REPLY

THE VALIDITY AND RELIABILITY OF GROUP JUDGMENTS

BY H. J. EYSENCK

(From the Psychological Laboratory, University College, London)

In a recent paper I attempted to show that the validity of æsthetic judgments increases as the number of judges increases, in accordance with a formula first given by Burt for reliability or self-consistency:

$$r_{\bar{k}g} = \sqrt{\frac{n\bar{r}_{kk'}}{1 + (n - 1)\bar{r}_{kk'}}}, \qquad (1)$$

where $r_{\bar{k}g}$ stands for the correlation of the average order with the 'true order,' $\bar{r}_{kk'}$ for the average intercorrelation, and n for the number of persons correlated. Using 900 preference-rankings of 12 uncolored pictures, I first showed that the average order of a 'standard' or 'criterion' group of 700 subjects correlated perfectly with the average order of an 'experimental' group of 200 subjects. Then, taking the 200 rankings singly and in groups of 5, 10, 20, and 50, I showed that as the size of the group increased, so the correlation of their average order with the true order, *i.e.* the order given by the 'standard group,' also increased as predicted by the formula (1).

This demonstration has been criticized by Mr. Babington Smith on four main grounds. He maintains (1) that the equation used was only an approximation, and that the full formula should have been used; (2) that the approximate equation depends on the applicability of the two-factor theory, and that this theory was not shown to apply; (3) that in any case the formula deals with reliability, and not with validity; and (4) that the results of certain experiments described by him disprove a claim he attributes to me, namely that "a measure shown to be reliable will in the long run be perfectly valid." Points (1) and (2) will be dealt with together, as they are really inseparable; points (3) and (4) also belong together logically. One or two further criticisms of minor importance will be dealt with in the course of the argument.[1]

It is of course quite true that the formula used is only an approximation, the full formula being given by Burt as:

$$r_{\bar{k}g} = \frac{n\bar{r}_{kg}}{\sqrt{n + n(n - 1)\bar{r}_{kk'}}}, \qquad (2)$$

which reduces to (1) since

$$\bar{r}_{kg} = \sqrt{\bar{r}_{kk'}} \text{ (approximately).} \qquad (3)$$

Now

$$\bar{r}_{kg} = \frac{\bar{r}_{kk} + (n - 1)\bar{r}_{kk'}}{\sqrt{n\bar{r}_{kk} + n(n - 1)\bar{r}_{kk'}}}, \qquad (4)$$

[1] My thanks are due to Mr. Babington Smith for his courtesy in letting me see his MS. before publication. I also want to record my gratitude to Professor C. Burt for kindly reading through and commenting upon the draft of this paper.

where $\bar{r}_{kk}$ is the communality ($\bar{r}_{kk} = \overline{r^2_{kg}} + \overline{r^2_{kg'}} + \overline{r^2_{kg''}} \cdots$). Equation (4) reduces to

$$\bar{r}_{kg} = \sqrt{\bar{r}_{kk'} + \frac{\bar{r}_{kk} - \bar{r}_{kk'}}{n}} \tag{5}$$

and from equation (5) it is apparent that the approximation is due to the fact that $\bar{r}_{kk} \neq$ but $> \bar{r}_{kk'}$. Substituting equation (5) in equation (2), we get

$$r_{\bar{k}g} = \sqrt{\frac{\bar{r}_{kk} + (n-1)\bar{r}_{kk'}}{1 + (n-1)\bar{r}_{kk'}}}. \tag{6}$$

This formula enables us to form an idea of the size of the error introduced through the fact that an approximation formula only is used, and that more than one factor may be present. As Burt points out: "Since analysis by multiple factors is a process of averaging deviations about preceding averages, the range of the correlations is reduced to rather more than half at each step; the treatment of the first factor as exclusively positive, instead of bipolar, produces the effect of missing a step" (2, p. 358). Hence we will not go far wrong if we assume that in the great majority of cases $\bar{r}_{kk} \leqslant 3/2\overline{r^2_{kg}}$; particularly as the latter terms in the progression indicated by Burt ($1 + \frac{1}{4} + \frac{1}{8} \cdots$) are very likely to be omitted because of lack of statistical significance (2, p. 357). (As Davies has shown in an analysis of all published tables of correlations between persons, in only 4 out of 48 researches was even the second factor statistically significant, while the third factor never reached the required level of significance (3).) We must also take into account the fact that $\overline{r^2_{kg}} > \bar{r}_{kk'}$ (if only slightly), but on the whole we may perhaps reasonably assume that in actual practice the value of $\bar{r}_{kk}$ will seldom if ever be larger than $2\bar{r}_{kk'}$, and in the majority of cases a good deal smaller.[2]

