EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1985, 45

THREE COEFFICIENTS FOR ANALYZING THE RELIABILITY AND VALIDITY OF RATINGS

LEWIS R. AIKEN
Pepperdine University, Malibu

Three numerical coefficients (V, R, and H) for analyzing the validity and reliability of ratings are described. Each coefficient, which ranges in value from 0 to 1, is computed as the ratio of an obtained to a maximum sum of differences in ratings, or as 1 minus that ratio. Computer programs for calculating the coefficients, their associated individual and cumulative right-tail probabilities, and the population mean and standard deviation of each coefficient are available. Tables of the individual and right-tail probabilities for selected values of the three coefficients are included for 2 to 7 rating categories and 2 to 25 items or raters. More complete probability tables can be generated for any value of c, m, or n with one of the computer programs provided. When the number of items or raters is large (>25), the right-tail probability associated with any value of a coefficient may be estimated by a z-score procedure. The three coefficients are applicable not only to validity and reliability (test-retest and internal consistency) determinations but also to the item analysis, agreement analysis, and cluster or factor analysis of rating-scale data.

BOTH unpublished and commercially-available scales and other psychometric devices yielding ordinal-level measurements appear to have increased in frequency of usage during recent years. Unfortunately, information concerning the reliability and validity of many of these instruments is either unavailable, insufficient, or inappropriate. The responsibility for making such information available and sufficient lies directly with the researchers and their mentors. On the other hand, these individuals are not necessarily responsible for the inappropriateness or inadequacy of statistical procedures used to assess and express the reliability and validity of their instruments.

For example, two of the most widely recommended methods for calculating reliability coefficients of rating scales—the Kuder-Richardson and Cronbach alpha formulas—assume measurement at an interval level in fairly large samples. And with regard to validity, fledgling researchers of educational and other psychosocial events are frequently at a loss to demonstrate the criterion-related or construct validity of their questionnaires, rating scales, and checklists. The "validity problem" in such cases is often hurriedly "solved" by having a small sample of "experts" attest to the fact that an instrument is a valid measure of whatever it is supposed to measure.

There is, of course, nothing wrong with using expert judgments if the judgments are made carefully and independently. In any event, such judgments are often the only kind of validity evidence obtainable by student-researchers. It would, therefore, seem useful to provide more appropriate statistical methods for analyzing data from validity judgments or ratings. With this in mind, over the past few years the writer has devised a set of procedures for computing and determining the statistical significance of a content validity coefficient (V) and two types of reliability coefficients—a repeatability coefficient (R) and a homogeneity coefficient (H)—for use with rating scales and other ordinally-scaled psychometric instruments. These are basically small-sample techniques for which the exact probability distributions have been determined, but formulas for computing large-sample approximations to the exact probabilities have also been worked out.

Procedures for computing and testing the significance of the V and R coefficients in one situation, that of assessing the content validity and reliability of judgments of single rating-scale items, were outlined in a previous paper (Aiken, 1980). Applications of V and R have subsequently been extended to a number of other situations, and an H coefficient has been devised as an alternative to traditional internal-consistency reliability coefficients. Several computer programs have also been prepared to compute the coefficients, generate the discrete probability distributions associated with them, and, in the case of large samples of items or raters, approximate the exact probabilities.

Validity Index

The procedure for determining a V coefficient begins with the ratings (judgments) of a single item by n raters (judges) or the ratings of m items by a single rater. Validity ratings can be made on any

convenient scale of c successive integers (e.g., 1, 2, 3, 4, 5 or 0, 1, 2, 3 or −3, −2, −1, 0, 1, 2, 3). Designating the integer assigned to the highest validity category as hi, the integer assigned to the lowest validity category as lo, and the rater's validity rating of the item as r, the r values are transformed to $s = r - \text{lo}$ if lo < hi and to $s = \text{lo} - r$ if hi < lo. The values of s for all raters (or items) are then added across the n raters or m items to yield S. When ratings of one item are made by n raters, the V coefficient for that item is computed as $V = S/[n(c - 1)]$. When ratings of m items are made by one rater, the V coefficient for that rater is computed as $V = S/[m(c - 1)]$. The range of both V coefficients is 0 to 1, a high value indicating that an item has high content validity (when ratings of a single item are made by n raters) or that a set of items has high content validity in the judgment of a single rater (when ratings of n items are made by one rater).

