II. FORMULAS FOR SCORING TESTS IN WHICH THE MAXIMUM AMOUNT OF CHANCE IS DETERMINED

George Frederick Miller,
Department of Education, University of Oklahoma

The element of chance is involved in all kinds of mental tests. When a child is asked which is the longer of two lines, or which is his left ear he might by chance point to the correct one. If in spelling he does not know whether it is m-a-i-n or m-a-i-n-e, he might get it right or wrong by chance.

In one type of test the amount of guessing that the examinee does is not revealed by either the results or the form of the test. If the question is, What is the capital of Pennsylvania?, the examinee might know it is 1 of 2 names, or he might know it is 1 of 10, 48, or any other number. His finished paper does not show to what extent chance was a factor in his answer. In this type of question, where the maximum amount of guessing is unknown, the only practical scoring is on the basis of the number of correct responses.

In another type of test the maximum number of chances is determined. In the true-false test, for instance, the guessing is limited to 2 items in each question, from which 1 may be chosen. In Test 3 of the Army Alpha Examination, the examinee bably be obtained by chance will be marked 0, and papers that that the examinee does is not revealed obpectively. If his paper correct, may be made. In this type the exact amount of guessing that the examinee does is not revealed objectively. If this paper is perfect, he might have guessed at none, or he might have guessed successfully at several. But his guesses are limited to a certain maximum number, so that they can be considered systematically in arriving at the score.

In the true-false test the common way of marking is to subtract the wrong from the right answers, not counting those questions not marked. This method is expressed by the formula $S = (S_1 - U) - 2 W$, in which S means the score that the examinee makes, S_1 means the maximum score for the test, U means the number of unmarked questions, and W means the number of wrong answers.

When each question contains 3 items from which 1 choice may be made, and only 1 item is the correct answer, what is the best method of scoring? The same principle should be applied in this case that is commonly used for the true-false score so the paper that has only the number right that would promay choose 1 of 3 items in each question and in Tests 7 and 8 of have a higher number of correct responses will be scored by the same formdla, and will be marked proportionately higher. Since in the case of 3 possibilities, in which about 1/3 of them rgiht, the paper with only 1/3 right should be marked 0. The formula for this case is: $S = (S_1 - U)\ 3/2\ W$, in which the letters have the same meanings as above. If 30 questions are marked in such a test, and only 10 of them are correct, the score is 0; if 25 of the 30 are correct, the score is 22.5. If the test is of a similar kind, but provides 1 choice out of 4, the fraction in the formula is 4/3 instead of 3/2. If 1 hoice out 5 is taken the fraction is 5/4, and if 1 out of 6 it is 6/5.

In order to make the formula more general other cases need to be considered. Instead of only one choice benig correct, as is assumed in all of the above cases, the test may be so constructed that two or more choices will be correct; but the same principle stated above holds,—when the number of correct responses is not greater than a chance marking would approximate, the score should be 0; and when the number of correct responses is greater than chance would approximate, a correct score is obtained by using the same formula that gives 0 when the number of correct responses is not greater than chance usually approximates. In cases where more than 1 choice is correct, the number correct must, of course, always be at least 1 less than the number of items from which the choices are made. Furthermore, if the choices permitted are greater than 1 out of 2, the value of the test is needlessly weakened by increasing the element of chance. For example, if 3 items are provided, 2 of which are correct and will count on the score if chosen, and 2 choices are permitted, then 1 of the choices will inevitably be correct. If the score in that test is found by subtracting the wrong answers from the right, the score will have to be either 0 or 2. Another illustration is the test in which 13 items from which to choose are given, 9 choices are permitted, and 9 correct answers are possible. It will be noticed that only 4 of the 9 choices can be incorrect. If the examinee makes all of the mistakes that he possibly can, and his wrong answers are subtracted from his right, he will still have 1 point to his credit. It might be assumed that no tests would ever be constructed allowing a greater proportion of choices than 1 out of 2, and that it is folly to mention such possibilities. But the two examples just mentioned are taken from a printed standard test[1]. The cases mentioned below that contain choices in greater proportions than 1 out of 2, are mentioned only in deriving a general formula, and not to endorse their use.

