Content Validity and Reliability of Single Items or Questionnaires

Lewis R. Aiken

Educational and Psychological Measurement 1980 40: 955
DOI: 10.1177/001316448004000419

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1980, 40

CONTENT VALIDITY AND RELIABILITY OF SINGLE ITEMS OR QUESTIONNAIRES

LEWIS R. AIKEN
Pepperdine University, Malibu

Procedures for computing content validity (V) and consistency reliability (R) coefficients and determining the statistical significance of these coefficients are described. These procedures, which employ the multinomial probability distribution for small samples and normal curve probability estimates for large samples, can be used in a variety of situations where judgments of the content validity of items or questionnaires are made on ordinal rating scales. Computing formulas for determining the statistical significance of V and R for large samples of raters and any number of rating categories are given. A computer program has been written to determine the right-tail probabilities associated with values of V and R obtained from ratings by N raters using c rating categories.

DECISIONS concerning the content validity of single items or questionnaires are typically expressed in terms of dichotomies (yes-no, agree-disagree, valid-invalid) or ranks (high validity-moderate validity-low validity). Social science researchers, however, often pay little attention to the statistical nature of these decisions, summarizing them with the statement that a majority or consensus of subject-matter experts agreed or failed to agree that the instrument in question is a valid measure of whatever it was intended to measure. Furthermore, the question of how reliable such decisions are may be entirely neglected.

The procedures described below for assessing the content validity and reliability of single items or the questionnaire as a whole are straightforward applications of multinomial and normal probability theory and are simple to use. The validity (V) and reliability (R) indexes described in this paper are certainly not the only statistics that can be employed for this purpose. Nevertheless, the V and R

coefficients do have the advantage of being applicable to both small and large samples of ratings made on any number of rating categories, and of ranging from 0 to 1.

Validity Index (V)

Assume that each of N raters inspects a single item or questionnaire and indicates, on a c-category ordinal rating scale (lowest validity category through highest validity category), his or her judgment of the content validity of the item or questionnaire. After all N raters have made their judgments, a scorer assigns a weight of 0 to each of the n_0 ratings falling in the lowest category, a weight of 1 to the n_1 ratings in the next higher category, and so on through a weight of $c - 1$ to each of the n_{c-1} ratings in the highest (cth) category. Then an index of content validity may be defined as

$$V = \sum_{i=1}^{c-1} \frac{in_i}{N(c-1)}. \tag{1}$$

Now if it is assumed that the probability of a single rating falling in the ith category is $1/c$, the multinomial probability formula

$$p(n_0, n_1, n_2, \cdots, n_{c-1}) = \frac{N!/c^N}{n_0!n_1!n_2! \cdots n_{c-1}!} \tag{2}$$

can be used to find the probability of a particular combination $(n_0, n_1, n_2, \cdots, n_{c-1})$ of the n_i's occurring at random. Combining the discrete probabilities associated with all possible values of V equal to or greater than the V computed from formula 1 yields the significance level of this V.

When there are only two rating categories, a given combination (n_0, n_1) of the n_i's yields a unique value of V. But when the number of rating categories is three or more, two or more different combinations of the n_i's can produce the same value of V. Therefore, the probabilities associated with the several n_i combinations yielding the same V value must first be summed to obtain the discrete probability associated with that value of V. Then, as in the two-category case, the discrete probabilities of all values of V equal to or greater than the value under consideration are combined to yield the significance level for V.[1]

[1] A FORTRAN program has been written to compute the right-tail probability associated with a specified value of V or R. Write to Lewis R. Aiken, PhD, Social Science Division, Pepperdine University, Malibu, California 90265 for a copy of the program and directions for its use.

When the number of raters is large, a normal approximation to the exact probability can be used. It can be shown that, for any number of categories (c), the probability under the standard normal curve to the right of

$$z = \frac{N(c-1)(2V-1)-1}{\sqrt{N(c-1)(c+1)/3}} \tag{3}$$

is a good approximation to the exact probability. The approximation improves as the number of raters and the number of rating categories increase.

Reliability Index (R)

Although information concerning the reliability of many psychometric instruments is obtained prior to the collection of validity data, in the present context the process is reversed. After it has been established that the index of content validity (V) is statistically significant, the investigator needs to determine whether the ratings are consistent (i.e., reliable) over time. To compute a consistency reliability coefficient, ratings of the content validity of the item or instrument on two separate occasions are required. Then, as indicated in the illustration of Table 1, these ratings are tabulated in the form of a $c \times c$ repeated ratings matrix.

