PERCENTAGE AGREEMENT AND PHI: A CONVERSION TABLE

LEWIS M. LEWIN AND JAMES A. WAKEFIELD, JR.

CALIFORNIA STATE COLLEGE, STANISLAUS

Studies in applied behavior analysis have used two expressions of reliability for human observations: percentage agreement (including percentage occurrence and percentage nonoccurrence agreement) and correlational techniques (including the phi coefficient). The formal relationship between these two expressions is demonstrated, and a table for converting percentage agreement to phi, or vice-versa, is presented. It is suggested that both expressions be reported in order to communicate reliability unambiguously and to facilitate comparison of the reliabilities from different studies.

DESCRIPTORS: reliability, observer agreement, phi, percentage agreement, methodology

Kelly (1977) reported that, although most studies in *JABA* use percentage agreement to assess observer reliability, there is some use of correlational techniques. In addition, several authors (Hartmann, 1977; Kratochwill & Wetzel, 1977) have offered advantages, with few disadvantages, for the use of correlational techniques. Yet Hartmann demonstrated that percentage agreement and correlations often yield markedly different values when applied to the same data, with percentage agreement seeming to overstate reliability compared to correlations.

The reason for the discrepancy between correlational indices of reliability and percentage agreement is that a correlation of .00 indicates randomly occurring agreement while a percentage agreement of 00 indicates absolutely no agreement. In fact, when a behavior is observed either very frequently or very infrequently by two observers, low percentage agreements are impossible, but low correlations are possible (and highly probable with inaccurate observations).

Thus, applied behavior analysts are likely to be faced with the problem of comparing the results of two studies that use different statistics to describe similar or identical situations. In

order to facilitate reporting and comparing of both correlational and percentage agreement statistics, a table is presented for converting percentage agreement scores to phi coefficients and vice versa. The table is based on a formula¹ presented by Wakefield (in press).

For economy, not all combinations of the rates of observations of the two observers are included. We have assumed one observer will be within 20 percentage points of the second observer with respect to the percentage of the total intervals in which behavior is judged to occur. This seems to be a reasonable practical assumption because high percentage agreements or phi's are not possible when the observation rates are greatly disparate. Wakefield's formulas, however, will convert phi to percentage agreement, or vice versa, with any degree of disparity between the rates of observation.

The user of the table should find the proportions of observation of two observers (P_1 and P_2) in the top rows of the table. The reported per-

$${}^1\phi = \sqrt{\frac{p_1q_1}{q_1p_1}} + \left(\frac{p_1q_1 + p_1q_2 + q_1q_2 + p_1p_2}{2\sqrt{p_1p_1q_1q_1}} \right) (\% + p_1 - p_2 - 1)$$

Where p_1 = proportion of occurrences for observer i , p_j = proportion of occurrences for observer j , $q_1 = 1 - p_1$, $q_j = 1 - p_j$, % = percentage agreement, and $p_i \neq p_j$.

Reprints may be obtained from Lewis M. Lewin or James A. Wakefield, Jr., Psychology Department, California State College, Stanislaus, Turlock California 95380.

Table 1
Phi Coefficients as a Function of Observer Rates ($P_1 + P_2$) and Percentage Agreement

$P_1(\text{obs. } 1)$	.9(.1)*			.8(.2)			.7(.3)			.6(.4)			.5
$P_2(\text{obs. } 2)$	.9(.1)	.8(.2)	.7(.3)	.8(.2)	.7(.3)	.6(.4)	.7(.3)	.6(.4)	.5	.6(.4)	.5	.4(.6)	.5
Percentage Agreement													
100	1.00	**	**	1.00	**	**	1.00	**	**	1.00	**	**	1.00
98	.89	**	**	.94	**	**	.95	**	**	.96	**	**	.96
96	.78	**	**	.87	**	**	.90	**	**	.92	**	**	.92
94	.67	**	**	.81	**	**	.86	**	**	.87	**	**	.88
92	.56	**	**	.75	**	**	.81	**	**	.83	**	**	.84
90	.44	.67	**	.69	.76	**	.76	.80	**	.79	.82	**	.80
88	.33	.58	**	.62	.71	**	.71	.76	**	.75	.78	**	.76
86	.22	.50	**	.56	.65	**	.67	.71	**	.71	.73	**	.72
84	.11	.42	**	.50	.60	**	.62	.67	**	.67	.69	**	.68
82	.00	.33	**	.44	.55	**	.57	.62	**	.62	.65	**	.64
80	-.11	.25	.51	.37	.49	.61	.52	.58	.65	.58	.61	.67	.60
78	**	.17	.44	.31	.44	.56	.48	.53	.61	.54	.57	.62	.56
76	**	.08	.36	.25	.38	.51	.43	.49	.57	.50	.53	.58	.52
74	**	-.00	.29	.19	.33	.46	.38	.45	.52	.46	.49	.54	.48
72	**	-.08	.22	.12	.27	.41	.33	.40	.48	.42	.45	.50	.44
70	**	-.17	.15	.06	.22	.36	.29	.36	.44	.37	.41	.46	.40

* $P_1 = .9$ or $.1$.

