr>
</tbody>
</table>

Table 3

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>( \tau )</th>
<th>( p_0 )</th>
<th>( .1 )</th>
<th>( .2 )</th>
<th>( .3 )</th>
<th>( .4 )</th>
<th>( .5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>.00</td>
<td>54.7</td>
<td>65.3</td>
<td>75.9</td>
<td>86.4</td>
<td>97.0</td>
<td>107.6</td>
<td></td>
</tr>
<tr>
<td>.25</td>
<td>69.9</td>
<td>80.4</td>
<td>91.0</td>
<td>101.6</td>
<td>112.1</td>
<td>122.7</td>
<td></td>
</tr>
<tr>
<td>.50</td>
<td>85.0</td>
<td>95.6</td>
<td>106.1</td>
<td>116.7</td>
<td>127.2</td>
<td>137.8</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>115.3</td>
<td>125.9</td>
<td>136.5</td>
<td>147.0</td>
<td>157.6</td>
<td>168.1</td>
<td></td>
</tr>
</tbody>
</table>

Thus, when \( \rho = \tau = 0 \) the multiple choice test should consist of at least 55 questions in order to satisfy the specified \((\varepsilon, p^*)\) requirement.
For a fixed $V(Y) = .035625$ and $n = 54.7$, Table 4 gives the $(\varepsilon, p^*)$ pairs that satisfy (6). When these points are plotted on a graph and connected by a smooth curve, we obtain the operating characteristic curve.

Table 4

<table>
<thead>
<tr>
<th>$\varepsilon$</th>
<th>.026</th>
<th>.029</th>
<th>.033</th>
<th>.037</th>
<th>.042</th>
<th>.050</th>
<th>.066</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p^*$</td>
<td>.70</td>
<td>.75</td>
<td>.80</td>
<td>.85</td>
<td>.90</td>
<td>.95</td>
<td>.99</td>
</tr>
</tbody>
</table>

For $p_0 = .8$ and the number of questions is fixed at $n = 55$, we can guarantee that $P\{|\bar{Y} - .050| < .8\} > .95$, or that $P\{|\bar{Y} - .037| < .8\} > .85$, or that $P\{|\bar{Y} - .026| < .8\} > .70$. Thus, by imposing a more stringent requirement that $\bar{Y}$ be closer to the subject's $p_0$ value, the corresponding level of confidence must decrease.

Remark 1. Since each score per question $S$ is the sum of five binomial variables, and there are 55 questions, computing $\bar{Y}$ is equivalent under our model to adding 275 independent variables. For such a large number of variables, the normal approximation to the sum of binomially distributed variables gives a very good approximation. If $\varepsilon$ were increased to $.1$ say, then (5) becomes

$$P\left(-.529\sqrt{n} < \frac{\bar{Y} - E(\bar{Y})}{\sqrt{V(\bar{Y})}} < .529\sqrt{n}\right) > .95,$$
in which case \( n = 13.7 \) when \( \tau = \rho = 0 \). Then we need a test with at least 14 questions. Even here, we are adding \( 5(14) = 70 \) binomial variates and the normal approximation yields accurate results.

**Remark 2.** In the previous development we started with a \( p_0 \) value and required a probability of at least \( P^* \) that the \( \bar{Y} \) score lie between \( p_0 - \varepsilon \) and \( p_0 + \varepsilon \). The result depended on the given \( p_0 \). We may also determine a minimum value of \( n \) that holds regardless of \( p_0 \). Obviously, the \( n \) value obtained will be conservative, i.e., larger than what would result for a specified \( p_0 \) value. This type of overall bound on \( n \) can be found by making \( V(\bar{Y}) \) as large as possible. The maximum value of \( V(Y) \) in (3) (when \( \tau = \rho = 0 \)) is achieved for \( p = p' = 1/2 \), in which case \( V(\bar{Y}) = k/[16(k-1)n] \). Hence the upper bound on \( n \) is the solution of

\[
\Phi(\varepsilon\sqrt{16(k-1)n/k}) = (1+P^*)/2.
\]

For example, with \( k = 5 \), \( P^* = .95 \), \( \varepsilon = .05 \), we obtain \( n = 120.0 \), which could be compared (say) with the value \( n = 54.7 \) which resulted for the given value \( p_0 = .8 \).