Determining the Statistical Significance of V

The discrete, right-tail probabilities associated with selected values of V for 2 to 7 rating categories (c) and 2 to 25 items (m) or raters (n) are given in Table 1. The two V values (for each c and m or n) selected for inclusion in the table are those having right-tail probabilities close to but not greater than the .01 and .05 levels, respectively.

By generating the complete table of probabilities associated with all possible V values for specified values of c and m or n, the ratings of m items by one rater or one item by n raters can be generalized to the ratings of m items by n raters. Assume, for example, that n values of V, each based on the four-category ratings assigned by one rater to five items, have been calculated. Then, using computer program "Proco"[1] with $c = 4$, $m = 5$, and right-tail probability limits of 0 and 1, a table of discrete and cumulative right-tail probabilities associated with every possible value of V is computed. Selecting a value of V toward the middle of this cumulative probability table, say $V_c = .53$, the corresponding right-tail probability—in this case $p_c = .50$—is used as the appropriate division point for the upper and lower groups. Then the actual number (k) of V values in the obtained

[1] Three computer programs, written in both MS-BASIC and FORTRAN-77, for computing and using the V, R, and H indexes are available from the writer on request. Write to: Lewis R. Aiken, Ph.D.; Social Science Division; Pepperdine University, Malibu; Malibu, CA 90265. Program 1 ("Comco") computes the V, R, and H coefficients. Program 2 ("Proco") computes the individual and right-tail probabilities for all possible values of V, R, and H, or a specified subset of the probabilities, for c = 2 to 7 and any value of m or n. Program 3 ("Popco") computes the population mean and standard deviation of V, R, and H.

TABLE 1
Right-Tail Probabilities (p) for Selected Values of the Validity Coefficient (V)

