A NOTE ON ESTIMATING STANDARD ERRORS OF FACTOR SCORES IN Q METHOD

William Stephenson
University of Missouri

The standard errors of factor loadings are well known (Burt, 1952; Harman, 1960) whereas those for factor *scores* have received little attention since Spearman's days (1927). The concern in the latter case is with the standard error of scores gained by an *individual* in R method, or by a Q-sample *statement* in the case of Q method. Estimates of standard errors of Q-sample scores are particularly important because of the factor interpretations and hypotheses they support, a matter frequently overlooked in comparisons of R and Q methodologies. It is proposed in this brief note to discuss standard error estimates for use in Q method and to give a reminder of the important difference between R and Q in the use to which such estimates are put.

Standard Error Estimates

The standard error of a score (t) gained by an individual in a mental test is

$$SE_t = s\sqrt{1 - r_{11}}$$

Operant Subjectivity, 1978(Jan), *1*(2), 29-37.

where r_{11} is the reliability coefficient of the test and s is its standard deviation. The same expression applies to each score gained by a statement or item of a Q sample.

In the simple case of one mental test applied to a sample of persons in R, or of one Q sort performed by one person in Q, no difficulties appear because the reliability coefficients can be determined empirically by test-retest. In factor analysis, however, whether in R or Q, factor scores are based on the summation of scores in several mental tests or Q sorts (as the case may be) and an estimate of the composite reliability is required for an error estimate. An approach to it is provided by Spearman's (1913) expression for the correlation of sums and differences.

Let there be p Q sorts, $t_1, t_2, t_3 \ldots t_p$ loaded in one and only one factor T. Let their retest Q sorts be $t_1', t_2' \ldots t_p'$ respectively. The reliability coefficients are $r_{t_1t_1'}, r_{t_2t_2'} \ldots r_{t_pt_p'}$, and the following composite reliability is required:

$$r_{(t_1 + t_2 + \ldots + t_p)(t_1' + t_2' + \ldots + t_p')}$$

Spearman's expression for this correlation of sums involves three correlation matrices, (a) for the $\frac{1}{2}p(p-1)$ correlations between the initial Q sorts ($t_1, t_2 \ldots t_p$), (b) for the $\frac{1}{2}p(p-1)$ coefficients for the retest Q sorts ($t_1', t_2' \ldots t_p'$), and (c) for the p x p correlation coefficients for each t correlated with each t'. It simplifies considerably to deal with the *means* of these three matrices.

Let the means be $\overline{r}_{tt}$, $\overline{r}'_{tt}$, $\overline{r}_{tt'}$, respectively for (a), (b), and (c). The Spearman (1913) expression for the required reliability coefficient is then as follows:

$$r_{(t_1 + \ldots t_p)(t_1' + \ldots t_p')} = \frac{p\bar{r}_{tt}'}{\sqrt{1 + (p-1)\bar{r}_{tt}}\ \sqrt{1 + (p-1)\bar{r}_{tt}''}} \quad \ldots [1]$$

On the assumption that persons doing Q sorts are likely to perform them comparably in the retest situation, influenced only by error, it may be assumed that

$$\bar{r}_{tt} = \bar{r}_{tt}'' \quad (= f^2) \quad \ldots [2]$$

Because the reliability of a mental test, or a Q sort, is usually greater than that due to communality (Stephenson, 1934), it cannot be assumed that $\bar{r}_{tt}' = f^2$. If $\bar{r}_{tt}' = h^2$, where $h^2 > f^2$, expression [1] is as follows:

$$r_{(t_1 + \ldots t_p)(t_1' + \ldots t_p')} = \frac{ph^2}{1 + (p-1)f^2} \quad \ldots [3]$$

Ordinarily different factors have different numbers of Q sorts composing them (p is different for each factor), and the reliabilities can be expected to be different for different individuals. It is burdensome to obtain a reliability coefficient for each Q sort, and even if one did its meaning would be open to doubt because comparable conditions are difficult to maintain in test-retest situations. Values for f and h can be approximated to as follows.