Applying the above principle to the case of 2 responses out of 3 being correct, 2/3 of the questions marked correctly would be the approximate result of chance, and should be scored 0. The fraction for 3 out of 5 would be 3/5, for 3 out of 6, 3/6, etc. The formulas for these cases taken in the order above, and let-

[1] Frasier, George Willard and Armentrout, Winfield D., "Standard Achievement Test in an Introduction to Education," Test IV.

ting U equal 0; are: $S = S_1 - 3\ W$; $S = S_1 - 5/2\ W$; $S = S_1 - 6/3\ W$; etc. Suppose, to illustrate, that a test contains 30 series of 5 items each, 3 of which, if marked correctly, count on the score, that 36 are wrong, and that U equals 0. Then substituting in the formula gives:

$$S = 90 - 5/2 \times 36, \text{ or } S = 0.$$

The principle of correcting scores for chance has just been applied to two types of series. The first was that 1 choice was allowed from any number of items, only 1 of which could count on the score. The second was that an indefinite number of choices was allowed from an indefinite number of items (the latter always being at least 1 greater than the former) and that the maximum number of items that could count on the score was equal to the number of choices. Two more types need to be considered.

Suppose, to illustrate type 3, that the series is 3 chosen from a group of 5, and not more than 2 of the 5 items can count on the score. The values for these 3 quantities, when the series is divided by 3, are : 1, 5/3, and 2/3, respectively. In this form the first 2 quantities are similar in form to those in the first type explained above, because the first is unity and the second comes within the term "any number of items". The third quantity, the maximum number in the series that can count on the score, has no effect on the fractional part of S_1 that chance marking wiil approximate. It affects only the value of S_1 and does not enter into the product subtracted from $S_1 - U$ in order to find the value of S. Then when the form 3 out of 5 and 2 right is reduced to the terms 1 out of 5/3 and 2/3 right the amount of chance involved in it can be calculated by the same process as the first type referred to before, which was of the form 1 out of 5 and 1 right. Now in the case of 1 chosen from 5, $\frac{5 \times W}{4}$ or $\frac{5 \times W}{5 - 1}$ is the product subtracted from $S_1 - U$. Applying the same principle to the case of 1 out of 5/3, the 5/3 must be substituted for 5, which gives $\frac{5/3 \times W}{5/3 - 1}$. Placing this term in the formula gives:

$$S = (S_1 - U) - \frac{5/3 \times W}{5/3 - 1}$$

The following problem will serve to illustrate this formula A test contains 15 series of 5 items each; 3 items in each series are to be marked, and a maximum of 2 of the 5 items are correct answers and count on the score if correctly marked. If W equals 12, and U equals 0, what is the score? Substituting in the formula gives:

$$S = 30 - \frac{5/3 \times 12}{5/3 - 1}, \text{ or } S = 0$$

A fourth type of series is possible. Suppose that the series is 2 chosen from a group of 5, and correct answers can be made from 3 of the 5 items. It can easily be shown that the element of chance is the same in this type as in the preceding one. About 3/5 of the maximum score would be obtained by chance. The maximum score for the series, 2, would also remain the same. So type 4 is solved by the same process as type 3.

In order to arrive at a general formula for all of the above types, it is necessary to use a symbol for the quotient resulting from the total number of items in the series divided by the number of choices, or in case the fourth type of series is not changed into the third type before solving, this quantity will be the reciprocal of the fraction expressing the part of the score, S_1, that chance marking will approximate. Let n represent that quantity[1]. Then the general formula becomes:

$$S = (S_1 - U) - \frac{nW}{n - 1} \qquad \text{I.}$$

The question that may arise is, What is the use of scoring by the principles of Formula I? This question is partly answered by an application of the formula to scoring some standard tests.

1. In the Army Alpha Examination where 1 choice out of 4 is provided as in Test 8, and only 1 of the 4 is correct, suppose that 4 questions are unmarked, and that 27 of the remaining 36 are wrong, what should be the score? Substituting in the formula we have:

$$S = (40 - 4) - \frac{4 \times 27}{4 - 1} \text{ or } S = 0$$

The authorized scoring in this case gives a score of 9 points, which means that the examinee probably gets 9 points more than he deserves.

2. The next illustration is taken from Test 3 of a "Standard Achievement Test on an Introduction to Education",

which was mentioned above. In this test 8 items in column A are to be paired with 8 in column B. Each item can be paired in only 1 way. If only 5 correct groupings are made, and U is 0, the formula gives:

$$S = 8 - \frac{8 \text{ X } 3}{8 - 1} \text{ or } S = 4\ 4/7$$

Гhe instructions given by the authors of that examination for scoring this test (R-W), would give a score of 2, when only 5 correct answers are made and no question left unmarked.

A further reason for correcting scores for probable chance marking is shown by the data that follow:

Table I

An illustration of the di‹ference in the use of the formula and the usual method of calculating scores, in the case of tests in which the maximum amount of chance is known.