5. **Comparisons of efficiency**

We now compare these results of Section 4 with those for the conventional response method (A-2 with \( m = 1 \)) in which only one alternative can be selected for each question. Let the score per question be \( S_1 = k - 1 \) if correct and \(-1\) if incorrect. The quantity \( Y = (S_1 + 1)/k \) transforms the score to the interval \([0,1]\), and we are interested in the average \( \bar{Y} \) of the \( n \) observed values. Here we use \( p_0 \) to denote the probability that the
subject answers any one question correctly. Comparisons are made by identifying the parameter \( p_0 \) in both models (subset selection and conventional) as the measure of ability we are trying to estimate. Note that \( p \) and \( p' \) do not enter into this formulation. In this case, the means and variances are

\[
E(S_1) = kp_0 - 1 \quad \text{and} \quad E(Y) = p_0 = E(\bar{Y}),
\]

\[
V(S_1) = k^2 p_0 (1-p_0), \quad V(Y) = p_0 (1-p_0), \quad V(\bar{Y}) = p_0 (1-p_0)/n.
\]

Then the problem is to find the smallest integer \( n \) such that

\[
P \left\{ \frac{\varepsilon \sqrt{n}}{\sqrt{p_0 (1-p_0)}} < \frac{(\bar{Y}-p_0)\sqrt{n}}{\sqrt{p_0 (1-p_0)}} < \frac{\varepsilon \sqrt{n}}{\sqrt{p_0 (1-p_0)}} \right\} \geq p^*.
\]

The required value of \( n \) for the conventional method, denoted \( n_c \), is

\[
n_c = p_0 (1 - p_0) \left[ \Phi^{-1} \left( \frac{1-p^*}{2} / \varepsilon \right) \right]^2,
\]

(8)

and the value of \( n \) (denoted \( n_s \)) using subset selection procedures is given by (7). As a measure of the relative efficiency of the two procedures we obtain

\[
e_{s,c} = \text{eff (subset selection with respect to conventional)} = \frac{n_c}{n_s} = \frac{p_0 (1-p_0)}{V(Y)},
\]

(9)
where $V(Y)$ is given by (3). Note that this efficiency is independent of $\varepsilon$ and $P^*$. When $\rho = \tau = 0$, the efficiency is

$$e_{s,c} = \frac{p_0(1-p_0)}{\frac{p'(1-p')}{4} + \frac{p(1-p)}{4(1-k-1)}} = \frac{4p_0(1-p_0)}{(p_0+\Delta)(1-p_0-\Delta) + \frac{(p_0-\Delta)(1-p_0+\Delta)}{k-1}}, \quad (10)$$

where $\Delta = \frac{1}{2}(p'-p)$. Table 5 gives some values of (10) for selected values of $p_0$ and $\Delta$.

<table>
<thead>
<tr>
<th>$p_0$</th>
<th>$\Delta$</th>
<th>Number Needed Conventional Procedure</th>
<th>Number Needed Subset Selection Procedure</th>
<th>Eff</th>
</tr>
</thead>
<tbody>
<tr>
<td>.50</td>
<td>0</td>
<td>384.2</td>
<td>120.0</td>
<td>3.20</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>384.2</td>
<td>120.0</td>
<td>3.20</td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>384.2</td>
<td>118.8</td>
<td>3.23</td>
</tr>
<tr>
<td></td>
<td>.10</td>
<td>384.2</td>
<td>115.2</td>
<td>3.33</td>
</tr>
<tr>
<td>LF</td>
<td></td>
<td>384.2</td>
<td>120.0</td>
<td>3.20</td>
</tr>
<tr>
<td>.80</td>
<td>0</td>
<td>245.9</td>
<td>76.8</td>
<td>3.20</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>245.9</td>
<td>75.1</td>
<td>3.28</td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>245.9</td>
<td>67.0</td>
<td>3.67</td>
</tr>
<tr>
<td></td>
<td>.10</td>
<td>245.9</td>
<td>54.7</td>
<td>4.49</td>
</tr>
<tr>
<td>LF</td>
<td></td>
<td>245.9</td>
<td>92.4</td>
<td>2.67</td>
</tr>
<tr>
<td>.90</td>
<td>0</td>
<td>138.3</td>
<td>43.2</td>
<td>3.20</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>138.3</td>
<td>38.0</td>
<td>3.28</td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>138.3</td>
<td>26.9</td>
<td>5.14</td>
</tr>
<tr>
<td></td>
<td>.10</td>
<td>138.3</td>
<td>15.4</td>
<td>9.00</td>
</tr>
<tr>
<td>LF</td>
<td></td>
<td>138.3</td>
<td>61.5</td>
<td>2.25</td>
</tr>
<tr>
<td>.95</td>
<td>0</td>
<td>51.4</td>
<td>16.1</td>
<td>3.20</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>51.4</td>
<td>14.2</td>
<td>3.62</td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>51.4</td>
<td>6.1</td>
<td>8.44</td>
</tr>
<tr>
<td>LF</td>
<td></td>
<td>51.4</td>
<td>34.6</td>
<td>2.11</td>
</tr>
<tr>
<td>.99</td>
<td>0</td>
<td>10.7</td>
<td>3.3</td>
<td>3.20</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>10.7</td>
<td>1.3</td>
<td>8.08</td>
</tr>
<tr>
<td>LF</td>
<td></td>
<td>10.7</td>
<td>5.3</td>
<td>2.02</td>
</tr>
</tbody>
</table>
Remarks about Table 5. (a) The values $\varepsilon = 0.05$ and $P^* = 0.95$ are used only to determine the sample sizes for each procedure. (b) The computations are based on asymptotic normal theory. (c) The abbreviation LF refers to the least favorable configuration among all values of $p$ and $p'$ for a given value of $p_0$. This configuration then gives the smallest possible efficiency for a given $p_0$ and, since $n_c$ is constant for fixed $p_0$, it also gives the largest value of $n_c$ for a given $p_0$. (d) For $p_0 = 0.99$ the number needed is generally too small for the normal approximation to be reliable. For both $p_0 = 0.95$ and 0.99 one-sided inequalities were used in the calculations, since $p_0 + \varepsilon \geq 1$.

