*Philippine Journal of Psychology*
1969, Vol. 2, No. 1, 34—38

# THE MEASUREMENT OF CONSENSUS IN PAIR-COMPARISON STUDIES[1]

LEONEL CAMPOS
*Ateneo de Manila University*

It is suggested that Kendall's coefficient of agreement, $u$, is not an appropriate measure of consensus in pair-comparison studies. A new index of consensus, $M(c)$, is described. It is further suggested that Kendall's own adaptation of the $X^2$ may be used, with minor modifications, to test the significance of a departure of $M(c)$ from a value of zero.

Kendall (1962) describes a coefficient of agreement, $u$, to be used in the context of pair-comparison studies. His formula may be written,[2]

$$u = \frac{2\sum_{i,j}\left[\binom{a_{ij}}{2}+\binom{a_{ij}}{2}\right]}{\binom{k}{2}\binom{n}{2}} - 1 \quad (1)$$

where

$a_{ij}$, the number of observers who, for any given comparison between an object, or stimulus $s_i$ and a second object, or stimulus $s_j$, decide that $s_i > s_j$;

$a_{ji}$, the number of observers who, for any individual comparison between $s_i$ and $s_j$, decide that $s_j > s_i$;

$n$, the total number of observers, i.e., $n = a_{ij} + a_{ji}$;

$k$, the number of objects, or stimuli, under consideration

$\sum_{i,j}$, the adding operation over all the comparisons between $s_i$ and $s_j$, and,

$\binom{a_{ij}}{2}, \binom{a_{ji}}{2}, \binom{k}{2}, \binom{n}{2}$

the combinations of $a_{ij}$, $a_{ji}$, $n$, and $k$, taken 2 observations at a time, respectively.

Formula (1) suffers from several deficiencies.

1. It yields, $-1 \leq u \leq 1$. That is, if $u > 0$, positive, it indicates agreement, and, if $u = 1.00$, complete unanimity is presumed to exist; if $u = 0.00$, it indicates "lack of agreement", and if $u < 0$, negative, it suggests "disagreement." The problem here is one of interpretation. What is the difference between "lack of agreement" and "disagreement"? It is of interest that Kendall also defined a *coefficient of concordance*, $W$, and gave it the range, $0 \leq W \leq 0$, since disagreement or contradiction, is illogical when more than two individuals are involved. But $u$ also gives negatives values, and yet, the basic rationale for $W$ and $u$ is essentially the same. It is true that as $n$ approaches infinity the negative values of $u$ shrink towards zero, but this seems to be a failure of the model to account for empirical data, rather than a desirable logical attribute.

2. $a_{ij}$ and $a_{ji}$ are complementary quantities, and are not expected to vary randomly with respect to each other, but formula (1) treats them as though they were entirely unrelated; finally,

3. Formula (1) implies that the effects of $a_{ij}$ and $a_{ji}$ combine additively to contribute to the observed degree of agreement; however, $a_{ij}$ indicates the number of individuals disagreeing — or not agreeing — with $a_{ji}$ other individuals, since the former prefer $s_i$ over $s_j$, while the latter cast their vote to

[1] This communication represents Report No. 5 of the Laboratory of Experimental Psychology of the Ateneo de Manila University.

[2] For details about the technique of pair-comparisons and about Kendall's $u$, the reader is referred to David (1963), Edwards (1957), Kendall (1962), or Moroney (1963).

favor $s_j$ over $s_i$. It is not apparent how these two quantities can give additive effects.

All these are serious short-comings. They suggest the possibility of a logical miscalculation in the formulation of (1).

There is an alternate way for arriving at a measure of consensus as obtained in pair-comparison studies. Consider the fact that for every comparison $(s_i, s_j)$, some individuals choose $s_i > s_j$. There are $a_{ij}$ of them; also some individuals choose $s_j > s_i$. Their number is $a_{ji}$. Clearly these two groups of observers are at odds in this respect. By the multiplication principle, the number of ways in which they can fail to agree with one another is, $(a_{ij})\ (a_{ji})$, the product of their numbers. If $n$ is even, no decisions are possible if $a_{ij} = n/2 = a_{ji}$. The maximal number of ways in which the two groups can disagree in this case turns out to be, $n^2/4$. Conversely, where there is complete unanimity, $a_{ij} = 0$, or $a_{ji} = 0$. In any case, $(a_{ij})\ (a_{ji}) = 0$, also. Therefore, the ratio of partial disagreement to total disagreement — or "no consensus" — may be expressed as

$$d = \frac{4(a_{ij})\ (a_{ji})}{n^2} \qquad (2)$$

and may be regarded as an index of disagreement for any single comparison of $s_i$ and $s_j$, in which $n$ observers participated with one trial per individual. $d = 0.000$ if either $a_{ij}$ or $a_{ji}$ is zero, indicating no disagreement. $d = 1.000$ if $a_{ij} = n/2 = a_{ij}$, indicating a tie, or a general lack of consensus. In general, $0 \leq d \leq 1$.

