THE SCORING OF MULTIPLE CHOICE QUESTIONNAIRES[1]

By Herman Chernoff

Stanford University

1. Summary. In many types of questionnaires the element of guessing is important. An approach to correcting for guessing is proposed. Here one regards the score assigned to a subject on a question as an estimate of the unknown value of the subject's knowledge on the question. This value is one when the subject knows the answer and is zero otherwise. To derive scores with minimum mean squared error, it is necessary to consider the responses of the whole population of subjects. Thus the score for a correct answer to a question depends on the proportion p of correct answers in the population. In the simplest model, we assume that a subject who knows the answer, responds correctly and that others select a response at random among r choices. Then an incorrect response is scored zero and a correct one is assigned a score of λ/p where λ is the proportion of the population who knew the answer and can be estimated using the relation $p = \lambda + (1 - \lambda)/r$.

The general approach is also applied to a variety of more complicated models, each of which are examples of a specified general formulation. These models include the "pairs of questions" model, the "partial knowledge" model and the "scaled questions" model.

While the method applies neatly to single questions, there are fundamental difficulties in extending it to obtain an overall or composite score for a subject on an examination consisting of many questions. This problem is discussed briefly and the simple minded procedure of totaling the scores for the individual questions is partially evaluated.

2. Introduction. We indicate a general approach to the scoring of examinations with multiple choice questions. The main idea may be motivated by referring to a test question which offers three choices, A, B, and C and for which we can assume that A is the correct answer. Suppose that when the results for all of the subjects are tabulated, the three choices A, B, and C have been selected with equal frequencies. If there is a large number of subjects, these frequencies indicate that at most a negligible portion of the subjects *knew* the answer. The customary procedure of scoring a plus one when a subject chooses A would be misleading. It would constitute the throwing of an irrelevant random "noise" into the total test score. Clearly, giving each subject a zero or throwing out the question altogether would be preferable. Suppose, on the other hand, that when the results are tabulated, A is selected 90% of the time, and B and

Received July 1, 1961; revised January 3, 1962.

[1] This work was supported in part by an Office of Naval Research Contract at Stanford University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

C, each 5% of the time. Then it would seem reasonable to assume that 85% of the subjects knew the answer and only 5% guessed it correctly. Here it would make much more sense to score a plus one for each subject who selected A.

Comparing the two cases above, it seems reasonable to expect that a proper procedure for scoring an individual should depend not only on the subject's choice, but on the frequencies of the choices for the entire population of subjects. Once this is granted, it remains to decide on an optimal scoring procedure. The general approach, which consists of regarding the score as an estimate whose mean squared error should be minimized, is discussed and is illustrated in Section 3 for the simplest model. In Section 4, a generalization where two questions are considered simultaneously is discussed. Then the method is generalized in Section 5. There it is seen that the necessary calculus is essentially that of conditional expectation or regression.

Sections 6, 7, and 8 are devoted to other variations of the model. Finally, in Section 9, the question of overall or composite scores is discussed and miscellaneous remarks are left for Section 10.

3. The simplest model. In the simplest model we assume that there are r possible responses to a question. A subject who knows the answer gives the correct answer. One who doesn't know the answer guesses and has probability $1/r$ of arriving at the correct response.

Suppose that the proportion of the population who know the answer is λ. The proportion who answer correctly is given by

$$p = \lambda + (1/r)(1 - \lambda). \tag{3.1}$$

For a large sample of respondents p and λ may be estimated accurately. Suppose that each individual who obtains the correct answer receives a score of $x = x_c$ and the others receive a score of $x = x_w$. We regard x as an estimate of the appropriate value which is one for an individual who knows the answer and zero for the others. The respondents can be partitioned into three groups. There is a proportion λ who know the answer and respond correctly. For them the error of estimate is $x_c - 1$. There is a second group, a proportion $p - \lambda$, who do not know the answer but guess correctly. For them the error of estimate is x_c since they deserve zero. Finally there is the proportion $1 - p$ who do not know the answer and who guess incorrectly. For them the error of estimate is x_w. Consequently the mean squared error is given by

$$V = \lambda(x_c - 1)^2 + (p - \lambda)x_c^2 + (1 - p)x_w^2.$$

Minimizing with respect to x_c and x_w we obtain

$$x_c = \frac{\lambda}{p} = \frac{rp - 1}{(r - 1)p}, \qquad x_w = 0 \tag{3.2}$$

giving a minimum value of

$$V = \lambda(p - \lambda)/p. \tag{3.3}$$

This compares favorably with

$$V^* = p - \lambda \tag{3.4}$$

which is the mean squared error for the usual scoring system which assigns a one to a correct answer and a zero to an incorrect answer.

The scoring method represented by equations (3.2) is subject to two criticisms which we shall now consider. First, one may remark that the scores obtained by this technique are perfectly correlated with those obtained by the usual method. Thus nothing is gained by using this new approach. The second criticism is that a subject's scores will depend not only on his response but also on the responses of the other subjects. In fact a subject who knows the answers to each question on an examination will receive very little credit if most of the other subjects are ignorant. This criticism was leveled by Lyerly [10] at a technique formerly proposed by Hamilton [6] to correct for guessing.

The first criticism oversimplifies the situation. If an examination consists of several questions the sum of a subject's scores under the proposed system will not correlate perfectly with the sum of the scores under the usual system. There is one exception to the last remark. That is the case where all the questions are equally difficult to the population of subjects in the sense that the corresponding λ are equal. Even in this case a modification that we shall soon propose will invalidate the first criticism.

The second criticism may be serious in those cases where it is desired to compare subjects with an objective standard (as expressed by the examination) and an underestimate of the subject's knowledge is worse than an overestimate. In other cases the criticism may be disregarded. If an underestimate is no worse than an overestimate, the proposed system is superior. If it is merely desired to rank the subjects, the proposed system is adequate. If it is merely desired to give the subject who answers all the questions correctly a perfect score, one may compare the total scores with the one that would be given to a subject with all answers correct.