No. of Items (m) or Raters (n)	Number of Rating Categories (c)											
	2		3		4		5		6		7	
	V	p	V	p	V	p	V	p	V	p	V	p
2							1.00	.040	1.00	.028	1.00	.020
3							1.00	.008	1.00	.005	1.00	.003
3			1.00	.037	1.00	.016	.92	.032	.87	.046	.89	.029
4					1.00	.004	.94	.008	.95	.004	.92	.006
4			1.00	.012	.92	.020	.88	.024	.85	.027	.83	.029
5			1.00	.004	.93	.006	.90	.007	.88	.007	.87	.007
5	1.00	.031	.90	.025	.87	.021	.80	.040	.80	.032	.77	.047
6			.92	.010	.89	.007	.88	.005	.83	.010	.83	.008
6	1.00	.016	.83	.038	.78	.050	.79	.029	.77	.036	.75	.041
7			.93	.004	.86	.007	.82	.010	.83	.006	.81	.008
7	1.00	.008	.86	.016	.76	.045	.75	.041	.74	.038	.74	.036
8	1.00	.004	.88	.007	.83	.007	.81	.008	.80	.007	.79	.007
8	.88	.035	.81	.024	.75	.040	.75	.030	.72	.039	.71	.047
9	1.00	.002	.89	.003	.81	.007	.81	.006	.78	.009	.78	.007
9	.89	.020	.78	.032	.74	.036	.72	.038	.71	.039	.70	.040
10	1.00	.001	.85	.005	.80	.007	.78	.008	.76	.009	.75	.010
10	.90	.001	.75	.040	.73	.032	.70	.047	.70	.039	.68	.048
11	.91	.006	.82	.007	.79	.007	.77	.006	.75	.010	.74	.009
11	.82	.033	.73	.048	.73	.029	.70	.035	.69	.038	.68	.041
12	.92	.003	.79	.010	.78	.006	.75	.009	.73	.010	.74	.008
12	.83	.019	.75	.025	.69	.046	.69	.041	.68	.038	.67	.049
13	.92	.002	.81	.005	.77	.006	.75	.006	.74	.007	.72	.010
13	.77	.046	.73	.030	.69	.041	.67	.048	.68	.037	.67	.041
14	.86	.006	.79	.006	.76	.005	.73	.008	.73	.007	.71	.009
14	.79	.029	.71	.035	.69	.036	.68	.036	.66	.050	.66	.047
15	.87	.004	.77	.008	.73	.010	.73	.006	.72	.007	.71	.008
15	.80	.018	.70	.040	.69	.032	.67	.041	.65	.048	.66	.041
16	.88	.002	.75	.010	.73	.009	.72	.008	.71	.007	.70	.010
16	.75	.038	.69	.046	.67	.047	.66	.046	.65	.046	.65	.046
17	.82	.006	.76	.005	.73	.008	.71	.010	.71	.007	.70	.009
17	.76	.025	.71	.026	.67	.041	.66	.036	.65	.044	.65	.039
18	.83	.004	.75	.006	.72	.007	.71	.007	.70	.007	.69	.010
18	.72	.048	.69	.030	.67	.036	.65	.040	.64	.042	.64	.044
19	.79	.010	.74	.008	.72	.006	.70	.009	.70	.007	.68	.009
19	.74	.032	.68	.033	.65	.050	.64	.044	.64	.040	.63	.048
20	.80	.006	.72	.009	.70	.010	.69	.010	.68	.010	.68	.008
20	.75	.021	.68	.037	.65	.044	.64	.048	.64	.038	.63	.041
21	.81	.004	.74	.005	.70	.010	.69	.008	.68	.010	.68	.009
21	.71	.039	.67	.041	.65	.039	.64	.038	.63	.048	.63	.045
22	.77	.008	.73	.006	.70	.008	.68	.009	.67	.010	.67	.008
22	.73	.026	.66	.044	.65	.035	.64	.041	.63	.046	.62	.049
23	.78	.005	.72	.007	.70	.007	.68	.007	.67	.010	.67	.009
23	.70	.047	.65	.048	.64	.046	.63	.045	.63	.044	.62	.043
24	.79	.003	.71	.008	.69	.006	.68	.008	.67	.010	.66	.010
24	.71	.032	.67	.030	.64	.041	.64	.035	.62	.041	.62	.046
25	.76	.007	.70	.009	.68	.010	.67	.009	.66	.009	.66	.009
25	.72	.022	.66	.033	.64	.037	.63	.038	.62	.039	.61	.049

sample of five V's falling at or above $V_c = .53$ is counted, and a binomial test with $n = 5$ and $p_c = .50$ is conducted to determine the probability that, by chance, k out of n observations will fall in the p_c group and $n - k$ observations in the $1 - p_c$ group. If the value of np_c is 5 or greater, the chi square distribution rather than the binomial may be used for this significance test.

The same procedure as that outlined above can be used for any value of c and m when the V's are computed across items, or for c and n when the V's are computed across raters. In the latter case, the complete probability distribution for specified values of c and n is generated, a convenient value of V (V_c) toward the center of the distribution is selected, and the right-tail probability (p_c) for V_c becomes the theoretical ("expected") probability for the upper group and $1 - p_c$ the theoretical probability for the lower group. Then the number (k) of V's in the sample falling in the upper group and the number ($m - k$) falling in the lower group are counted. Finally, a binomial or chi square test is conducted to determine the probability that, by chance, k out of m observations will fall at or above V_c and $m - k$ observations will fall below V_c.[2]