Since $d$ is a measure of the extent to which lack of unanimity exists, the magnitude of consensus may be defined by

$$c = 1 - d \qquad (3)$$

for any comparison $(s_i, s_j)$. Now if we have made $m = k(k-1)/2$ pair-comparisons, the average $c$, $M(c)$ is

$$M(c) = \frac{\sum_{i,j} c}{m} \qquad (4)$$

Then from formula (3) we get

$$M(c) = \frac{\sum_{i,j} (1-d)}{m} \qquad (5)$$

which upon substitution and simplification finally gives

$$M(c) = 1 - \frac{4 \sum_{i,j} (a_{ij})\ (a_{ji})}{n^2\, k(k-1)/2}$$

Formula (5) represents a measure, or an *index of consensus*, averaged over all the comparisons of pairs, $(s_i, s_j)$. It applies when $n$ is an even number. If $n$ is an odd number, no ties are possible but minimal consensus obtains when $a_{ij} = (n+1)/2$, and $a_{ji} = (n-1)/2$, or vice versa, in which case, $(a_{ij})\ (a_{ji}) = (n^2-1)/4$. By reasoning similar to that which led to (5), we arrive at

$$M(c) = 1 - \frac{4 \sum_{i,j} (a_{ij})(a_{ji})}{(n^2-1)k(k-1)/2} \qquad (6)$$

to cover the case in which $n$ is an odd number. It can be easily verified that (5) and (6) always yield values of $M(c)$ such that, $0 \leq M(c) \leq 1$.

*An example of the computation of* $M(c)$. A set of 10 words (Agitolalia, Catelectrotonus, Decibel, Goniometer, Leptosome, Nares, Paralexia, Schizothymia, Sciosophy, and Tautophone) were combined according to the pair-comparison strategy. Twenty-eight sophomores were then asked to choose from each pair that word which was more familiar, or which "sounded" more familiar to them. The list of pairs given to the *S*s is shown in Table 1. The *S*s indicated their preference by putting an *X* on the blank space provided between the words, on the side of the word of their choice. For the purpose of getting $M(c)$, it is enough to define $a_{ij}$ as the number of choices assigned to the word on the left, regardless of the identity of the word. Such procedure is logically faulty, but is otherwise economical and arithmetically exact.

Defining $a_{ij}$ in the manner described above is economical in the sense that

TABLE 1

LIST OF WORDS AS PRESENTED TO S.

On the right of each pair is the corresponding $a_{ij}$, $a_{ji}$, and $(a_{ij})$ $(a_{ji})$. (See text for more complete description).