Finally, the following modification will eliminate the first criticism and reduce the effect of the second when it is applicable. If the population of subjects is large, one may subdivide them or stratify them into subgroups and score each subgroup separately. If the stratification is performed according to some criterion which correlates with ability to answer the questions, the new procedure would have improved performance.

For example suppose that the population is divided into three groups of equal size. In the first almost no one knows the answer, in the second half of the subjects know the answer, and in the third almost all of the subjects know the answer. The scores for correct answers in these subgroups would be 0, $(1 + r^{-1})^{-1}$, and 1 respectively. The overall mean squared error would be $[6(r + 1)]^{-1}$ compared with $[2(r + 1)]^{-1}$ if the population is not stratified.

If stratification is applied, the scores are no longer perfectly correlated with those of the usual procedure even when the examination consists of one question.

Soundly applied stratification reduces the effect of the second criticism when it applies. One may note however that stratification could lead to unhappy individual consequences if a few very capable subjects were, by chance, included in the wrong stratum. Even this possibility can be avoided by variations of the following expedient. Obtain total scores without using stratification. Then use these total scores to stratify the population and recalculate new improved scores.

The last suggestion has the "advantage" of objectivity. The final score depends only on the responses of the subject and the population and not on any other extraneous criteria. One can elaborate on it. First one could estimate a regression of the λ for an individual item or question on the total scores. Second one could break up the examination into relevant subportions and then apply a multiple regression. It may even be feasible to apply some form of factor analysis. Finally one could iterate and use the new improved scores to stratify and then to compute further improved scores.

4. Pairs of questions. In this section we shall examine a generalization of the simplest model where the scores on a question depend on the responses to a pair of questions. This model will serve to lay the groundwork for a proper generalization of the method applied in the preceding section. It also represents one attack on the problem of scoring responses so as to reduce the effect of the ignorance of some subjects on the scores of other subjects.

Let λ_{11} be the proportion of the subjects who know the answer to both questions, λ_{10} be the proportion who know the answer to the first only, λ_{01} be the proportion who know the answer to the second only, and λ_{00} be the proportion who don't know either answer. Similarly we define p_{cc}, p_{cw}, p_{wc}, and p_{ww} in terms of the proportions who answer the questions correctly (c) and incorrectly (w). We are assuming that subjects who don't know the answer select a response at random.

Then,

$$
\begin{aligned}
p_{cc} &= \lambda_{11} + \frac{\lambda_{10}}{r} + \frac{\lambda_{01}}{r} + \frac{\lambda_{00}}{r^2}, \\
p_{cw} &= \lambda_{10}\frac{(r-1)}{r} + \lambda_{00}\frac{(r-1)}{r^2}, \\
p_{wc} &= \lambda_{01}\frac{(r-1)}{r} + \lambda_{00}\frac{(r-1)}{r^2}, \\
p_{ww} &= \lambda_{00}\left(\frac{r-1}{r}\right)^2.
\end{aligned}
\tag{4.1}
$$

Let x_{icc}, x_{icw}, x_{iwc}, x_{iww} be the scores assigned for the ith question to individuals whose results for the pair of questions are cc, cw, wc, and ww respectively, $i = 1, 2$. Then, the minimum mean squared error of the estimate x_{icc}

from the appropriate value (which is one for individuals who know the answer to the ith question and 0 for the others) is obtained as follows:

Each subject belongs to one of nine nonempty classes determined by his knowledge and his answers. To each is assigned a value due to the knowledge and a score due to the pair of answers. Minimizing the mean squared error consists of minimizing

$$V_1 = \left(\lambda_{11} + \frac{\lambda_{10}}{r}\right)(x_{1cc} - 1)^2 + \left(\frac{\lambda_{01}}{r} + \frac{\lambda_{00}}{r^2}\right)x_{1cc}^2 + \frac{(r-1)\lambda_{10}}{r}(x_{1cw} - 1)^2$$
$$+ \lambda_{00}\frac{(r-1)}{r^2}x_{1cw}^2 + \left[\lambda_{01}\frac{(r-1)}{r} + \lambda_{00}\frac{(r-1)}{r^2}\right]x_{1wc}^2 + \lambda_{00}\left(\frac{r-1}{r}\right)^2 x_{1ww}^2$$

and

$$V_2 = \left(\lambda_{11} + \frac{\lambda_{01}}{r}\right)(x_{2cc} - 1)^2 + \left(\frac{\lambda_{10}}{r} + \frac{\lambda_{00}}{r^2}\right)x_{2cc}^2 + \frac{r-1}{r}\lambda_{01}(x_{2cw} - 1)^2$$
$$+ \lambda_{00}\frac{(r-1)}{r^2}x_{2cw}^2 + \left[\lambda_{10}\frac{r-1}{r} + \lambda_{00}\frac{r-1}{r^2}\right]x_{2wc}^2 + \lambda_{00}\left(\frac{r-1}{r}\right)^2 x_{2ww}^2.$$

We obtain

$$\begin{aligned} x_{1cc} &= \frac{\lambda_{11} + (\lambda_{10}/r)}{p_{cc}} & x_{1wc} &= 0 \\ x_{1cw} &= \frac{(r-1)\lambda_{10}/r}{p_{cw}} & x_{1ww} &= 0 \\ x_{2cc} &= \frac{\lambda_{11} + (\lambda_{01}/r)}{p_{cc}} & x_{2wc} &= \frac{(r-1)\lambda_{01}/r}{p_{wc}} \\ x_{2cw} &= 0 & x_{2ww} &= 0 \end{aligned} \tag{4.2}$$

which yield minimizing values of

$$\begin{aligned} V_1 &= \frac{\left[\lambda_{11} + \frac{\lambda_{10}}{r}\right]\left[\frac{\lambda_{01}}{r} + \frac{\lambda_{00}}{r^2}\right]}{p_{cc}} + \frac{\left[\frac{(r-1)\lambda_{10}}{r}\right]\left[\frac{\lambda_{00}(r-1)}{r^2}\right]}{p_{cw}}, \\ V_2 &= \frac{\left[\lambda_{11} + \frac{\lambda_{01}}{r}\right]\left[\frac{\lambda_{10}}{r} + \frac{\lambda_{00}}{r^2}\right]}{p_{cc}} + \frac{\left[\frac{(r-1)}{r}\lambda_{01}\right]\left[\frac{\lambda_{00}(r-1)}{r^2}\right]}{p_{cw}}. \end{aligned} \tag{4.3}$$

Thus by combining two questions, one may obtain some reduction in the mean squared error for each question. Technically this reduction is obtained by using more information; namely the combined scores. To accomplish this requires more computation and an extension of the model. This extension involves assuming independence of the guesses for individuals who don't know the answers.