Taking $\bar{r}_{kk} = 2\bar{r}_{kk'}$ as the probable maximum value, and substituting this in equation (6), we get

$$r_{\bar{k}g} = \sqrt{\frac{(n+1)\bar{r}_{kk'}}{1 + (n-1)\bar{r}_{kk'}}} \tag{7}$$

and hence the probable maximum error in taking $\bar{r}_{kg} = \sqrt{\bar{r}_{kk'}}$ is

$$\sqrt{\frac{(n+1)\bar{r}_{kk'}}{1 + (n-1)\bar{r}_{kk'}}} - \sqrt{\frac{n\bar{r}_{kk'}}{1 + (n-1)\bar{r}_{kk'}}}, \tag{8}$$

[2] The value actually occurring in the experiment criticized is of course much smaller than this. As evidence for this contention we may cite the fact that the errors actually observed (*i.e.* the differences between the theoretical and actual values in Fig. 1 in my original article) are not even half the maximum error. Seeing that the inaccuracy introduced through the use of the approximate formula and through the presence of any second, third, etc. factors only influences the result by changing the values in the leading diagonal of the correlation matrix, we may remember Thurstone's observation that "the diagonal entries . . . may be given *any value between zero and unity without affecting the results markedly*, especially when the number of variables is large" (4, p. 108. My italics).

which reduces to

$$\text{Max. Error} = \sqrt{\frac{\bar{r}_{kk'}[(2n+1) - 2\sqrt{n^2+n}]}{1+(n-1)\bar{r}_{kk'}}}. \tag{9}$$

From equation (9) it is then possible, provided we know the average intercorrelation, to calculate the probable maximum error which can be introduced by using the approximate equation (1) instead of the full formula (2).

If we have the $r_{\bar{k}g}$ value calculated by the approximate equation, then equation (9) can be rewritten in the following form which will be more convenient for ascertaining the maximum amount by which the value of $r_{\bar{k}g}$ from the approximate equation is likely to be increased:

$$\text{Max. Error} = r_{\bar{k}g}\sqrt{\frac{2n+1}{n} - 2\sqrt{\frac{n+1}{n}}}. \tag{10}$$

In Table 1, below, are given the maximum errors so calculated for various numbers of persons and for various values of $r_{\bar{k}g}$, as calculated by the approximate formula. This table may be of interest as showing the maximum amount of error to which anyone using Table 2 in my original paper would be liable in the ordinary course of investigation.

TABLE 1

$r_{\bar{k}g}$:	Number of Rankings:				
	5	10	20	50	200
.10	.01	.00	.00	.00	.00
.20	.02	.01	.00	.00	.00
.30	.03	.01	.01	.00	.00
.40	.04	.02	.01	.00	.00
.50	.05	.02	.01	.00	.00
.60	.06	.03	.01	.01	.00
.70	.07	.03	.02	.01	.00
.80	.08	.04	.02	.01	.00
.90	.09	.04	.02	.01	.00

There are several points which deserve mention in connection with the preceding argument. The first is that the maximum error introduced by using the approximate formula is small (except when only 5 subjects are used; but *cf.* below). In the analysis of results such as one is likely to obtain in the ordinary course of investigation the actual errors found are of the order .01. Thus in five experiments carried out by the present writer in an endeavor to strengthen and make statistically significant the second factor (thus increasing the amount of error over the usual), the following results were reached on the average:

TABLE 2

$\overline{r^2}_{kg} = .382$	$\bar{r}_{kk} = .494$	$r_{\bar{k}g} = .942$ (approximate formula)
$\overline{r^2}_{kg'} = .112$	$\bar{r}_{kk'} = .343$	$r_{\bar{k}g} = .955$ (full formula)
	$N = 15$	Error = .013

Although the error involved in using an approximate formula may only be small, theoretically of course it is always preferable to use the exact formula. In practice, however, one must compare the gain in accuracy with the amount of extra work involved in the use of the full formula.