| | Pair | | $a_{ij}$ | $a_{ji}$ | $(a_{ij})$ $(a_{ji})$ |
|---|---|---|---|---|---|
| 1. | catelectrotonus | agitotalia | 12 | 16 | 192 |
| 2. | decibel | catelectrotonus | 28 | 0 | 0 |
| 3. | goniometer | decibel | 0 | 28 | 0 |
| 4. | leptosome | goniometer | 12 | 16 | 192 |
| 5. | nares | leptosome | 20 | 8 | 160 |
| 6. | schizothymia | nares | 18 | 10 | 180 |
| 7. | sciosophy | schizothymia | 10 | 18 | 180 |
| 3. | goniometer | sciosophy | 21 | 7 | 147 |
| 9. | paralexia | tautophone | 16 | 12 | 192 |
| 10. | agitolalia | decibel | 1 | 27 | 27 |
| 11. | catelectrotonus | goniometer | 2 | 26 | 52 |
| 12. | decibel | leptosome | 26 | 2 | 52 |
| 13. | goniometer | nares | 11 | 17 | 187 |
| 14. | leptosome | schizothymia | 11 | 17 | 187 |
| 15. | nares | sciosophy | 18 | 10 | 180 |
| 16. | schizothymia | tautophone | 16 | 12 | 192 |
| 17. | sciosophy | paralexia | 8 | 20 | 160 |
| 18. | goniometer | agitolalia | 18 | 10 | 180 |
| 19. | leptosome | schizothymia | 22 | 6 | 132 |
| 20. | nares | sciosophy | 2 | 26 | 52 |
| 21. | schizothymia | goniometer | 12 | 16 | 192 |
| 22. | sciosophy | leptosome | 8 | 20 | 160 |
| 23. | tautophone | nares | 12 | 16 | 192 |
| 24 | paralexia | schizothymia | 18 | 10 | 180 |
| 25. | agitolalia | leptosome | 9 | 19 | 171 |
| 26. | catelectrotonus | nares | 5 | 23 | 115 |
| 27. | decibel | schizothymia | 28 | 0 | 0 |
| 28. | goniometer | sciosophy | 22 | 6 | 132 |
| 29. | leptosome | tautophone | 12 | 16 | 192 |
| 30 | nares | paralexia | 18 | 10 | 180 |
| 31. | nares | agitolalia | 20 | 8 | 160 |
| 32. | schizothymia | catelectrotonus | 19 | 9 | 171 |
| 33. | sciosophy | decibel | 1 | 27 | 27 |
| 34. | tautophone | goniometer | 16 | 12 | 192 |
| 35. | paralexia | leptosome | 17 | 11 | 187 |
| 36. | agitolalia | schizothymia | 8 | 20 | 160 |
| 37 | catelectrotonus | sciosophy | 12 | 16 | 192 |
| 38. | decibel | tautophone | 28 | 0 | 0 |
| 39. | goniometer | paralexia | 13 | 15 | 195 |
| 40. | sciosophy | agitolalia | 16 | 12 | 192 |
| 41. | tautophone | catelectrotonus | 20 | 8 | 160 |
| 42. | paralexia | decibel | 1 | 27 | 27 |
| 43. | agitolalia | tautophone | 7 | 21 | 147 |
| 44. | catelectrotonus | paralexia | 2 | 26 | 52 |
| 45. | paralexia | agitolalia | 23 | 5 | 115 |
| | | Sums: | 619 | 541 | 6135 |

we have to count only the $X$'s on the left. $a_{ji}$ can then be found by subtracting $a_{ij}$ from $n$. That is, $a_{ji} = n - a_{ij}$. These savings in terms of labor are made possible by the lack of directionality of the product, $(a_{ji})(a_{ij})$. In Table 1 the $a_{ij}$'s and $a_{ji}$'s are lined up with the item to which they correspond. Similarly, the products, $(a_{ij})(a_{ji})$ are also given, and from the column of products we obtain, $\sum_{i,j} (a_{ij})(a_{ji}) = 6135$.

Since $k = 10$, $k(k-1)/2 = 45$, $n = 28$ an even number, we use formula (5):

$$M(c) = 1 - \frac{4(6135)}{(784)(45)} = .3045$$

In $M(c)$ we have avoided the pitfalls that plague $u$ (formula (1) ), as discussed above. In addition, defining M($c$) as an average presents the distinct advantage of allowing us to find a "partial" $M(c)$—call it $m(c)$—which may be estimated over any number $m$ of comparisons, not necessarily $k(k-1)/2$. Suppose, for instance, that we are interested in the behavior of $s_1$ in reference to all other stimuli. Here, $m = (k-1)$, since $s_1$ enters only into $k-1$ comparisons. Accordingly, $M(c)$ becomes,

$$m(c) = 1 - \frac{4\sum_{1,j} (a_{1j})(a_{j1})}{n^2(k-1)} \tag{7}$$

if $n$ is even; if $n$ is odd, $n^2$ becomes $n^2 - 1$, and the denominator of the last term is $(n^2-1)(k-1)$. That is, formula (7) indicates, that we may select any comparisons we wish and average $c$ (formula (3) ), over them, to obtain what resembles a sort of "item analysis" for pair-comparison data.