For an examination consisting of only two questions the method proposed in

this section is an improvement over the suggestion in Section 3 of stratifying by total scores. Its efficacy depends on the extent to which the knowledge of the answers to the two questions are correlated.

In principle the model and results of this section can be extended to examinations of arbitrary length. They would give results superior to those based on stratification using only criteria involving responses on the examination. Unfortunately this extension is not very practical. The computational difficulties grow enormously as the number of questions increases. Furthermore the number of subjects necessary to estimate reliably the appropriate λ's also grows enormously. As a result, the author would recommend using this model or an extension of it only on a portion of the examination consisting of a small number of strongly "related" questions. The word related is used in the sense that knowledge of the answers are correlated.

Suppose that the two questions are of unequal difficulty but of equal value. Then the subject who answers the easier question correctly and the other incorrectly will have a higher total score than the subject who misses the easier question but answers the other correctly. Some reflection may be required to make this fact seem non paradoxical.

A couple of similar comments were suggested to me by F. M. Lord. One is that a subject who does well on questions 2, 3, $\cdots$ should (other things being equal) receive a better score on question 1 than a subject who does poorly on questions 2, 3, $\cdots$. This is presumably correct if we extend the model of this section to cover several questions and if knowledge of the answers to various questions are positively correlated.

In Section 7 an extreme case of a generalization of the pair of questions model, called the scaled model, is discussed. In this case knowledge of the answer to one question implies knowledge of the answers to the preceding questions and it is relatively easy to analyze several questions simultaneously.

The reader may have noticed that the model of this section is related to that used in configural analysis (see [9]).

5. A general formulation. The models of Sections 3 and 4 suggest the following general formulation which will be applied to several other models in Sections 6, 7 and 8.

Let w represent a random variable associated with a random subject. In our applications it will generally be a vector valued random variable which will characterize the knowledge of the subject and his responses. In Section 4, $w = (w_1, w_2, w_3, w_4)$ could be used to describe a subject where $w_1 = 1$ if he knows the answer to the first question and 0 otherwise, $w_2 = 1$ if he knows the answer to the second question and 0 otherwise, and w_3 and w_4 are his responses to the two questions.

Let z represent the *observed responses*. In the above example z can be regarded as the subvector (w_3, w_4). In general z will be some function of w, i.e.,

$$z = f(w). \tag{5.1}$$

We are interested in the value of the subject's knowledge since that is what is to be estimated. In the above application the value is $w_1 + w_2$ (assuming the two questions are equally important). The value can not be observed directly. It is to be estimated from the observed data represented by z. In any case the value v is also some function of w, i.e.,

$$v = g(w). \tag{5.2}$$

Now we select an assigned score x to serve as the estimate of v. Since x is to depend on the observed data, we write

$$x = h(z). \tag{5.3}$$

We shall select the scoring system h which minimizes the mean squared error

$$V = E\{(x - v)^2\}. \tag{5.4}$$

Applying the properties of conditional expectation, we note that

$$V = E\{E[(x - v)^2 \mid z]\} \tag{5.5}$$

and hence is minimized by selecting

$$x = h(z) = E\{v \mid z\}, \tag{5.6}$$

i.e., x is the regression of v on z. To apply this method of scoring it is necessary that enough be known about the probability distribution of w so that the above regression can be evaluated. In the models of Sections 3 and 4, it was possible to use the data for the population of subjects to estimate the probability distribution of w. It would suffice to know the joint distribution of v and z.

For the above method of scoring, the minimum variance is

$$V = E\{v^2\} - E\{x^2\}. \tag{5.7}$$

For an arbitrary scoring system $x^* = h^*(z)$, the mean squared error is increased by

$$V^* - V = E\{(x^* - x)^2\} \tag{5.8}$$

since

$$V^* = E\{(x^* - v)^2\} = E\{(x^* - x)^2\} + E\{(x - v)^2\} + 2E\{(x^* - x)(x - v)\}$$

and

$$\begin{aligned} E\{(x^* - x)(x - v)\} &= E\{E[(x^* - x)(x - v) \mid z]\}, \\ &= E\{(x^* - x)E[(x - v) \mid z]\}, \\ &= 0. \end{aligned}$$

This general model and these results can be applied to the examination as a whole or to individual items or to sections of the examination. We defer some discussion till Section 9 on total scores.

6. Partial knowledge. The simple model of Section 3 does not apply in many cases. One variation introduced by Horst [1] and discussed by Gulliksen ([5], pp. 1–486) and Solomon [12] allows for partial knowledge.

In Horst's model which we shall call the *simple ordering of responses* model, there is an order of the r possible responses. When they are arranged in order, a subject's knowledge will consist of knowing that the correct response is among the first i possible responses for some i between 1 and r. The subject selects one of these i responses at random. If λ_i denotes the portion of the population who can limit their answer to the first i responses, the proportion who select the jth response is

$$p_j = \frac{\lambda_j}{j} + \frac{\lambda_{j+1}}{j+1} + \cdots + \frac{\lambda_r}{r} \qquad 1 \leqq j \leqq r. \tag{6.1}$$

A large sample of responses will enable the investigator to estimate the p_j's directly. These will determine the appropriate order (if it is not already known) because the lower index in the appropriate order is associated with a larger p_j. Finally

$$\lambda_j = j[p_j - p_{j+1}] \qquad 1 \leqq j \leqq r. \tag{6.2}$$

In particular, we have the known result that $\lambda_1 = p_1 - p_2$, i.e., the percentage who know the answer is the difference in the observed proportions for the two most favored responses.