In the case of the experiment criticized by Mr. Babington Smith, it would have been necessary to calculate 19,900 correlations, and to factorize, by a special iterative procedure, 40 tables containing 10 correlations each, 20 tables containing 45 correlations each, 10 tables containing 190 correlations each, 4 tables containing 1,225 correlations each, and 1 table containing 19,900 correlations.[3] It must be left to the reader to judge whether the possible gain in accuracy would have justified the amount of work required.

The true position seems to be this: the errors involved in using the approximation are appreciable only when n is very small. But rather than attempting to derive accurate data from such very small samples by means of refined statistical procedure—an illusory accuracy in any case, because of sampling errors—it would be better to increase the number of subjects, and thus decrease the size of the error involved in using the approximate formula. The number of subjects need not be unmanageably large to achieve this object—as Table 2 shows, even with 15 people the average error involved was only .01 in 5 experiments. We would agree with Mr. Babington Smith, however, that in cases where for some reason or other only few subjects were available, or in cases of particular theoretical interest, where accuracy is of supreme importance, the exact formula should be used, rather than the approximation.

Two minor points in this connection may be of interest. The errors introduced through the use of the approximate formula are on the conservative side; they underestimate the correlation with the 'true order.' Thus when Table 2 of my original paper is used, as I suggested, to "enable the investigator to judge at a glance which way the results are tending, or how many subjects to use," he will always be on the safe side by following the guidance of the table implicitly. Secondly, it would have been quite impossible to construct a table of this kind by using the full formula, as the full formula contains too many unknowns.

The second major point in this discussion is connected with Mr. Babington Smith's second criticism. It will be clear to those who have followed the mathematical argument that any factor other than the first, general, factor can influence the result only by increasing the value of $\bar{r}_{kk'}$, *i.e.* by increasing (very slightly) the amount of inaccuracy introduced through the use of the approximate formula. (This effect has been taken into account in equation (7) and the subsequent discussion.) That means that if the two-factor theory does not hold, our results are affected quantitatively;

[3] Mr. Babington Smith, in a private communication, has objected to this argument because he maintains that by following his procedure it is unnecessary to calculate all these correlations. But that is not really material to the argument; as presented by Burt, the full formula does require calculation of these values, and later developments were not available when I made my decision as to which of his two formulae to use. It is this decision which Mr. Babington Smith has criticised.

Mr. Babington Smith seems to suggest that it affects the result qualitatively, by making the formula altogether inapplicable.

Thus we conclude that the formula used was an approximation, and as an *approximate* formula applicable whether or not the two-factor theory holds. It is admitted that when very few subjects are used in an experiment, it is advisable to use the full formula rather than the approximation, although even then it would seem preferable to increase the number of subjects rather than to attempt to reach a rather spurious accuracy by means of fuller statistical calculations. As the number of subjects increases, however, the errors involved decrease rapidly whether the two-factor theory holds or not; hence when dealing with 15 or more subjects the use of the approximate formula would seem quite justified, especially as the errors introduced are on the conservative side.

Mr. Babington Smith's third criticism raises a point which is even more important than those already considered, and will probably be of wider interest. He maintains that "Mr. Eysenck's formula deals with reliability and not with validity," and appears to consider unfounded my claim to have shown that the *validity* of æsthetic judgments increases as the number of judges increases.

We can meet this criticism in two different ways, firstly as applied to our experiment, and secondly, as applied to the general case. Validity is defined as the correlation of some test with an accepted outside criterion, and although such a criterion was provided in our experiment that would not always be so. In the case of our experiment, the average order produced by the 'standard' group of 700 subjects was the criterion against which the orders of the 200 subjects in the 'experimental' group, singly and in groups, were validated. This is what Hull would call a 'subjective judgment criterion,' which according to him takes its place beside the 'product' and the 'action' criterion (5, p. 375). The critic may possibly doubt the soundness of this criterion; as Guilford points out "students of æsthetics especially are often of the opinion that the combined judgments of the masses should count for little as compared with the judgments of a single 'expert' " (6, p. 259). It seems, however, that such a criticism is hardly tenable in the face of the results of experiments reported by Dewar (7), Bulley (8), Semeonoff (9) and others who found a very strong tendency for the average judgments of large numbers of laymen, even of very young children, to agree perfectly with the judgments of experts. (Fechner also comments upon this effect.) In terms of the theory of the general factor of æsthetic appreciation, as outlined by Burt (10), we would say that the expert is more highly saturated with this general factor than the layman is likely to be, but that we are dealing with the same factor in both cases. Hence the judgment of a group cf 700 people may be regarded as at least as good as that of a group cf 10 experts, say, and probably as better.