Large-Sample Test for the Mean of V

When the sample of items or raters is large (m or $n > 25$), the central limit theorem can be applied to determine the statistical significance of the mean value of V. For example, assume that n raters rate m items on c rating categories. Compute n values of V, each based on the ratings of the m items provided by a single rater, and calculate the mean ($\bar{V}$) of those n values. Then, assuming independent ratings across raters, it can be shown that the expected mean (μ) and standard deviation (σ) of the V index are $\mu = .5$ and $\sigma = .5\sqrt{(c + 1)/[3m(c - 1)]}$, respectively. Applying the central limit theorem, $z = .2(\bar{V} - .5)\sqrt{3mn(c - 1)/(c + 1)}$. If z is greater than 1.645 (.05 level) or 2.33 (.01) level, it is concluded that the set of items, and hence the entire scale or questionnaire, has significant content validity. By substituting n for m in the formula for the

[2] The following procedure (after Jones and Fiske, 1963) is also appropriate for testing the significance of several values of V in combination:

a. Use a complete table of right-tail probabilities to determine the probability (p_i) associated with each of the n values of V (V_i).
b. Convert each p_i to a common logarithm (l_i).
c. Add the n (or m) values of l_i and multiply the sum by -4.605 to yield $\chi^2 = -4.605 \Sigma \log p_i$.
d. Test the significance of this chi square with $2n$ degrees of freedom.

population standard deviation and m for n in the formula for z, the significance of the mean of m values of V, each based on n items, can also be tested. Note that the formula for z turns out to be the same in both cases.

Repeatability Coefficient

In a previous paper (Aiken, 1980), the writer described a procedure and a coefficient (R) for determining the test-retest reliability, or "repeatability," of content validity ratings. Using this procedure, n raters rate an item or instrument twice (with a suitable interval in between) on a c-integer scale.[3] The sum (S) of the absolute values of the differences between the ratings of each rater on the two occasions is then computed. Next the R coefficient is calculated as $R = 1 - S/[n(c - 1)]$.

Determining the Statistical Significance of R

The right-tail probabilities associated with selected values of R for $c = 2$ to 7 and m or $n = 2$ to 25 are given in Table 2. As in the case of Table 1 for the V coefficient, the R values in Table 2 are those having right-tail probabilities close to but not exceeding .05 or .01. A complete table of V and R values and their associated discrete and right-tail probabilities can be generated with an MS-BASIC or FORTRAN-77 computer program available from the writer (see footnote 1).

Like the V coefficient, R may also be computed from the ratings of m items by one rater on two occasions: $R = 1 - S/[m(c - 1)]$. The statistical significance of this R can be determined similarly by referring to Table 2, although this time the table is entered with m, the number of items, rather than n, the number of raters. Furthermore, whether R is based on ratings assigned by a single rater to m items or the ratings assigned by n raters to a single item, the statistical significance of a sample of R values can be obtained by the same procedure as that described above for testing the significance of a sample of V's. A value of $R(R_c)$ toward the center of the theoretical probability distribution of R for specified values of c and m or c and n is selected, and its corresponding right-tail probability

[3] It is not necessary for the ratings on the two occasions to involve the same number of rating categories or even the same items, although they typically do. If the items rated on the second occasion are different from but parallel to those rated on the first occasion, R is a parallel-forms reliability coefficient.