				Scores from					
No. of tests	**Army Alpha Group Ex-amination**			Otis Group Intelligence Sca e, Form A By			Haggerty Intelligence Examination, Delta 2		
	Chance	Actual trial	Formula and trial	Chance	Actual trial	Formula and trial	Chance	Actual trial	Formula and trial
2	—	—	—	5	6	1	—	—	—
3	5	5	0	—	—	—	—	—	—
5	—	—	—	—	—	—	5	6	1
6	—	—	—	—	—	—	10	11	1
7	10	17	0	5	5	0	—	—	—
8	10	15	5	4	1	0	—	—	—
9	—	—	—	8	7	0	—	—	—
10	—	—	—	10	9	0	—	—	—
Totals	25	27	5	32	28	1	15	17	2

Explanations: 1. "Chance" here means the score that would probab'y be made by chance. In the Army Alpha Examination,

*The formula may be simplified by letting *n* be reciprocal of the fractional part of the score, $S_{,}$ that will be *wrong* by chance marking. The formula then becomes,

$$S = (S_{,} - U) - nW.$$

Test 3 there are 16 questions, and the choice is 1 out of 3 in each question. So the probable score by chance is 5 1/3.

2. "Actual Trial" means that a chance marking was actually made for the tests. Cards were used. In the case of Test 3 just mentioned, 16 cards were numbered 1, 16 were numbered 2, and 16, were numbered 3. They were shuffled, and 16 o. them drawn. The items were marked in the order in which the cards were drawn.

3. The score made by the actual trial was used in the formula to correct it for chance. I. the chance scores were used in the formula they, of course, would all give 0.

It is very significant to notice some of the results obtained from some intelligence tests when no compensation for chance is made. The Army Alpha Examination, for example, gives a score o. 25 by pure chance (Table I). In practice the grade of C— was given for scores from 25 to 44 on that examination, and the grades of D and D— for lower than 25. Now it is evident that score 25 if made only on tests 3, 7, and 8 is the probable 0 mark. and the lowest C— individual, as well as all below him, were of such low grade intelligence that they could not be measured by the test and the method of scoring. Scores of 25 and lower may be due entirely to chance, and have no significance, unless they are made on tests other than 3, 7 and 8; or unless credit for intelligence is given for merely holding a pencil and marking in certain places at random in tests 3, 7, and 8.

The Army Examination was the forerunner of numerous intelligence examinations of a similar kind, which followed the same method of scoring—right answers minus wrong in tests where the choice of 1 out 2 was given, and counting on the score all those right in all other tests.

Another instance from Table 1 may be taken from the Otis Group Intelligence Scale, in which chance marking gives a score of 32. This mark of 32 means, according to the norms worked out for the test, a Binet mental age of 8.6 years. If compensation is made for chance, the score of 32, if made entirely on Tests 2, 7, 8, 9 and 10, becomes 0. It would be absurd to suppose that all children who make 32 or lower on the test have 0 intelligence. They might make some of the score on Tests other than 2, 7, 8, 9 and 10, or have intelligence that the examination does not measure. Because a 15 pound child weighs 0 on scales used for weighing freight cars and graduated in no smaller units than 100 pounds, does not prove that the child has no weight.

In case only the relative standing of a group of individuals is desired, and a single test is used, the whole number of right answers used as the score serves the purpose as well as correcting the scores for chance. In a true-false test of 50 questions, for example, if the correct scores run from 25 to 50, the person who makes 25 is at the bottom of the list, whether his score is recorded as 25 or 0; and all of the others remain in the same relative positions but not separated by the same number of points, regardless of whether the formula is used or not. But if scores are obtained from various tests, unlike in the probable effect of chance on the scores, the importance of adjusting the scores to correct them for chance becomes evident. If scores are not corrected for chance, a score of 50 may be made by one person on tests where chance, is practially eliminated and by another on tests where 1/3 of the score may be due to chance The same scores of 50 for each would indicate very different amounts of intelligence.

The distorted relative values obtained from the usual method of scoring is further illustrated by the scores made by three individuals on the Army Alpha Examination.

Table II

Actual records of three examinees on the Army Alpha Examination, each scored in two ways, by the formula I and according to the instructions in the manual.

	Individuals					
	A		B		C	
No. of Tests	Regular scores	Scores by formula	Regular scores	Scores by formula	Regular scores	Scores by formula
1	9	9	3	3	9	9
2	12	12	8	8	2	2
3	15	15	5	5	4	3
4	38	38	11	11	6	6
5	20	20	5	5	7	7
6	13	13	4	4	7	7
7	37	36	2	0	7	4
8	23	17	17	13	3	0
Totals	167	160	55	49	45	38
Percentage of loss due to formula marking	4		10		16	

These meager data indicate that the lower the score the greater the percentage due to chance. By correcting scores for probable chance marking, the bright are separated from the dull more distinctly, and the effectiveness of the test is improved.