But as pointed out above, this case is rather exceptional, and in general such an outside criterion will not be available. It is possible even then, however, to argue with Guilford that "the determination of the validity of test items . . . may be carried out *without the use of an outside criterion.*" As he points out, "Especially with personality tests, for which it is difficult to find a valid objective criterion, it has been customary to *let the test become its own criterion.*" (6, p. 451. My italics.) That is precisely what would be done in the ordinary case to which I would suggest applying Burt's formula: *the criterion would be derived from the judgments themselves.* In a precisely analogous fashion, 'g' was derived from the intercorrelations of so-called tests of intelligence. It is then open to us to identify 'g' with a term of common reference, such as 'intelligence,' just as more recently 'T,' the general factor of æsthetic judgment, has been identified with 'good taste' (11). In this way we arrive at an exact, operational definition of such vague everyday terms as 'beauty' and 'intelligence' which in their ordinary form are useless for scientific discussion. This suggested use of the word 'validity' has good authority on its side, and seems to avoid both the extreme views advanced in this sphere: the ultra-conservative view of Babington Smith, who would restrict the use of the term unduly, and the rather too revolutionary view of Carr, who would abolish all differentiation between reliability and validity (12).

Guilford goes on to say that the chief criticism against this procedure, as applied to personality tests, is that we do not know what the real dimensions or variables of personality are. Such a criticism can hardly be levelled against this procedure in the field of æsthetics, as due largely to the work of Professor Burt and his students the fundamental bases of æsthetic judgments are no longer unknown. (In the field of temperament, the position has also improved considerably, due mainly to the important work done by Guilford himself in his search for Personality Factors.) As regards æsthetics, it would appear that first of all we are dealing with a general factor which enters into our judgments of painting, poetry, sculpture, designs, music, prose, photography, colors, and even odors (10, 11, 13). Secondly, we deal with bipolar 'type' factors which seem to be closely connected with emotional and temperamental characteristics (14, 15, 16). Apart from these factors which go to make up the 'communality' of each person's judgment we have the 'unique' specific and error factors which characterize one person at all times (specific) and one person at any one particular time (errors). Associations, familiarity, etc., are such specific factors; mood, environment, etc., are error factors.

Of all these factors, it is only the general factor that is of importance for our argument; the other factors will cancel out in the long run, as Fechner pointed out long ago (*Vorschule*, 1, p. 194). In attempting to derive the criterion from the judgments themselves, we are trying to do *implicitly* what has been attempted elsewhere *explicitly* (17): to make a psychological analysis of the bases of judgment actually employed by the subjects, *i.e.* of the general factor. Such an attempt is similar in its aim to Professor Spearman's famous analysis of 'g' into his neogenetic laws, although it cannot of course be claimed that it has advanced as far towards its goal.

This leads us to the last point made by Mr. Babington Smith. He maintains that "when Mr. Eysenck's claim is restated in the form that 'a measure shown to be reliable will in the long run be perfectly valid' it is more easily seen to have over-reached itself," and goes on to claim that "examples are given . . . where there is a significant and even high degree of agreement between judges with respect to rankings the sum of which differs from the criterion." Now properly qualified Mr. Babington Smith's statement is a correct interpretation of my view. In my original paper I said that "one great difficulty in experimentation of this kind is that the criterion . . . is not given externally, as in the case of the weights, but has to be deduced from the experimental data themselves" (1, p. 650). This qualifies the general statement against which Mr. Babington Smith argues, and restricts its application merely to cases in which the criterion has to be deduced from the experimental data themselves. In cases of this kind it is indeed true that reliability and validity do come to the same thing, hence in this restricted sense I agree with Mr. Babington Smith's statement. But all his experiments do contain an external criterion, such as the weight of the containers, or the I.Q. of the person whose picture is judged for his intelligence, and hence it cannot be admitted that they have any bearing on the discussion, however interesting they may be in themselves.