Consider the numbers (n_{ij}) in the cells of the 3 × 3 matrix of Table 1. As shown in the top row of the matrix, six of the seven experts whose first rating was "high validity" continued on occasion 2 to rate the item as having high validity, and one expert

TABLE 1
Repeated Validity Ratings Matrix

		Rating on Occasion 2			
		Low Validity	Moderate Validity	High Validity	
Rating on Occasion 1	High Validity	0	1	6	7
	Moderate Validity	1	2	2	5
	Low Validity	2	1	0	3
		3	4	8	

changed his rating from "high validity" to "moderate validity." The numbers in the cells of the rows labeled "moderate validity" and "low validity" are interpreted similarly.

A consistency reliability coefficient for the case of three rating categories is defined as

$$R = 1 - \frac{\sum_{j=0}^{2} \sum_{i=0}^{2} n_{ij}|i - j|}{2N}. \tag{4}$$

Applying formula 4 to the ratings in Table 1 gives

$$R = 1 - \frac{2(0) + 1(1) + 0(2) + 1(1) + 2(0) + 2(1) + 0(2) + 1(1) + 6(0)}{2(15)} = .833.$$

Note that perfect consistency ($R = 1.00$) occurs only when all ratings fall in the cells of the minor (lower left to upper right) diagnonal of the matrix, and that R decreases as ratings fall in cells increasingly distant from the minor diagonal.

Observe that there are $c = 3$ cells on the minor diagonal of Table 1, a total of $2(c - 1) = 4$ cells in the first off-diagonal group, and $2(c - 2) = 2$ cells in the second off-diagonal group. Since there are $c^2 = 9$ cells in the entire matrix, the probability of a rating falling in a cell of the minor diagonal is 3/9; the probability of a rating falling in a cell of the first off-diagonal group is 4/9; and the probability of a rating falling in a cell of the second off-diagonal group is 2/9. Summing the n_{ij}'s in all cells of the minor diagonal, $d_2 = n_{00} + n_{11} + n_{22} = 10$. The n_{ij}'s in the cells of the first off-diagonal group sum to $d_1 = n_{01} + n_{10} + n_{12} + n_{21} = 5$, and the n_{ij}'s in the second off-diagonal group to $d_0 = n_{02} + n_{20} = 0$. The subscripts on the d_k's refer to scoring weights assigned to the n_{ij}'s comprising the d_k's in the computing formula for R. Then using the multinomial formula, the probability of the above combination of the d_k's occurring at random is computed as

$$p(0, 5, 10) = \frac{15!}{0!5!10!}(2/9)^0(4/9)^5(3/9)^{10} = .00088.$$

As in computing the validity index, however, different combinations of the summed ratings may yield the same value of R. Furthermore, the cumulative probability in the right tail of the R dis-

tribution is required for a statistical test of the significance of R (see footnote 1).

The above method for computing R and determining its statistical significance can be extended to more than three rating categories. When there are c categories,

$$R = \frac{\sum_{j=0}^{c-1} \sum_{i=0}^{c-1} n_{ij}|i-j|}{N(c-1)} . \tag{5}$$

Generalizing the logic of the above example to more than three categories yields the following multinomial formula for computing the probability of a given pattern of ratings falling at random in the cells of the c × c repeated ratings matrix:

$$p(d_0, d_1, d_2, \cdots, d_{c-1}) = \frac{N!2^{(N-d_{c-1})}}{d_{c-1}!c^{(2N-d_{c-1})}} \prod_{k=0}^{c-2} \frac{(k+1)^{d_k}}{d_k!} . \tag{6}$$

As with the V coefficient, when the number of raters is large a normal approximation to the exact probability associated with a given value of R can be used. It can be shown that, for any number of categories and a large sample of raters, the probability under the standard normal curve to the right of

$$z = \frac{2N(c-1)(3cR - 2c + 1) - 3c}{\sqrt{2N(c-1)(c+1)(c^2+2)}} \tag{7}$$

is a good approximation to the exact probability. The approximation improves as the number of raters and rating categories increase.