Scoring according to Formula I is the universal practice for tests in which there is a choice of 1 out of 2 (the true-false test), but for tests in which the choice is other than 1 out of 2, the general practice is to count the number right as the score. If chance is recognized in calculating the score in tests where the choice is 1 out of 2, why should it not be considered in determining the score when the choice is 1 out of 3, 1 out of 4, 2 out of 5, and in other cases where the maximum amount of chance is known? It is evident that the advantage from guessing is greater where there are only 2 items from which 1 is to be chosen, than where there are three or more from which 1 is to be chosen. The greater the number from which a choice is made the less probability there is of guessing the right answer. If the number is about 50 or more, as in paired vocabularies, the total right is about the same as the score computed on the basis of the above principle. If, for instance, the test contains 50 questions, and only 25 of the answers are correct the score is 24.5; but if only 25 of 50 questions are correct in a true-false test, the score is 0.

Conditions sometimes arise in tests with a fixed number of choices and a fixed number of items from which to choose that are not provided for by Formula I. If the number of choices that the examinee actually makes is greater than the number allowed, how should the score be calculated? A deduction in the score proportionate to the advantage gained by extra choices should be made. If the total number of items in a test equals 15, the number allowed to be chosen is 5, and W equals 0, it is evident that if 15 choices are made the score would be 0. Also if the choices actually made are only 5, and W is 0, the score would be perfect, or 5. The range between 15 and 5 points is 10. If the examinee marks extra points to the extent of all the range, he is penalized 5 points, or 5/10 of a point for 1 unit of the range. If the number actually marked is 7 the amount deducted would be 5/10 (7-5) or 1. Expressed in the form of an equatoin the process of deduction from the score is: S = 5 — 5/10 (7-5). If n_1 is the number of markings allowed in a series, if N is the sum of the number of markings made in excess of n_1, for each series of the test, and if n_2 is the total number of items in a series, then the equation in general terms is:

$$S = S_1 - \frac{n_1 N}{n_2 - n_1} \qquad \text{II.}$$

This formula is to be used for cases in which W and U both are 0. If W and U do not equal 0, Formula I and Formula II must be combined, which gives:

$$S = (S_1 - U) - \frac{nW}{n - 1} - \frac{n_1 N}{n_2 - n_1} \qquad \text{III.}$$

Elements may be omitted from Formula III to meet various needs. If U equals 0, U is omitted; if W equals 0, $\frac{nW}{n - 1}$ is omitted; and if N equals 0, $\frac{n_1 N}{n_2 - n_1}$ is omitted.

For a simple application of Formula III, suppose that 30 questions are given in a true-false test, that 40 of the 60 items from which choices may be made are marked instead of the 30 allowed, and that 5 are marked incorrect'y. S_1 will then equal 30, $U = 0$, $n = 2$, $W = 5$, $n_1 = 30$, $N = 10$, and $n_2 = 2$. Substituting in Formula III gives:

$$S = 30 - \frac{2 \times 5}{2 - 1} - \frac{1 \times 10}{2 - 1} \text{ or } S = 10$$

This score is the same as the one obtained by the common practice of scoring the true-false test by counting all questions that are marked both ways as unmarked, and subtracting the wrong answers from the right. If all tests in which the maximum amount of chance is determined were the simple true-false kind, the above formulas would not be necessary. When, however, the series contain many items, some of which are omitted by the examinee, some over-marked, some marked correctly and some incorrectly; when several choices are permitted in 1 question, or when 1 choice is permitted from several correct items; and when several questions are included in a series; the problem becomes tco complex to be solved by simple inspection. The following scheme will indicate the method of finding the values of U, W, and N.

Explanations

1. The whole plan represents a test of 11 series, with 7 items in each series. From each series of 7 the examinee is to choose 2.

2. IR means an item that if chosen, or marked, will count 1 point on the score, provided the whole number chosen in the

series does not exceed 2, in which case the additional IR marked counts 1 overmarking (1 unit on the value o: N.)

3. I is used to represent items other than IR. If an I is chosen, or marked, it will count: (a) wrong if the number of marks for the series does not exceed the number allowed by thc directions for making the test, or 2; (b) overmarked, if the number of marks for the series exceeds the number allowed by the directions for marking the test.

4. The directions to the examinee for marking the test are: "Encircle 2 correct items (words, figures, symbols, etc.) out of the seven in each of the series."