Let us relate this example to the general formulation of Section III. For each subject let $w = (I, J)$ where I is the group from which the subject made his selection and J is his choice. The observed datum is $z = J$. The value $v = g(w)$ will ordinarily depend on the unknown I. Thus we may represent this dependence by $v = v_I$. The investigator who regards partial knowledge as worthless will let $v_1 = 1$ and $v_2 = v_3 = \cdots = v_r = 0$. For those cases where partial knowledge is regarded as valuable some other assignment can be made.

According to this model,

$$\begin{aligned} P\{w = (i,j)\} &= \lambda_i/i && \text{for } 1 \leqq j \leqq i, \\ &= 0 && \text{otherwise.} \end{aligned} \tag{6.3}$$

To derive the optimal score $x = E\{v \mid z\}$ we first compute the conditional probability of w given $z = J = j$

$$P\{w = (i,j) \mid J = j\} = \frac{\lambda_i/i}{\sum_{\alpha=j}^{r} \lambda_\alpha/\alpha} = \frac{\lambda_i}{ip_j} \quad \text{for} \quad i = j, \quad j+1, \cdots, r. \tag{6.4}$$

Thus the score corresponding to $J = j$ is

$$x_j = E\{v \mid J = j\} = \frac{1}{p_j} \sum_{i=j}^{r} \frac{\lambda_i v_i}{i} \tag{6.5}$$

and

$$V = \sum_{i=1}^{r} \lambda_i v_i^2 - \sum_{j=1}^{r} p_j x_j^2 \tag{6.6}$$

while the "standard" procedure of assigning v_j to a choice of j would lead to an increase of

$$V^* - V = \sum_{j=1}^{r} p_j(x_j - v_j)^2 \tag{6.7}$$

For the important special case where $v_1 = 1$ and $v_2 = v_3 = \cdots = v_r = 0$,

$$\begin{aligned} x_1 &= \lambda_1/p_1 = (p_1 - p_2)/p_1 , \\ x_2 &= x_3 = \cdots = x_r = 0. \end{aligned} \tag{6.5a}$$

Here we have

$$V = \lambda_1 - \lambda_1^2/p_1 = [p_2(p_1 - p_2)]/p_1 \tag{6.6a}$$

and

$$V^* - V = p_2^2/p_1 . \tag{6.7a}$$

The simple ordering of responses is not the only way in which partial knowledge can be represented. Other more complex models could be applied to yield what we shall call *complex orderings*. In general let us assume that the subject's knowledge can be represented by a set S of possible responses from which he selects one at random and among which is the correct response.[2] If $S_1 , S_2 , \cdots , S_{r'}$ are possible sets one may attach values $v_1 , v_2 , \cdots , v_{r'}$ to the corresponding knowledge. In some cases it is possible to use the observed frequencies of response to estimate $\lambda_1 , \lambda_2 , \cdots , \lambda_{r'}$ where λ_i is defined as the proportion of subjects whose knowledge is represented by S_i. We present one example. Let $S_i = \{a\}$, $S_2 = \{a, b\}$, $S_3 = \{a, c\}$, $S_4 = \{a, b, c, d\}$. From the observed frequencies of responses p_a , p_b , p_c , p_d, we estimate $\lambda_1 , \lambda_2 , \lambda_3$, and λ_4 by solving

$$\begin{aligned} p_a &= \lambda_1 + \tfrac{1}{2}\lambda_2 + \tfrac{1}{2}\lambda_3 + \tfrac{1}{4}\lambda_4 , \\ p_b &= \tfrac{1}{2}\lambda_2 + \tfrac{1}{4}\lambda_4 , \\ p_c &= \tfrac{1}{2}\lambda_3 + \tfrac{1}{4}\lambda_4 , \\ p_d &= \tfrac{1}{4}\lambda_4 . \end{aligned} \tag{6.8}$$

While $r' \leqq r$ is necessary for the λ's to be estimable, this condition is not sufficient.

If the λ's are estimable the general approach of section 5 is applicable. In those examples of complex ordering where the λ's are not consistently estimable,[3]

[2] Still more complicated models can be constructed where the choice among the elements of S are not equally probable or even where S does not contain the correct answer.

[3] The appropriate technical term is *unidentified*.

a minimax approach may be applied to the mean squared error. That is, $x = h(z)$ could be selected to minimize

$$\max_{\theta \in \Omega} E\{(x - v)^2\} \tag{6.9}$$

where Ω is the class of all distributions of w consistent with the observed distribution of z. Note that for given θ and given $x = h(z)$

$$E_\theta\{(x - v)^2\} = E\{[x - E_\theta(v \mid z)]^2\} + E_\theta\{[v - E_\theta(v \mid z)]^2\}. \tag{6.10}$$

The first expectation on the right hand side does not have a θ subscript because it can be computed by using the observed distribution of z. Now fix θ and let $v^*(\theta)$ represent a point in $r + 1$ dimensional space whose coordinates are $E_\theta\{v \mid z = 1\}$, $E_\theta(v \mid z = 2), \cdots, E_\theta(v \mid z = r)$, $[E_\theta\{[v - E_\theta(v \mid z)]^2\}]^{\frac{1}{2}}$. Let x^* be the point whose coordinates are $x_1 = h(1)$, $x_2 = h(2), \cdots, x_r = h(r)$, 0. Then $E_\theta\{(x - v)^2\}$ is the squared distance between x^* and $v^*(\theta)$. The minimax estimate has a simple geometric interpretation. It corresponds to the point on a given r dimensional plane in $r + 1$ dimensional space for which the maximum distance to the points of a specified set is as small as possible. The specified set is $\{v^*(\theta): \theta \in \Omega\}$. We shall not discuss complex orderings further.