*The significance of $M(c)$.* The next problem is that of evaluating the significance of $M(c)$. In reference to $u$, Kendall (1962) has shown that the quantity,

$$X^2 = \frac{4}{n-2}\left\{\left(\sum_{i,j}\left[\binom{a_{ji}}{2}+\binom{a_{ij}}{2}\right]\right. - \frac{1}{2}\binom{k}{2}\binom{n}{2}\frac{n-3}{n-2}\right\} \tag{8}$$

is distributed as $X^2$ with degrees of freedom

$$df = \frac{2\binom{k}{2}\binom{n}{2}}{(n-2)^2} \tag{9}$$

Since,

$$\sum_{i,j}\left[\binom{a_{ji}}{2}+\binom{a_{ij}}{2}\right] = \binom{k}{2}\binom{n}{2} - \sum_{i,j}(a_{ij})\quad(a_{ji}) \tag{10}$$

and since, when $n$ is even we can extract from formula (5),

$$\sum_{i,j}(a_{ij})\ (a_{ji}) = \frac{\left[1 - M(c)\right]\binom{k}{2}n^2}{4} \tag{11}$$

by appropriate substitutions involving formulas (11), (10), and (8)—working in a reverse order—we obtain, after much simplification (we omit the algebra),

$$X^2 = \frac{nk(k-1)}{2(n-2)^2}\left[M(c)\ (n^2-2)\ + 1\right] \tag{12}$$

if $n$ is an odd number, then, from (6) we get,

$$\sum_{i,j}(a_{ij})\ (a_{ji}) = \frac{\left[1 - M(c)\right](n^2-1)}{4}\binom{k}{2} \tag{13}$$

Again, suitable substitutions and algebraic calisthenics eventually yield—the intermediate steps are omitted,

$$X^2 = \frac{k(k-1)(n-1)}{2(n-2)^2}\left[M(c)\ (n+1)\ (n-2) + 2\right] \tag{14}$$

For both (13) and (14) the degrees of freedom remain as stated in formula (9).

It is of interest to notice that formulas (12) and (14) never give $X^2 = 0.000$, even if $M(c) = 0.000$. This is not a deficiency acquired in the process of translating (8) into (12) and (14). If, when $n$ is even, we let $a_{ij} = n/2 = a_{ji}$, holding them constant over all $k(k-1)/2$ comparisons, formula (8) becomes, after simplification,

$$X^2 = \frac{nk(k-1)}{2(n-2)^2} \qquad (15)$$

and if, when $n$ is odd we let, $a_{ij} = (n+1)/2$, and $a_{ji} = (n-1)/2$, holding constant over all $k(k-1)/2$ comparisons, formula (8) becomes, after simplification,

$$X^2 = \frac{k(k-1)(n-1)}{(n-2)^2} \qquad (16)$$

Both (15) and (16) are identical to the residuals of (12) and (14), respectively, when $M(c) = 0.000$. Now if $k$ remains small, while $n$ is increased, (15) and (16) approach zero as $n$ grows large. For this reason we suggest that the constants 1 and 2 be dropped from (15) and (14), respectively. This way, if $n$ is small, the investigator places himself on the conservative side in estimating probabilities related to $X^2$ while if $n$ is large, it will not make any difference anyway. Accordingly, formula (15) becomes,

$$X^2 = \frac{nk(k-1)(n^2-2)}{2(n-2)^2} M(c) \qquad (17)$$

for $M(c)$ computed with $n$, even. If $M(c)$ is computed with $n$, odd, formula (16) becomes,

$$X^2 = \frac{k(k-1)(n^2-1)}{2(n-2)} M(c) \qquad (18)$$

for a conservative estimate of the probability that $M(c)$ deviates seriously from a value of zero.

Finally, to complete the example started above, we find that given the values $n = 28$, $k = 10$, and $M(c) = 0.3045$, formula (9) gives

$$df = \frac{(2)(10)(9)(29)(28)}{(4)(676)} = 50.3254$$

that is, we have 50 degrees of freedom (rounding off), while formula (7) gives

$$X^2_{50} = \frac{(28(10)(9)(784-2)(0.3045)}{(2)(676)} = 443.8312$$

This value of $X^2$ with $df = 50$ has $p < .001$, and can be interpreted as suggesting that the magnitude of consensus obtaining for this group, with these particular set of stimuli, implies a definite trend towards agreement about the ordinal arrangement of the stimulus words in terms of their familiarity to the group.

## REFERENCES

DAVID, H. A. The method of paired comparisons. *Griffin's Statistical Monograph & Courses.* London: Griffin, 1963.

EDWARDS, A. L. *Techniques of attitude scale construction.* New York: Appleton-Century-Crofts, 1957.

KENDALL, M. G. *Rank correlation methods.* (2nd ed.) London: Griffin, 1962.

MORONEY, M. J. *Facts from figures.* (3rd ed.) Middlesex: Penguin Books, 1956.