The table shows that for all values of the parameters there is a considerable saving in the number of items needed. The efficiency values given by the least favorable configuration are the smallest possible. Thus, for example, with $p_0 = 0.50$, there is an efficiency of over 320%, whereas with $p_0 = 0.99$ there is an efficiency of over 200%.

The results given apply only for $\rho = \tau = 0$; however, similar results hold for other values of $\rho$ and $\tau$.

6. More Than One Correct Choice

In this section we generalize the discussion to the case of exactly $r$ correct alternatives for each question. (We consider only the case $r \leq k/2$, which suffices for most purposes.) This type of model might occur when there are two or more equivalent forms of a correct answer, and the equivalence may not be recognized. For example, $\frac{1}{\sqrt{2}} = 0.707\ldots = \frac{\sqrt{2}}{2} = \frac{1}{1.414\ldots}$, or $\sin 1.5709\ldots = \sin \pi/2 = \sin 90^\circ = 1$ are equivalent correct answers. Similarly, each of several synonyms for the same word would be a correct answer.
In the conventional procedure, \( R_c(r) \), the subject selects exactly \( r \) choices and receives one point for each correct choice. For the subset selection procedure, \( R_s(r) \), the score \( S_r \) per question is

\[
S_r = (k-r) \sum_{i=1}^{r} V_i - r \sum_{j=1}^{k-r} (1-W_j),
\]

(11)

where \( V_i = 1 \) if the \( i \)th correct choice is selected, zero otherwise and \( W_j = 1 \) if the \( j \)th incorrect choice is screened out, zero otherwise. This is equivalent to giving \( k-r \) points for each correct choice selected and taking off \( r \) points for each incorrect choice selected. It follows from (11) that

\[
E(S_r) = r(k-r)[p - (1-p')] = r(k-r)(p + p' - 1).
\]

(12)

The statistic \( S_r \) has a range from \(-r(k-r)\) to \(+r(k-r)\) and to change the range to the unit interval \([0,1]\) we make the transformation to \( Y_r \) as

\[
Y_r = \frac{S_r + r(k-r)}{2r(k-r)},
\]

(13)

which has expectation \( p_0 = (p + p')/2 \). The variance of \( Y_r \) is

\[
\text{Var}(Y_r) = \frac{\sqrt[p]{(1-p)p'(1-p')}}{2} + \frac{\rho p(1-p)}{4} + \frac{\eta p'(1-p')}{4} + \frac{p'(1-p')(1-\eta)}{4r} + \frac{p(1-p)(1-p)}{4(k-r)},
\]

(14)

where \( \eta = \text{Corr}(W_i, W_j) \) is the common correlation between any two of the \( r \) correct alternatives, \( \rho = \text{Corr}(V_i, V_j) \), and \( \tau = \text{Corr}(W_i, V_j) \). Thus for \( Y_r \) (and the mean \( \bar{Y}_r \) over \( n \) questions) we have
\[ E(Y_r) = E(\bar{Y}_r) = p_0, \quad \text{Var}(\bar{Y}_r) = \text{Var}(Y_r)/n. \quad (15) \]

The number of questions required and the efficiency are computed in a manner similar to that of Section 5 and the result is

\[
\text{Eff(subset selection with respect to conventional selection when } r \text{ choices are correct)}
= \text{Eff}(R_s(r), R_c(r)) = \frac{n_{cr}}{n_{sr}}
\]

\[
= \frac{4 p_0 (1-p_0) [1 + (r-1)n]/r}{2 \sqrt{p(1-p)p'(1-p')} + np'(1-p') + (1-n)p'(1-p') + (1-p)p(1-p) + \frac{(1-p)p(1-p)}{k-r}}.
\]