7. Questions which can be scaled. In Section 4 we discussed a model where the results on two questions were combined. One object was to capitalize on the correlations among the knowledge of the answers to several questions. The analysis for two questions was more difficult than for a single question. We indicated two objections to the extension to many questions. These were the difficulty of analysis and the large sample size required to obtain good estimates of the necessary parameters.

There is a special extreme case for which these objections do not apply. This is the case where the questions test knowledge which can be scaled. More precisely, suppose that there are s questions such that any subject who knows the answer to the ith question knows the answer to all the preceding ones. If he knows the answer to only the first i, he answers them correctly but guesses at random among the r choices on each of the subsequent questions. Let λ_i be the proportion of subjects in this category for $i = 1, 2, \cdots, s$ and let λ_0 be the proportion who don't know the answer to any of the questions. Let

$$w = (I, J, K_1, K_2, \cdots, K_L) \tag{7.1}$$

where $I \geqq 0$ indicates the number of questions whose answers the subject knows, $J \geqq I$ indicates the number of correct answers before a wrong one appears and $K_1, K_2, \cdots, K_L$ are the other questions, if any, which are answered correctly. The subject's response may be represented by

$$z = (J, K_1, K_2, \cdots, K_L). \tag{7.2}$$

The value associated with the subjects knowledge can be represented by $v = v_I$. One may use $v = I$ although other measures may be appropriate on occasion.

We have

$$P\{w = (i, j, k_1, k_2, \cdots, k_l)\} = \lambda_i \left(\frac{1}{r}\right)^{j+l-i} \left(\frac{r-1}{r}\right)^{s-j-l}, \tag{7.3}$$

$$P\{z = (j, k_1, k_2, \cdots, k_l)\} = \sum_{i=0}^{j} \lambda_i r^{-(s-i)}(r-1)^{s-j-l}. \tag{7.4}$$

$$P\{w = (i, j, k_1, \cdots) \mid z = (j, k_1, \cdots)\} = \lambda_i r^i \Big/ \sum_{i=0}^{j} \lambda_i r^i \tag{7.5}$$

for $0 \leqq i \leqq j$.

Then, given $z = (j, k_1, k_2, \cdots, k_l)$, the appropriate score is

$$x_j = \sum_{i=0}^{j} v_i \lambda_i r^i \Big/ \sum_{i=0}^{j} \lambda_i r^i \tag{7.6}$$

where λ_i may be estimated using the fact that

$$p_j = P\{J = j\} = \frac{r-1}{r^{j+1}} \sum_{i=0}^{j} \lambda_i r^i. \tag{7.7}$$

and hence

$$\lambda_j = (r-1)^{-1}\{rp_j - p_{j-1}\} \tag{7.8}$$

where the p_j may be directly estimated from the observed frequencies of the random variable J. As may have been expected, $K_1, K_2, \cdots, K_L$ are not involved in the estimation of λ and do not influence x.

The mean squared error corresponding to the optimal procedure is given by

$$V = \sum_{i=0}^{r} \lambda_i v_i^2 - \sum_{j=0}^{r} p_j x_j^2. \tag{7.9}$$

8. A model involving reexamination. An example which was brought to the author's attention consisted of a Spanish Language examination which was given to a group of subjects before and after special training. Each question offered three choices and it seemed clear from the nature of the examination that the simple ordering partial knowledge model was appropriate. On the other hand one would expect a definite relationship between the knowledge before and after the training. That is, one would expect that on the whole, students would know at least as much for the second examination as they would for the first. Then it would make sense to consider a question before training and one after training as a pair of questions to which the models of scaling and simple ordering could be applied.

However data were unavailable to substantiate the expectation that subjects did not forget between the examinations which were to be given six months apart. Therefore it was decided to drop the scaling assumption out of fear that if it did not hold, the results would be questionable. The following model was applied.

Let $1, 2, \cdots, r$ be the natural order of choice among the possible responses. Let $S_1 = \{1\}$, $S_2 = \{1, 2\}$, $S_3 = \{1, 2, 3\}$, $\cdots$ represent the possible knowledge of a subject responding to the question. We represent the combined knowledge and response of a subject to the question on both examinations by

$$w = (I_1, I_2, J_1, J_2) \tag{8.1}$$

where the subject knows that the answer is in S_{I_1} on the first examination, knows that the answer is in S_{I_2} on the second examination, and responds J_1 on the first and J_2 on the second examination. Let v_i be the value associated with S_i. Thus the subject deserves a value $v_{(1)} = v_{I_1}$ on the first examination a value $v_{(2)} = v_{I_2}$ on the second. In the above mentioned Spanish examination, the investigator was satisfied only with complete knowledge, i.e., $v_1 = 1$, $v_2 = v_3 = 0$. The responses are denoted by $z = (J_1, J_2)$.

Let

$$\lambda_{i_1,i_2} = P\{I_1 = i_1, I_2 = i_2\}. \tag{8.2}$$

Then

$$P\{w = (i_1, i_2, j_1, j_2)\} = \frac{\lambda_{i_1,i_2}}{i_1 i_2} \quad \text{for} \quad 1 \leqq j_1 \leqq i_1, 1 \leqq j_2 \leqq i_2, \tag{8.3}$$

and

$$p_{j_1,j_2} = P\{z = (j_1, j_2)\} = \sum_{\substack{j_1 \leqq i_1 \leqq r \\ j_2 \leqq i_2 \leqq r}} \lambda_{i_1,i_2}(i_1 i_2)^{-1}.$$

The appropriate scores to be assigned to a subject for the two tests are given by

$$x_1 = E\{v_{I_1} \mid z = (j_1, j_2)\} \tag{8.4}$$

and

$$x_2 = E\{v_{I_2} \mid z = (j_1, j_2)\}. \tag{8.5}$$

If $z = (j_1, j_2)$,

$$x_1 (p_{j_1,j_2})^{-1} \left\{ \sum_{\substack{j_1 \leqq i_1 \leqq r \\ j_2 \leqq i_2 \leqq r}} v_{i_1} \lambda_{i_1,i_2}(i_1 i_2)^{-1} \right\}.$$