With respect to f^2, the mean correlation of a p x p matrix involving only one factor can be calculated from the mean of the factor loadings of the p variables. If the mean loading is k, then $f^2 = k^2$ (from the fundamental expression for the division of scores into additive factors).

The correlation h^2 cannot be less than f^2 or greater than 1.00 (corresponding to perfect composite correlation). The minimum is too low because of specificities, and the maximum is of course too high. Pragmatically, therefore, one may choose a value for h which is halfway between the minimum and maximum. It is readily calculated from expression [3] by setting the composite reliability at 1.00:

$$1 = \frac{ph^2}{1 + (p-1)f^2}$$

$$h^2 = \frac{1 + (p-1)f^2}{p} \qquad \ldots [4]$$

The difference between this value of h^2 and f^2 is halved and added to f^2 for a guessed value for h^2 for introduction into [3]. To judge by empirical data, very accurate measurements result when six or more Q sorts are added--weighting each with Spearman weights (Spearman, 1927)--where their loadings on the factor are of order 0.60 to 0.70. It is an advantage, therefore, to have standard errors on the small side rather than too large.

The values provided by [3] with this guessed value for h^2 are highly conservative, and in practice smaller standard errors seem to be allowable. This can be said not only because data are meaningful for differences smaller than this guessed value would allow as significant, but because small-sample tests for replication error in variance analysis (Stephenson, 1953) for Q-sample data also indicate that data are significant when expression [3] indicates that they are not (at twice SE levels of significance). For this reason we have found that h^2 can safely be set at higher than the above midway-procedure recommends. In practice, the following empirical expression has been found to be satisfactory when Q samples are of order n = 50 and factors are defined by six to eight

Q sorts, loaded 0.60 to 0.70 on the one factor:

$$r_{pp} = \frac{0.80p}{1 + 0.80(p-1)} \qquad \ldots [5]$$

An Example

The need for concern about standard errors of statement scores stems from their use in the interpretation of Q factors. The same need does not arise in R methodology because interpretation of factors in R proceeds only in terms of mental test factors and their factor loadings: The test scores gained by an individual are used for personnel selection, counseling and the like pragmatic purposes and not for any more detailed inductions. In Q method, on the contrary, interpretation of data begins with consideration of Q-sort factors and their factor loadings, but then penetrates much further into the data in terms of the scores gained in the factors by each statement of the Q sample. Three objectives are served in this way: (i) It is determined how far the initial interpretation (based on Q-sort variables and their factor loadings) is validated by the factor scores gained by each Q-sample statement; (ii) new inductions are made possible, giving rise to hypotheses which had not been considered previously; and (iii) a table of factor scores constitutes a Q-factor model against which additional hypotheses can be tested. The difference between R and Q is nowhere more noteworthy than in this concern with the minutia of factor scores: It corresponds to the fact that items of a population in R are analytic propositions whereas those in Q are synthetic propositions.

Because so much depends in Q upon objectives (i), (ii), and (iii), it is important to have estimates of the standard errors of scores to which attention has been given above.

Consider, for example, the following statement of a Q sample used in a study of attitudes toward public medicine (Stephenson, 1965):

Statement	Normalized Factor Scores A	B	C	D	E
It is time to realize that medical care is already as important as food, housing and clothing in the preservation and enjoyment of life.	1.00	3.00	-0.05	0.50	0.05

One cannot draw further inferences without knowledge of the standard errors of these scores. From the table of factors (i.e., variables and their factor loadings) for the study from which this item of data was taken, factors A to E could be given the following initial explanations: Factor A was for adults of poor socio-economic circumstances, B for medical personnel, C for English-speaking non-Americans (Canadians, British, Indians, Australians), D for individuals in higher income and socio-economic brackets, and E for housewives with particular interests in home nursing. All such interpretation is inference from known (analytic) attributes of the *persons* performing the Q sorts. Most interpretation in sociometric and R studies is based upon such attributes.