Mr. Babington Smith has objected to this view on the grounds that until the exact weights (in his weight lifting experiment) are known the case is exactly similar to those cases in which there is no external criterion, and that therefore the suggested differentiation must break down. That is not correct. We deal roughly with two different classes of judgments: On the one side, we have those where an 'objective,' outside criterion is either available or possible; judgments of weights would be included here. On the other side, we have judgments where the criterion can only consist, or be derived from, the judgments themselves; judgments of beauty would come under this head. For if, with St. Thomas Aquinas and most modern psychologists, we define the beautiful as "Id cuius ipsa apprehensio placet," the *apprehensio* is clearly the only criterion which we can have, by the very nature of the case. This *apprehensio* is expressed as a judgment, and a study of these judgments is the only possible way of studying the beautiful. We must of course take care in our experiments that it is really the *ipsa apprehensio* which pleases, and not some outside effect, such as prestige value ("What a lovely picture—it must have cost a lot of money"). Thus the difference between the two classes lies in the *possibility*, not in the *availability* of an external criterion.

The same argument applies to Mr. Babington Smith's contention that his experiments prove that "errors in human estimates do not necessarily cancel out." This is not a novel statement, and it does not contradict what I myself have said. My claim was that *chance* errors tended to cancel out in the long run; Mr. Babington Smith is dealing with *systematic* errors. A systematic error in factor analysis would form part of the communality of the test (or of the person, when persons are correlated); chance errors would form part of the uniqueness of the test (or the person). But while a systematic error would indeed be an error when we are dealing with an outside criterion, we can hardly regard it as in any sense an error when dealing with something that has no outside criterion. There it would be of the nature of a 'type'-factor (15) or of a 'group'-factor (18); nobody would regard the verbal group-factor as a systematic *error*, for instance!

Thus the only *errors* we deal with in our analysis are chance errors, *i.e.* errors which cancel out by definition.

It would be interesting to follow up further the very suggestive criticisms brought forward by Mr. Babington Smith, but we must forbear doing so, and can only refer the reader to Burt's extensive discussion of these and similar points (10). Mr. Babington Smith's objections to our view seem to be based fundamentally on a philosophic view which regards æsthetic judgments as subjective; indeed, Mr. Babington Smith would extend this view to 'most other judgments.' As this is a philosophic question, there can be little gain in pursuing it any further here; in fact, like so many other philosophic questions, it would appear to be mainly a question of definition, not of fact. On the facts, there can, I think, be little dispute, and it is to be hoped that agreement may be reached on this basis.

(Manuscript received April 9, 1941)

References

1. Eysenck, H. J., The validity of judgments as a function of the number of judges, *J. exp. Psychol.*, 1939, **25**, 650–654.
2. Burt, C., *The Factors of the Mind*, London: Univ. of London Press, 1940.
3. Davies, M., The general factor in correlations between persons, *Brit. J. Psychol.*, 1939, **29**, 404–421.
4. Thurstone, L. L., *Primary Mental Abilities*, Chicago: University of Chicago Press, 1939.
5. Hull, C. L., *Aptitude Testing*, World Book Company, Yonkers, 1928.
6. Guilford, J. P., *Psychometric Methods*, New York: McGraw-Hill Book Co., 1936.
7. Dewar, H., A comparison of tests of artistic appreciation, *Brit. J. educ. Psychol.*, 1938, **8**, 29–49.
8. Bulley, M., *Have you good taste?*, London: Methuen, 1933.
9. Semeonoff, B., Further developments in a new approach to the testing of musical ability, *Brit. J. Psychol.*, 1940, **31**, 145–161.
10. Burt, C., The psychology of art (In: *How the Mind Works*, by C. Burt *et al.*), London: Allen & Unwin, 1933.
11. Eysenck, H. J., The general factor in æsthetic judgments, *Brit. J. Psychol.*, 1940, **31**, 94–102.
12. Carr, H. A., The reliability vs. the validity of test scores, *Psychol. Rev.*, 1938, **45**, 435–440.
13. Williams, E. D., *et al.*, Tests of literary appreciation, *Brit. J. educ. Psychol.*, 1938, **8**, 265–284.
14. Burt, C., The factorial analysis of emotional traits, *Char. and Person.*, 1939, **7**, 238–254; 285–299.
15. Eysenck, H. J., 'Type'-factors in æsthetic judgments, *Brit. J. Psychol.*, 1940, **31**, 262–270.
16. Eysenck, H. J., Some factors in the appreciation of poetry, and their relation to temperamental qualities, *Char. and Person.*, 1940, **9**, 160–167.
17. Eysenck, H. J., The empirical determination of an æsthetic formula, *Psychol. Rev.*, 1941, **48**, 83–92.
18. Eysenck, H. J., Critical notice of Primary Mental Abilities, by L. L. Thurstone, *Brit. J. educ. Psychol.*, 1939, **9**, 270–275.