5. Of course, in practice the letters IR and I would be replaced by meaningful items preceded by a statement, as: Mark 2 of the :ollowing that are usually liquids: map, fish, ink, country, milk, water, putty.

No. of				Items				Values of		
series								U	W	N
1	I	(I)	IR	I	(IR)	IR	I	0	1	0
2	I	I	(I)	I	IR	IR	IR	1	1	0
3	IR	(I)	I	I	I	(IR)	(IR)	0	0	1
4	(IR)	I	I	IR	IR	I	I	0	0	0
5	I	IR	(I)	(IR)	(I)	IR	I	0	1	1
6	I	IR	I	I	I	IR	(IR)	0	1	0
7	(IR)	I	(IR)	(IR)	I	I	I	0	0	1
8	I	I	I	I	(IR)	IR	(IR)	0	0	0
9	I	(I)	(IR)	(I)	IR	IR	I	0	1	1
10	I	(IR)	IR	I	IR	I	I	1	0	0
11	(IR)	(IR)	(IR)	I	I	I	I	0	0	1
Totals								2	5	5

In addition to the values o: U, W, and N expressed as totals in the scheme, values for the other letters of Formula III when applied to the scheme are: $S_1 = 22$, $n = 7/3$, $n_1 = 2$, and $n_2 = 7$. Substituting the values in Formula IIII makes:

$$S = (22 - 2) - \frac{7/3 \times 5}{7/3 - 1} - \frac{2 \times 5}{7 - 2} \text{ or } S = 9\ 1/4$$

Confusion in the understanding and use of the above formulas will be prevented by keeping in mind the following definitions of the terms and symbols used in this article.

(A) Terms

1 Item means the word, figure, or other symbol that forms the smallest part of an examination. For example, Test 8 of the Army Alpha Examination contains 160 items, 4 in each of the (40) questions.

2. Question means an item that counts on the score i: correctly marked; or the question is a group of items only one of which counts on the score. In the scheme of 11 series given just above there are 22 questions, 2 for each series.

3. A series is a group of items that are marked as a unit. The different series in various tests do not have a fixed number

of items, but all the series in a given test must be uniform in the total number of items each contains, in the number of choices permitted, in the number of items that count on the score, and in the number of items from which correct choices can be made. Tests are usually so constructed that a series contains only 1 item that counts on the score, only 1 that may be chosen, only 1 that can serve as a correct answer, and 2, 3, 4, or 5, items from which a choice is made. In case not more than 1 item in the series can count on the score, the series means the same as a question.

4. Test means a number of uniform series, provided it is a test in which the maximum amount of chance can be determined. Test is also used to designate various other kinds of groups of questions. Test 1 of the Army Alpha Examination is an example of one of the other kinds.

5. Examination means a group of tests. It is the largest unit considered in this article. In printed form it si the whole booklet or folder. Scale is sometimes used in this sense, as "Otis Group Intelligence Scale." Test is also used in this sense, as, "Standard Achievement Test on an Introduction to Education." The last named contains 4 major divisions, each of them called also "Tests."

(B) Symbols

Note:—The letters used are of two kinds: (1) Capitals, which designate values for the test as a whole, and (except S_1) whose values generally cannot be found by multiplying the number of series by some other factor. (2) Small letters, which designate values for each series of the test, and are constant for each test.

1. S means the actual score that an examinee makes on any test.

2. S_1 means the total number of items of a test that can count on the score. S_1 is the maximum score for a test, and is equal to the number of questions in the test.

3. U means the number of items in a test that count on the score if marked, minus the number that is marked. To find the value of U each series of items must be checked separately, and the number unmarked in each series added to make U.

4. W means the number of items that count on the score that are not marked, and instead of which some other item is marked. The value of W must be found by checking each series separately as described for U.

5. N means the sum of the number of markings made in excess of n_1 for each series of the test.

6. n means the reciprocal of the fraction that expresses the part of the score (either for a series or the test) that chance marking will approximate.

7. n_1 means the number of items in a series permitted to be marked by the directions for marking the test.

8. n_2 means the total number of items in a series.

Conclusions

1. The current practice is scoring tests in which the maximum amount of chance is known, by counting the number of right answers as the score (except in the true-false test), contains a fault that can readily be corrected.

2. The correction can be made in most instances (in which a choice of 1 out of 3, 4, 5, and the like items is allowed) by Formula I, which is:

$$S = (S_1 - U) - \frac{nW}{n - 1}$$

3. Formula III is a general one, which will serve where other corrections are desired. This formula is:

$$S = (S_1 - U) - \frac{nw}{n - 1} - \frac{n_1 N}{n_2 - n_1}$$