TABLE 2
Right-Tail Probabilities (p) for Selected Values of the Repeatability Coefficient (R)

No. of Items (m) or Raters (n)	Number of Rating Categories (c) 2		3		4		5		6		7	
	R	p	R	p	R	p	R	p	R	p	R	p
2							1.00	.040	1.00	.028	1.00	.020
3							1.00	.008	1.00	.005	1.00	.003
3			1.00	.037	1.00	.016	.92	.046	.93	.028	.94	.018
4					1.00	.004	1.00	.002	.95	.006	.96	.003
4			1.00	.012	.92	.027	.88	.044	.90	.023	.88	.036
5			1.00	.004	.93	.008	.95	.003	.92	.006	.90	.009
5	1.00	.031	.90	.032	.87	.035	.85	.040	.84	.044	.83	.048
6			1.00	.001	.94	.002	.92	.004	.90	.005	.89	.006
6	1.00	.016	.92	.012	.83	.040	.83	.035	.83	.032	.83	.031
7			.93	.005	.90	.004	.89	.004	.89	.004	.86	.010
7	1.00	.008	.86	.024	.81	.044	.82	.030	.80	.048	.81	.038
8	1.00	.004	.88	.010	.88	.006	.88	.004	.85	.008	.85	.006
8	.88	.035	.81	.038	.79	.046	.81	.026	.80	.035	.79	.044
9	1.00	.002	.89	.004	.85	.007	.83	.010	.84	.006	.83	.009
9	.89	.020	.83	.018	.78	.047	.78	.046	.78	.047	.78	.048
10	1.00	.001	.85	.008	.83	.008	.82	.009	.82	.010	.83	.006
10	.90	.011	.80	.027	.77	.048	.78	.039	.78	.035	.78	.033
11	.91	.006	.86	.003	.82	.009	.82	.008	.82	.007	.82	.007
11	.82	.033	.77	.037	.76	.048	.77	.033	.76	.044	.77	.036
12	.92	.003	.83	.006	.81	.010	.81	.007	.82	.006	.81	.009
12	.83	.019	.75	.049	.75	.047	.75	.050	.77	.033	.76	.038
13	.92	.002	.81	.009	.79	.010	.81	.006	.80	.008	.80	.010
13	.77	.046	.77	.026	.74	.047	.75	.042	.75	.040	.76	.040
14	.86	.006	.82	.004	.81	.005	.80	.005	.80	.006	.80	.007
14	.79	.029	.75	.034	.74	.046	.75	.036	.74	.048	.75	.042
15	.86	.004	.80	.007	.80	.005	.78	.009	.79	.008	.79	.008
15	.80	.018	.73	.043	.73	.045	.73	.050	.75	.037	.74	.043
16	.81	.011	.78	.010	.79	.005	.78	.008	.79	.006	.78	.009
16	.75	.038	.75	.024	.73	.044	.73	.042	.74	.043	.74	.044
17	.82	.006	.79	.005	.78	.006	.78	.007	.78	.008	.78	.010
17	.76	.025	.74	.030	.73	.043	.74	.036	.73	.050	.74	.045
18	.83	.004	.78	.007	.78	.006	.78	.006	.77	.010	.77	.010
18	.72	.048	.72	.037	.72	.042	.72	.048	.73	.039	.73	.046
19	.79	.010	.76	.009	.77	.006	.76	.009	.77	.008	.77	.007
19	.74	.032	.71	.045	.72	.040	.72	.041	.73	.044	.73	.047
20	.80	.006	.78	.005	.77	.006	.76	.008	.76	.010	.77	.008
20	.75	.021	.72	.026	.72	.039	.72	.035	.72	.049	.72	.047
21	.81	.004	.76	.007	.76	.006	.76	.007	.76	.008	.76	.008
21	.71	.039	.71	.032	.71	.038	.71	.046	.72	.039	.72	.048
22	.77	.008	.75	.009	.76	.006	.75	.010	.75	.009	.76	.009
22	.73	.026	.70	.038	.71	.037	.72	.040	.72	.044	.72	.048
23	.78	.005	.76	.005	.75	.006	.75	.008	.76	.007	.75	.009
23	.70	.047	.70	.045	.71	.035	.71	.050	.71	.048	.72	.048
24	.79	.003	.75	.006	.75	.006	.75	.007	.75	.008	.75	.010
24	.71	.032	.71	.028	.71	.034	.71	.043	.72	.039	.72	.048
25	.76	.007	.74	.008	.75	.006	.74	.010	.74	.010	.75	.010
25	.72	.022	.70	.032	.71	.033	.71	.037	.71	.043	.71	.048

(p_c) is noted. Then a binomial or chi square test is conducted to determine the probability that k out of n (or m) obtained values of R will, by chance, fall at or above R_c (in the p_c group) and $n - k$ will fall below R_c (in the $1 - p_c$ group).