Table 6 gives selected values of \( n_{sr} \) and \( \text{Eff}(R_s(r), R_c(r)) \) for the case \( p = \tau = \eta = 0. \)

Table 6

<table>
<thead>
<tr>
<th>( p_0 )</th>
<th>( \Delta )</th>
<th>( 2 )</th>
<th>( 3 )</th>
<th>( 4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Number} )</td>
<td>( \text{Needed} )</td>
<td>( \text{Eff} )</td>
<td>( \text{Number} )</td>
<td>( \text{Needed} )</td>
</tr>
<tr>
<td>0.50</td>
<td>0</td>
<td>80.0</td>
<td>2.40</td>
<td>80.0</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>80.0</td>
<td>2.40</td>
<td>80.0</td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>79.2</td>
<td>2.42</td>
<td>79.2</td>
</tr>
<tr>
<td></td>
<td>.10</td>
<td>76.8</td>
<td>2.45</td>
<td>76.8</td>
</tr>
<tr>
<td>LF</td>
<td>80.0</td>
<td>2.40</td>
<td>80.0</td>
<td>1.60</td>
</tr>
<tr>
<td>0.80</td>
<td>0</td>
<td>51.2</td>
<td>2.40</td>
<td>51.2</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>50.8</td>
<td>2.42</td>
<td>51.6</td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>48.5</td>
<td>2.53</td>
<td>52.3</td>
</tr>
<tr>
<td></td>
<td>.10</td>
<td>44.2</td>
<td>2.78</td>
<td>51.9</td>
</tr>
<tr>
<td>LF</td>
<td>52.4</td>
<td>2.40</td>
<td>52.4</td>
<td>1.57</td>
</tr>
<tr>
<td>0.90</td>
<td>0</td>
<td>28.8</td>
<td>2.40</td>
<td>28.8</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>28.3</td>
<td>2.45</td>
<td>29.3</td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>25.5</td>
<td>2.72</td>
<td>30.6</td>
</tr>
<tr>
<td></td>
<td>.10</td>
<td>20.5</td>
<td>3.38</td>
<td>30.7</td>
</tr>
<tr>
<td>LF</td>
<td>30.7</td>
<td>2.25</td>
<td>30.7</td>
<td>1.50</td>
</tr>
<tr>
<td>0.95</td>
<td>0</td>
<td>10.7</td>
<td>2.40</td>
<td>10.7</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>10.3</td>
<td>2.50</td>
<td>11.1</td>
</tr>
<tr>
<td></td>
<td>.05</td>
<td>8.1</td>
<td>3.17</td>
<td>12.2</td>
</tr>
<tr>
<td>LF</td>
<td>12.2</td>
<td>2.11</td>
<td>12.2</td>
<td>1.41</td>
</tr>
<tr>
<td>.99</td>
<td>0</td>
<td>2.2</td>
<td>2.40</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>1.8</td>
<td>3.03</td>
<td>2.7</td>
</tr>
<tr>
<td>LF</td>
<td>2.7</td>
<td>2.02</td>
<td>2.7</td>
<td>1.35</td>
</tr>
</tbody>
</table>
As for the case with one correct answer, the efficiency is considerable even when there are three correct answers. When the number of correct answers is too large as in the case \( r = 4 \) out of \( k = 5 \), the efficiency is favorable to the conventional procedure. What is surprising is that even in this most unfavorable case the efficiency is larger than expected.

7. Condition Needed to Use Asymptotic Normal Approximation

To justify the use of the asymptotic normal approximation it is quite useful to find the relation that must hold between \( \eta, \rho \), and \( \tau \) in order that the joint distribution of the \( W_i \) (\( i = 1,2, \ldots, a \)) and the \( V_j \) (\( j = 1,2, \ldots, b \)) should have a positive definite correlation matrix. If this positive definite condition does not hold, then there is a degeneracy in the limit as \( n \to \infty \) and the usual asymptotic normal approximation is not valid. Here we are replacing \( r \) by \( a \) and \( k - r \) by \( b \), for convenience of notation.

If \( \text{Corr}(W_i, W_j) = \eta, \ i,j = 1, \ldots, a; \text{Corr}(V_i, V_j) = \rho, \ i,j = 1, \ldots, b, \) and \( \text{Corr}(W_i, V_j) = \tau, \ i = 1, \ldots, a, \ j = 1, \ldots, b, \) then the correlation matrix is positive definite if and only if

\[
\eta < 1, \ \rho < 1, \quad [1 + (a-1)\eta][1 + (b-1)\rho] - \tau^2 ab > 0 .
\]

This correlation matrix is well known; for a list of references see, e.g., Olkin (1974).

ACKNOWLEDGEMENT

The authors are grateful to Murray Aitkin, Clyde Coombs and Fred Lord for their comments, suggestions and encouragement.
REFERENCES