However

$$\sum_{i_2=j_2}^{r} \lambda_{i_1,i_2}(i_1 i_2)^{-1} = p_{i_1,j_2} - p_{i_1+1,j_2}$$

where a p with $r + 1$ for a subscript is assumed to be zero. It follows that if $z = (j_1, j_2)$

$$x_1 = (p_{j_1,j_2})^{-1} \left\{ \sum_{i_1=j_1}^{r} v_{i_1}[p_{i_1,j_2} - p_{i_1+1,j_2}] \right\} \tag{8.6}$$

and

$$x_2 = (p_{j_1,j_2})^{-1}\left\{\sum_{i_2=j_2}^{r} v_{i_2}[p_{j_1,i_2} - p_{j_1,i_2+1}]\right\}. \tag{8.7}$$

It may be noted that $(i_1 i_2)^{-1}\lambda_{i_1,i_2}$ can be computed by taking second differences. Then

$$\lambda_{i_1,i_2} = (i_1 i_2)[p_{i_1,i_2} - p_{i_1+1,i_2} - p_{i_1,i_2+1} + p_{i_1+1,i_2+1}]. \tag{8.8}$$

We illustrate with an artificial example with $r = 3$, $v_1 = 1$, $v_2 = .2$, $v_3 = 0$, where the frequency of observations (J_1, J_2) is given by the following table which also includes the marginal distributions

p_{j_1,j_2}

j_1 \ j_2	1	2	3	
1	.340	.130	.035	.505
2	.180	.100	.025	.205
3	.100	.070	.020	.190
	.620	.300	.080	

The tables for $p_{i_1,j_2} - p_{i_1+1,j_2}$ and $p_{j_1,i_2} - p_{j_1,i_2+1}$ are given below.

$p_{i_1,j_2} - p_{i_1+1,j_2}$

i_1 \ j	1	2	3
1	.160	.030	.010
2	.080	.030	.005
3	.100	.070	.020

$p_{j_1,i_2} - p_{j_1,i_2+1}$

j_1 \ i_2	1	2	3
1	.210	.095	.035
2	.080	.075	.025
3	.030	.050	.020

We compute x_1 and x_2 for each possible value (j_1, j_2) of z.

x_1

j_1 \ j_2	1	2	3
1	.518	.277	.314
2	.089	.060	.040
3	.000	.000	.000

x_2

j_1 \ j_2	1	2	3
1	.674	.146	.000
2	.528	.150	.000
3	.400	.142	.000

The table for λ_{i_1,i_2} and the associated marginal probabilities $\lambda_{i_1\cdot}$ and $\lambda_{\cdot i_2}$ is

also presented below

λ_{i_1,i_2}

i_1 \ i_2	1	2	3	
1	.130	.040	.030	.200
2	.100	.100	.030	.230
3	.090	.300	.180	.570
	.320	.440	.240	

In specific applications it may pay to compute the λ's merely to determine whether the model is consistent with the data. If some of the λ's are negative one may ask whether the model fails to apply or whether sampling variation was responsible for the discrepancy.

We may compute

$$V_1 = E\{v_{(1)}^2\} - E\{x_1^2\} = .1029 \tag{8.9}$$

and

$$V_2 = E\{v_{(2)}^2\} - E\{x_2^2\} = .1107 \tag{8.10}$$

In this example the mean squared error derived from using the naive scores x_1^* and x_2^* which assign value v_j to response j are given by

$$V_1^* = E\{(x_1^* - v_1)^2\} = .2712 \tag{8.11}$$

and

$$V_2^* = E\{(x_2^* - v_2)^2\} = .2240. \tag{8.12}$$

Finally, if the relationship between the two questions is ignored and the simple ordering partial knowledge model is applied to each we would have scores given by $x_1^{**} = .442$, .075, and 0 for $J_1 = 1$, 2, and 3 respectively and $x_2^* = .587$, .147 and 0 for $J_2 = 1$, 2, and 3. The mean squared error for this procedure is given by

$$V_1^{**} = .1090$$

and

$$V_2^{**} = .1174.$$

In this example, combining the two questions does not lead to much improvement over considering the questions separately.

The special case where $v_1 = 1$, $v_2 = v_3 = \cdots = v_r = 0$ gives rise to much

simpler formulae for the scores. Here, if $z = (j_1, j_2)$, we have

(8.6a) $$x_1 = (p_{1,j_2})^{-1}(p_{1.j_2} - p_{2,j_2}), \quad \text{if } j_1 = 1,$$
$$x_1 = 0 \quad \text{if } j_1 \neq 1,$$

(8.7a) $$x_2 = (p_{j_1,1})^{-1}(p_{j_1,1} - p_{j_1,2}), \quad \text{if } j_2 = 1,$$
$$x_2 = 0 \quad \text{if } j_2 \neq 1,$$

(8.9a) $$V_1 = \lambda_{1.} - \sum_{j_2=1}^{r} (p_{1,j_2})^{-1}(p_{1,j_2} - p_{2,j_2})^2,$$

(8.10a) $$V_2 = \lambda_{.1} - \sum_{j_1=1}^{r} (p_{j_1,1})^{-1}(p_{j_1,1} - p_{j_1,2})^2.$$

9. Composite scores. Consider an examination consisting of a considerable number of individual questions. How should a subject be assigned a composite score for the examination as a whole? At several points in this paper it may have seemed to be implicitly implied, that simply totaling the scores for the individual questions or items in the examination is the proper and natural procedure. Indeed, this procedure is far from unreasonable and it will be evaluated incidentally later in this section. However it is important to note that the method of totaling, innocent though it seems, is ordinarily non-optimal. This non-optimality arises principally from the possible interdependence of different questions. Thus knowledge of the answer of one question is ordinarily correlated with knowledge of the answer to another. Another issue, which arises and is briefly discussed at the end of this section, is the possibility that composite scores are desired, not so much to evaluate a subject's knowledge, as, to discriminate among subjects of different ability.