Large-Sample Test for the Mean of R

If the sample of raters (when R is based on the ratings of m items by a single rater) or items (when R is based on the ratings of a single item by n raters) is large, the central limit theorem may be applied to determine the right-tail probability associated with the mean R value. For example, assume that each value of R is based on the ratings of m items by a single rater. Compute the mean, $\bar{R}$, of the n obtained values of R. Now it can be shown that the expected (population) mean and standard deviation of R are $\mu = (2c - 1)/(3c)$ and $\sigma = \sqrt{(c + 1))(c^2 + 2)/[2m(c - 1)]/(3c)}$. If the computed value of $z = \sqrt{n}\ (\bar{R} - \mu)/\sigma$ is equal to or greater than a specified critical, right-tail value of z, it is concluded that the ratings of the m items are reliable. By substituting n for m in the above formulas for the standard deviation and m for n in the formula for z, it can be determined whether the mean of m values of R, each computed across n raters, is statistically significant. This is a similar, but not identical, way of demonstrating the test-retest reliability, or "repeatability" of the instrument.[4]

Homogeneity Coefficient

Obtaining successive ratings on two occasions is not always the simplest nor even the most accurate way of determining the reliability of a procedure or instrument. Two rater x situation occasions are never identical, and hence, as with the test-retest approach to determining the reliability of a test, the R coefficient may involve significant errors of measurement. Although internal consistency procedures for assessing reliability are also influenced by errors of measurement, they have the advantage of greater efficiency and possess some of the characteristics of the parallel forms approach. For these reasons, an internal-consistency method of determining the reliability of ratings would seem to have much to recommend it.

---

[4] Although the two approaches (across raters or across items) of computing V and R yield identical mean values of these statistics, the z values associated with the two values of $\bar{V}$ or $\bar{R}$ are equal only when $m = n$. Consequently, it is possible for the mean of the V coefficients computed across raters to be statistically significant when the mean of the V coefficients computed across items is not and vice versa.

TABLE 3
Right-Tail Probabilities (p) for Selected Values of the Homogeneity Coefficient (H)