In Q method one proceeds much further, however, in terms of the minutia of the Q-sample *statements* as such. Thus, with respect to objective (i) above, one can explain the high score (3.00) gained by factor B by the fact that one expects doctors to be motivated toward medical care, and this confirms the initial explanation of B as being medically-oriented. The low score gained by D relative to A, however, is puzzling, and one wonders whether the explanation is to be found in a *dependency* on doctors which is common to both A and D, but perhaps greater for A than D. This raises a new idea or hypothesis (objective (ii) above) that even the well-to-do feel dependent upon doctors. As for use of factor scores as a model (objective (iii) above), this is a little more complicated, but it consists essentially of expressing alternative hypotheses as Q sorts which are entered into the table of factor scores for A to E, with which they are then compared. Thus an authority on public medicine, Dr. Alan Gregg (1956), made certain proposals about 'Great Medicine,' amounting to a theoretical position on public medicine. Dr. Gregg would have scored the above Q statement +3.00--it was a key theme of his viewpoint. His overall viewpoint is readily represented as a Q sort, using the Q sample providing factors A to E. We found that such a Q sort correlated positively with B but not with the other factors, thus testing the hypothesis that only medically-oriented persons have ideas approximating to those of 'Great Medicine.'

Such inferences, with respect to objectives (i), (ii), and (iii) above, abound in all interpretation of data in Q.

The standard error for factors A to E in the aforementioned study were as follows:

Factor	No. of Q sorts Summed	Mean Factor Loading (f)	Composite[a] Reliability [3]	Composite[a] Reliability [5]	Standard[a] Error [3]	Standard[a] Error [5]
A	7	0.68	0.928	0.966	0.268	0.186
B	6	0.61	0.890	0.960	0.332	0.200
C	4	0.59	0.845	0.941	0.393	0.243
D	2	0.58	0.751	0.889	0.499	0.333
E	5	0.63	0.883	0.952	0.342	0.218

[a]According to expressions [3] and [5], *supra.*

The factors are uncorrelated. The standard errors are in standard terms; expression [3] uses h^2 at midway between its minimum and maximum values; expression [5] depends only on p. There is a considerable difference between the values provided by expressions [3] and [5], being less the larger the mean factor loading and the larger the number of Q sorts entered into the factor. For reasons mentioned earlier, [5] is the recommended estimate for Q samples of size $n = 50$, loadings of size 0.60 to 0.70, and when six to eight Q sorts define the factor.

In the case of the statement with which this example began, the different of 0.50 between scores on D and A has a standard error of amount $\sqrt{0.186^2 + 0.333^2} = 0.381$. The difference is therefore not significant. The scores on A and C differ by an amount 1.05, the standard error of the difference being 0.306: the difference is thus significant.

A number of pragmatic and theoretical problems have been ignored above: it is believed, however, that enough has been indicated to place the statistical rules for inductions at this miniscule level on a sound enough practical footing.

REFERENCES

Burt, C. Tests of significance in factor analysis. *British Journal of Psychology, Stat. Sect.*, 1952, *5*, 109-133.

Gregg, A. *Challenges to contemporary medicine.* New York: Columbia University Press, 1956.

Harman, H.H. *Modern factor analysis.* Chicago: University of Chicago Press, 1960.

Spearman, C. Correlations of sums and differences. *British Journal of Psychology*, 1913, *5*, 417-426.

Spearman, C. *The abilities of man.* London: Macmillan, 1927.

Stephenson, W. Factorizing the reliability coefficient. *British Journal of Psychology*, 1934, *25*, 211-216.

Stephenson, W. *The study of behavior: Q-technique* *and its methodology.* Chicago: University of Chicago Press, 1953.

Stephenson, W. Q-factor models exemplified by a study of attitudes toward public medicine. Unpublished manuscript, University of Missouri, 1965.

Xeroxed copies of the JINNI source deck, plus program write-up, are available for $1.50 (postage included). JINNI is a factor score, FORTRAN IV computer program written for a Burroughs B5700 and is explicitly designed for Q-sort studies. Among other features, statement scores distinguishing each factor from the others, or scores indicating consensus, are determined using expression [5] from the preceding article. The program does not perform a factor analysis, but uses as input the factor matrix as generated elsewhere, e.g., following varimax or judgmental rotation. Contact S.R. Brown, Political Science Department, Kent State University, Kent, OH 44242.