Let us digress briefly to point out that, if it is desired to evaluate a subject's knowledge, the general formulation of Section 5 applies to the entire examination considered as a unit. However, unless the examination consists of few questions or of a few sections to each of which the scale model of Section 7 is applicable, practical difficulties would make it unfeasible to apply the method of Section 5. In fact these difficulties are exactly those mentioned in Section 4 where it was indicated that it would be unfeasible to extend the two question model very far.

How should we proceed if we discard the idea of applying the general formulation to the entire examination? One approach that has been applied consists of adopting some strongly parametrized assumptions and then applying standard techniques such as maximum-likelihood to the data. In such cases the fundamental approach of this paper is essentially irrelevant. Maximum-likelihood estimation has been applied with forms of the probit and logit model by Lord [8] and Birnbaum [1, 2, 3]. Although they were not mainly interested in multiple choice questionnaires, these were considered. On the other hand the models they applied avoided an important aspect of the possibility of interdependence of different questions.

A second approach which we propose here is to let the composite score be the sum of individual scores which are assigned so as to minimize the mean squared deviation of the sum from the overall value of the subject's knowledge.

To illustrate let us apply this approach to the case of an examination consisting of k questions, for each of which the simplest model with r choices applies. As before p_i and λ_i represent respectively the proportions of the population of subjects who obtain the correct response and who know the answer to the ith question. Also

$$\lambda_{ij} = \lambda_i\lambda_j + \theta_{ij} \tag{9.1}$$

represents the proportion who know the answer to both the ith and jth questions and can be estimated by observing p_i, p_j, and p_{ij}, the proportion who answer both questions correctly. In fact

$$p_{ij} = p_ip_j + \theta_{ij}[1 - (1/r)]^2 \tag{9.2}$$

if we assume that responses to questions for which the subject doesn't know the answer are random and independent of one another. Let us also assume that the value of a subject's knowledge can be expressed by

$$v = \sum_{i=1}^{k} v_i \tag{9.3}$$

where $v_i = a_i$ if the subject knows the answer to the ith question and 0 otherwise.

For the sake of convenience, we may express the subject's total or composite score by

$$x^* = \delta_0 + \sum_{i=1}^{k} (x_i + u_i) = \delta_0 + \sum_{i=1}^{k} x_i^* \tag{9.4}$$

where $x_i^* = x_i + u_i$, $u_i = \delta_i$ if the ith answer is correct and 0 otherwise, and x_i is the appropriate score for the single question model. That is, $x_i = a_i\lambda_i/p_i$ for a correct answer and 0 otherwise. Then the mean squared deviation of the composite score from the value of the subject's knowledge is given by

$$\begin{aligned} V^* &= E\{\delta_0 + \sum (x_i^* - v_i)\}^2 \\ &= (\delta_0 + \sum_i \delta_ip_i)^2 + E\{\sum_i (x_i - v_i) + (u_i - \delta_ip_i)\}^2. \end{aligned}$$

Now

$$\begin{aligned} \sum_i E\{(x_i - v_i) + (u_i - \delta p_i)\}^2 &= \sum_i E\{x_i - v_i\}^2 + \sum_i E\{u_i - \delta_ip_i\}^2 \\ &= V + \sum_i p_i(1 - p_i)\delta_i^2 \end{aligned}$$

where

$$V = \sum_i a_i^2 \lambda_i \left(1 - \frac{\lambda_i}{p_i}\right) \tag{9.5}$$

represents the sum of the mean squared deviations for optimal responses to individual questions. Finally, for $i \neq j$,

$$E\{(x_i - v_i + u_i - \delta_i p_i)(x_j - v_j + u_j - \delta_j p_j)\}$$
$$= E\{(x_i^* - v_i)(x_j^* - v_j)\} - \delta_i \delta_j p_i p_j$$
$$= \theta_{ij} \left\{\frac{a_i}{rp_i} - \delta_i\left(1 - \frac{1}{r}\right)\right\}\left\{\frac{a_j}{rp_i} - \delta_j\left(1 - \frac{1}{r}\right)\right\}$$

and we have

$$V^* = V + (\delta_0 + \sum_i \delta_i p_i)^2 + \sum_i p_i(1 - p_i)\delta_i^2 \sum_{i \neq j} \theta_{ij} \left\{\frac{a_i}{rp_i} - \delta_i\left(1 - \frac{1}{r}\right)\right\}\left\{\frac{a_j}{rp_i} - \delta_j\left(1 - \frac{1}{r}\right)\right\}. \tag{9.6}$$

It follows that in the case of independence, $\theta_{ij} = 0$ for $i \neq j$, the optimal score would be $x = \sum x_i$. In fact the mean squared error for x is easily obtained, even when $\theta_{ij} \neq 0$, by setting all the δ_i equal to zero in equation (9.6)

Note that in the general case δ_0 should be selected to make the mean error zero. With this adjustment, it is desirable to minimize

$$\sum_i p_i(1 - p_i)\delta_i^2 + \sum_{i \neq j} \theta_{ij} \left\{\frac{a_i}{rp_i} - \delta_i\left(1 - \frac{1}{r}\right)\right\}\left\{\frac{a_j}{rp_j} - \delta_j\left(1 - \frac{1}{r}\right)\right\}$$

whose partial derivative with respect to δ_{i_0} is given by

$$2\delta_{i_0} p_{i_0}(1 - p_{i_0}) - 2\left(1 - \frac{1}{r}\right) \sum_{j \neq i_0} \theta_{i_0 j}\left[\frac{a_j}{rp_j} - \delta_j\left(1 - \frac{1}{r}\right)\right].$$

Hence if the θ_{ij} are small and positive, as is often assumed to be the case, the optimal δ_i would tend to be positive.

It should be emphasized that in the procedure evaluated above we have severely limited our choice of scoring methods and we have no simple way of evaluating how serious a loss this limitation imposes.