No. of Items (m) or Raters (n)	Number of Rating Categories (c) 2		3		4		5		6		7	
	H	p	H	p	H	p	H	p	H	p	H	p
3							1.00	.040	1.00	.028	1.00	.020
4									1.00	.005	1.00	.003
4			1.00	.037	1.00	.016	1.00	.008	.85	.035	.83	.038
5					1.00	.004	1.00	.002	.87	.007	.89	.004
5			1.00	.012	.78	.033	.75	.040	.73	.035	.72	.050
6			1.00	.004	.81	.010	.86	.003	.78	.010	.81	.005
6	1.00	.031	.72	.037	.67	.046	.72	.024	.67	.041	.67	.046
7			1.00	.001	.72	.010	.75	.009	.73	.007	.72	.008
7	1.00	.016	.75	.014	.67	.030	.62	.041	.63	.036	.61	.050
8	1.00	.008	.78	.005	.71	.006	.67	.007	.67	.009	.69	.009
8			.56	.033	.56	.030	.56	.046	.57	.045	.57	.050
9	1.00	.004	.65	.009	.67	.007	.65	.009	.66	.007	.65	.009
9	.60	.039	.55	.031	.53	.047	.52	.048	.54	.047	.55	.046
10	1.00	.002	.64	.006	.60	.010	.60	.010	.62	.008	.62	.009
10	.64	.021	.52	.028	.48	.045	.51	.046	.51	.050	.53	.047
11			.60	.006	.58	.009	.58	.009	.59	.009	.59	.010
11	.67	.012	.50	.031	.49	.032	.48	.046	.51	.040	.50	.050
12	.69	.006	.56	.009	.55	.010	.56	.010	.57	.010	.57	.010
12	.44	.039	.44	.034	.44	.049	.47	.048	.48	.045	.49	.048
13	.71	.003	.52	.007	.52	.009	.55	.009	.55	.009	.56	.010
13	.48	.022	.43	.031	.43	.049	.45	.045	.47	.047	.48	.046
14	.73	.002	.51	.009	.51	.009	.52	.010	.53	.010	.54	.010
14	.51	.013	.41	.037	.42	.046	.44	.048	.45	.048	.46	.049
15	.54	.007	.48	.009	.50	.009	.51	.009	.52	.009	.52	.010
15	.36	.035	.39	.030	.40	.048	.43	.046	.44	.047	.45	.048
16	.56	.004	.48	.007	.48	.008	.49	.010	.51	.010	.51	.010
16	.39	.021	.38	.040	.39	.049	.41	.049	.43	.047	.44	.049
17	.58	.002	.47	.006	.46	.009	.48	.010	.49	.010	.50	.010
17	.28	.049	.35	.043	.39	.046	.41	.046	.42	.049	.44	.048
18	.44	.008	.42	.010	.45	.009	.47	.010	.48	.010	.49	.010
18	.31	.031	.33	.046	.37	.049	.40	.050	.41	.049	.43	.048
19	.47	.004	.41	.010	.44	.010	.46	.009	.48	.010	.49	.010
19	.33	.019	.33	.040	.37	.043	.39	.046	.41	.048	.42	.050
20	.49	.003	.40	.010	.43	.010	.45	.010	.47	.010	.48	.010
20	.25	.041	.32	.047	.36	.049	.38	.050	.40	.050	.41	.049
21	.38	.007	.39	.010	.42	.010	.45	.009	.46	.010	.47	.010
21	.27	.027	.33	.042	.35	.048	.38	.048	.40	.048	.41	.048
22	.40	.004	.37	.010	.41	.010	.44	.010	.45	.010	.46	.010
22	.30	.017	.30	.050	.35	.049	.37	.049	.39	.050	.40	.050
23	.42	.003	.37	.010	.41	.010	.43	.010	.45	.010	.46	.010
23	.23	.035	.30	.043	.34	.050	.37	.044	.38	.050	.40	.048
24	.34	.007	.36	.010	.40	.010	.42	.010	.44	.010	.45	.010
24	.25	.023	.30	.048	.34	.049	.36	.050	.38	.048	.39	.049
25	.36	.004	.36	.010	.39	.010	.42	.010	.43	.010	.44	.010
25	.19	.043	.29	.048	.33	.047	.36	.049	.38	.049	.39	.050

Computing the H Coefficient

The homogeneity coefficient (H), an internal consistency reliability coefficient for rating data, is, like the V and R coefficients, computed either across n raters or across m items. When computed across raters, H is a measure of agreement among the raters (or judges) as to how a specific item should be rated (or judged). When computed across items, H is a measure of the similarity of a single person's ratings of a set of m items. In the former case, the $n(n-1)/2$ absolute values of the differences among the ratings of the n raters are computed; in the latter case the $m(m-1)/2$ absolute values of the differences among the ratings assigned by one rater to m items are computed. Then the H coefficient, which ranges from 0 to 1, is computed as $H = 1 - 4S/[(c-1)(m^2 - j)]$, when the sum ($S$) of the absolute values of the differences is computed across items, or as $H = 1 - 4S/[(c-1)(n^2 - j)]$ when S is computed across raters. Note that $j = 0$ if m (or n) is even and $j = 1$ if m (or n) is odd.