If $P_1 = .9$, $P_2 = \{.9, .8, .7\}$.

If $P_2 = .1$, $P_2 = \{.1, .2, .3\}$, etc.

**Impossible occurrence.

centage agreement may be found in the left-hand column. The tabled number where the row and column intersect is the corresponding phi coefficient. Stars in the table indicate impossible occurrences.

Redundancy in the table has been eliminated by placing in parentheses observer rates which result in the same phi coefficient as the rates listed next to the parenthesized numbers. For example, if two observers observed a behavior 90% of the time ($P_1 = P_2 = .9$) and they had 94% agreement, the phi coefficient would be .67. The same phi (.67) would be observed if $P_1 = P_2 = .1$ and percentage agreement were 94%. This value appears in the first column, which is given two different sets of labels (i.e., $P_1 = .9$ and $P_2 = .9$ or $P_1 = .1$ and $P_2 = .1$).

As a practical example of the use of the table, consider an observer recording disruptive behavior in 80% (P_1) of observed intervals and another observer recording disruptive behavior in 70% (P_2) of the intervals. If they had a percentage agreement of 90, the corresponding phi

would be .76. If they had a percentage agreement of 80, the phi would be .49.

As illustrated in Table 1, it is possible to have high percentage agreement and low phi. When this occurs, phi is the appropriate index of reliability, and percentage agreement is merely a spuriously high number that includes forced agreements (forced by the observer rates) and chance agreements (from inattentive, uncertain, or uncooperative raters) along with actual agreements. For example, if two observers observe appropriate eating behavior in 90% of their observations of a retarded subject ($P_1 = P_2 = .90$), their percentage agreement *cannot* be lower than 80%. If both randomly made up the data (independently) without actually observing, they would most likely have a percentage agreement of 82%. Obviously the phi of .0 in the table is a better expression of the reliability of these observations than is a number that adds 80% forced agreements to 2% chance agreements.

In practical situations, the exact values of the observer rates (P_1 and P_2) and percentage agree-

ment may not appear in the table. When percentage agreement or one of the observer rates (P_1 or P_2) is between two tabled values, linear interpolation can be satisfactorily used to compute phi. However, when any two of the three values of P_1 , P_2 , or percentage agreement require interpolation, the formula in footnote 1 should be used instead.

DISCUSSION

There are advantages to reporting phi in addition to percentage agreement. The correlational indices of reliability are less spuriously inflated than percentage agreement by changes in the frequency of the target behavior that in turn change the frequency of forced agreements and are thus more accurate across a wide range of frequencies. This characteristic has the advantage of allowing comparison of reliability indices both between and within studies even when the frequency of behavior varies—a situation commonly encountered by applied behavior analysts.

With JABA's traditional indices of interobserver reliability (percentage agreement and effective percentage agreement), two different indices may be required depending on the percentage of intervals in which the behavior occurs (Hartmann, 1977). When the percentage of intervals in which the behavior occurs is greater than 80% or less than 20% (both are arbitrary cut-off points), the percentage of nonoccurrence or percentage of occurrence, respectively, is considered more appropriate than percentage agreement. Thus, indices of reliability from two studies are not comparable when the percentage of intervals in which the target behavior occurs differs between studies. In contrast, correlational indices of reliability can be legitimately com-

pared across studies even when the percentage of intervals in which the target behavior occurs differs *between* studies.

The same problem of noncomparability for between-study comparisons occurs for within-study comparisons. The requirement of different reliability indexes (percentage agreement or effective percentage agreement) depending on the rate of behavior has produced, as Hartmann (1977) states, "ambiguity of usage" (p. 113). As several authors (Hartmann; Kratochwill & Wetzel, 1977) have indicated, it is often difficult to decide which is the most appropriate index of reliability to use when different rates of behavior occur during an experiment. Hartmann has recommended that the use of effective percentage statistics be restricted to situations in which extreme rate behaviors do not undergo substantial changes in rate. In contrast, the phi coefficient represents a far less drastic distortion of obtained interobserver agreement as the frequency of behavior changes *within* a study.

REFERENCES

- Hartmann, D. P. Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis*, 1977, **10**, 103-116.
- Kelly, M. B. A review of the observational data-collection and reliability procedures reported in the Journal of Applied Behavior Analysis. *Journal of Applied Behavior Analysis*, 1977, **10**, 97-101.
- Kratochwill, T. R., and Wetzel, R. J. Observer agreement, credibility, and judgment: Same considerations in presenting observer agreement data. *Journal of Applied Behavior Analysis*, 1977, **10**, 133-139.
- Wakefield, J. A., Jr. Relationship between two expressions of reliability: Percentage agreement and phi. *Educational and Psychological Measurement*, in press.

Received November 4, 1977

Final acceptance November 15, 1978