Thus far we have assumed that the main objective of the examination score is to evaluate the subject's knowledge. There are occasions when this is not the primary objective. In many cases where a single composite score is desired the objective is to discriminate among students of different ability. Then, quite a different approach to scoring may be appropriate. For example, we propose here to use as our discriminant function a composite score of the form

$$D = \sum_i \alpha_i x_i \tag{9.7}$$

where the x_i is the score which would be appropriate for estimating v_i, the value of the ith part of an examination consisting of k *independent* parts. In fact let us select the α_i so as to minimize

$$V_D = \sum_i \alpha_i^2 E(x_i - v_i)^2 \tag{9.8}$$

subject to the condition that

$$\mu_D = \sum \alpha_i E(v_i) \tag{9.9}$$

be a fixed specified value. This criterion leads to the use of

$$\alpha_i = E(v_i)/E(x_i - v_i)^2. \tag{9.10}$$

This proposal is subject to criticism. An argument could be made for it if one were to assume that the difference between two subjects were expressed by a factor θ so that the ratios of their v_i's, v_{i1}/v_{i2} is approximately θ.

This last assumption is questionable. For individual questions the ratio v_{i1}/v_{i2} can vary enormously. Even in considering groups of questions v_{i1}/v_{i2} will tend to be one for a group of easy questions and may be quite different for a group of difficult questions. The assumption of independence is also questionable in practice.

Aside from doubts about the assumptions, one should note that for easy questions, $E(x_i - v_i)^2$ may be small compared to $E(v_i)$. Then a subject who missed an easy question would suffer severely. While there is justification for this, one should consider tempering the "punishment" in case the question was missed for reasons outside the model such as a slip of the pen or an oversight.

Finally note that the equations of this section apply even in those case where stratification of the population preceded the application of scoring.

10. Miscellaneous remarks. The method proposed in this paper involves a combination of two ideas. First, the performance of the population can be used to estimate the distribution of knowledge or ability among the population. Second, knowing the distribution of knowledge among the population one may apply minimum mean squared error estimation to estimate the subject's knowledge.

This type of combination appeared in a problem given on matriculation examinations at Columbia University and Stanford University. This problem originated, I believe, with Howard Raiffa, as a means of exposing the foundation of Robbins' Theory of Empirical Bayes Procedures. It evolved to the following form "A tubetester always reads good for a good tube and reads good for 25% of the bad tubes. A million tubes are tested. The cost of rejecting a good tube is equal to that of accepting a bad tube. What procedure is called for if 240,000 of the tubes receive good readings?" This problem seemed rather difficult and exotic to many students.

The possibility of using the performance of the population to help score individuals arose in a method suggested by Hamilton [6] to correct for guessing and in a predecessor of this method due to Calandra [4]. As we indicated in Section 3, this use was subjected to some criticism.

The idea of applying minimum mean squared error, and consequently regression theory, was previously introduced by Sitgreaves [11] in a different model,

which involved a combination of the probit type of model used by Lord [8] and that of questions which were scaled.

When enough of a parametric structure is put into the model it becomes feasible to attack the problem of design of questionnaires. The author has done some preliminary work for a model where the probability that an individual know the answer to the ith question is

$$\lambda_i = 1 - e^{-\theta t_i}$$

with t_i a measure of the "ease" of the question and θ a measure of the ability of the subject. Similar work was done for multiple choice questionnaires incidentally in the attack on the design of non-multiple choice questions by Lord [8], Birnbaum [3]. The general problem of design is one of the principal subjects of "Studies in Item Analysis and Prediction," edited by Solomon [13].

The validity of a test plays an important role in most studies of item analysis. In this paper we have ignored this question and there is no essential need to introduce it. The fundamental issue here concerns how to estimate what the test does measure. The question of whether the test measures what we would like it to can be treated separately. In fact if a study were to show how a criterion C is related to knowledge of the answers to items on the test, this study could presumably be used to derive good values v_i for portions of the test.

REFERENCES

[1] BIRNBAUM, A. (1957). Efficient design and use of tests of a mental ability for various decision-making problems. Technical Report, USAF School of Aviation Medicine Project No. 7755-23, Teachers College, Columbia University, 1–25.

[2] BIRNBAUM, A. (1957). On the estimation of mental ability. Series Report No. 15. USAF School of Aviation Medicine Project No. 7755-23, Teachers College, Columbia University, 1–31.

[3] BIRNBAUM, A. (1957). Further considerations of efficiency in tests of mental ability. Technical Report No. 17, USAF School of Aviation Medicine Project No. 7755-23, Teachers College, Columbia University, 1–39.

[4] CALANDRA, A. (1941). Scoring formulas and probability considerations. *Psychometrika* **6** 1–9.

[5] GULLIKSEN, H. (1950). *Theory of Mental Tests*. Wiley, New York.

[6] HAMILTON, C. H. (1950). Bias and error in multiple choice tests. *Psychometrika* **15** 151–168.

[7] HORST, P. (1933). The difficulty of a multiple choice test item. *J. Educ. Psych.* **24** 229–232.

[8] LORD, F. M. (1953). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika* **18** No. 1 57–76.

[9] LUBIN, A. and OSBURN, H. G. (1957). A theory of pattern analysis for the prediction of a quantitative criterion. *Psychometrika* **22** No. 1 63–73.

[10] LYERLY, S. B. (1951). A note on correcting for chance success in objective tests. *Psychometrika* **16** 21–30.

[11] SITGREAVES, R. (1961). Optimal test design in a special testing situation. 29–45. *Studies in Item Analysis and Prediction*, Ed. H. Solomon. Stanford Univ. Press.

[12] SOLOMON, H. (1955). Item analysis and classification techniques. *Third Berkeley Symp. Math. Stat. and Prob.* **5** 169–184 Ed. J. Neyman. Univ. of California Press.

[13] SOLOMON, H. (1961). *Studies in Item Analysis and Prediction*. Stanford Univ. Press.