Determining the Statistical Significance of H

The statistical significance of either H coefficient can be determined by referring to Table 3. As with Table 1 for the V coefficient and Table 2 for the R coefficient, only two values of H and the associated right-tail probabilities are given for each value of c and m (or n)—those closest to but not exceeding $p = .01$ and .05. A complete table of H values and their associated probabilities can be computed with an available computer program (see footnote 1).

Large-Sample Test for the Mean of H

Methods identical to those described previously for the V and R coefficients may be used to determine whether a small sample (m or $n < 25$) of H values is statistically significant. Assuming a large sample of H values, the central limit theorem can be applied to determine the significance of H. When the individual values of H are computed across m items, the population mean of H is $\mu = [2(c+1) + (m+j)(c-2)]/[3c(m+j)]$, and the population standard deviation of H is

$$\sigma = \frac{2}{(3c)} \sqrt{\frac{.2(c+1)(m+j-1)[c^2(m+3) - 2(2m-9)]}{[(c-1)(m+j)(m^2-j)]}}.$$

Then $z = \sqrt{n}\,(\bar{H} - \mu)/\sigma$, where $\bar{H}$ is the mean of n values of H. If the computed value of z is greater than or equal to the preset critical value of z, it is concluded that the items are significantly homogeneous.

When the individual H values are computed across raters rather than across items, m is changed to n in the above formulas for μ and σ and n is changed to m in the formula for z. A statistically significant z value in this second case reveals nothing about the internal consistency or homogeneity of the rated instrument itself. It indicates, rather, that the n raters were in accord in their ratings of the m items. In this case z is interpreted similarly to Kendall's (1970) coefficient of concordance.

Other Coefficients and Procedures

It should be noted that neither R nor H is limited in use to content validity ratings; the reliability of any procedure involving ratings may be assessed by these methods. Furthermore, V, R, H, and similar coefficients can be used for other purposes. In the computation of R, for example, if the items on the second occasion constitute a "validity criterion," then R becomes a criterion-related validity coefficient. And when R is computed across raters for all $m(m - 1)/2$ pairs of items, the resulting matrix of R values provides a starting point for an ordinal-level cluster or factor analysis of the ratings.

An application in the area of item analysis involves the computation of an I coefficient having features in common with both V and H. Determination of I for a specified item and rater begins with the computation of the sum (S) of the absolute values of the differences between the rating assigned to that item and the ratings assigned to all other $m - 1$ items. Then I is computed as $I = 1 - S/[(m - 1)(c - 1)]$. The statistical significance of I may be determined from Table 1 by using $m - 1$ rather than m as the number of items. The process of generalizing across raters is similar to that for V, with the condition that m becomes $m - 1$ in all formulas.

Computationally similar to I is an "agreement coefficient" (A), a measure of the extent to which a given rater agrees with the ratings assigned to an item by the other $n - 1$ raters. The computation of A for a specified item/rater combination begins by determining the sum of the absolute values of the $n - 1$ differences between the rating assigned by the designated rater to that item and the ratings assigned to the item by the other $n - 1$ raters. Then A is computed as $A = 1 - S/[(n - 1)(c - 1)]$ and tested for significance as a V index in which n is changed to $n - 1$.

Finally, the V and H coefficients can be used to detect the tendencies to be lenient and to succumb to a halo effect in making ratings. Uniformly high V indexes, across items or across raters, are suggestive of leniency errors. A high H coefficient, computed across

a group of items that ostensibly assess different characteristics, may point to the presence of a halo effect. This is especially true when the high H coefficient is accompanied by a high V computed on the same set of items.

REFERENCES

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 40, 955–959.

Jones, L. V. and Fiske, D. W. (1963). Models for testing the significance of combined results. *Psychological Bulletin*, 50, 375–382.

Kendall, M. G. (1970). *Rank correlation methods* (4th ed.) London: Charles Griffin.