Validity: Q vs. R

Marten Brouwer

ABSTRACT: This study deals with political ideology, in the form of Q sorts of 80 self-referent statements provided by more than 50 respondents, the P set being cross-sectional. The resulting data set was split up into three equivalent P samples, each of which was submitted independently to Q factor analysis. Also, it was split up into four equivalent item samples, each of which was submitted independently to R factor analysis. Comparison of Q analysis loadings with R analysis z-scores and of Q analysis z-scores with R analysis loadings shows a limited amount of cross-validity. With respect to (meta-)reliability, the three cases of Q factor analysis are strikingly similar; the four cases of R factor analysis, however, turn out to be rather dissimilar.

Author's address: Hobbemakade 27, 1071 XK Amsterdam, The Netherlands. Marten Brouwer is Professor of Political Psychology, University of Amsterdam.

Earlier versions of this paper were presented at the Q conference of the International Society for the Scientific Study of Subjectivity, October 24, 1992, Columbia, MO; and at a conference of the Dutch Association for Political Psychology (NVPP), November 6, 1992, Leiden, The Netherlands.

Introductory Remarks re Q

Although Q methodology had its start in 1935, it has never reached appropriate status in mainstream social science. Yet, it is alive and kicking, even though its main proponent William Stephenson died in 1989. The International Society for the Scientific Study of Subjectivity and its journal *Operant Subjectivity* (now in its 16th volume) testify to that.

The original idea (not out of line with the ideas of Stephenson's teacher, the famous statistician Spearman) seems very simple: correlating persons instead of tests. Superficially, this looks like a rather elementary operation: inverting an "item by subject" matrix to a "subject by item" matrix. There is much more to it, however, with regard to the philosophy of science involved. Summarily, two (not unrelated) points should be referred to here.

First: to take a full string of subjective opinions voiced by one particular subject with respect to an acknowledged topic of interest (like political ideology) implies that the researcher should be aware of the subjective, i.e., self-referent character of such opinions. We are not dealing here with statements of fact, verifiable by external observations, but with statements of communicated feeling (I like, or dislike, such and such ideas or persons or whatever). This focus on "self" has a long tradition in the social sciences. Stephenson himself often mentioned William James as a precursor. We might add here that James was in turn influenced by the French personalist Renouvier (see Kimble et al., 1991, pp. 17-18); and Renouvier by his friend from Brittany, the philosopher Lequier whose ideas on the "libre arbitre" of the self may well have contributed to his suicide in 1862 (see Turpin 1978).

Second: thinking in terms of collections of self-referent statements, which come to light by the act of measurement, is reminiscent of quantum mechanics. Stephenson was a physicist himself, close to Rutherford and other luminaries. He wrote many essays on the implications of the similarity (and not just the analogy) between his Q methodology and quantum mechanics, stressing the fact that orthogonal factors in the

analysis of subjectivity do correspond with the "states" of a phenomenon as distinguished in quantum theory (see *inter alia,* Stephenson, 1983; Brown, 1992).

Introductory Remarks re Validity

Traditionally, validity has not been an issue in Q. In fact, it has often been argued that "the concept of validity has very little status since there is no outside criterion for a person's own point of view" (Brown, 1980, pp. 174-175). On many occasions, Stephenson himself has also written and said that validity (and reliability) should not be held to be relevant to problems in Q. It may seem rather adventurous, therefore, nevertheless to engage in a project investigating the validity and reliability of data gathered with proper Q methodology.

Validity refers to the relationship of an instrument to the phenomena it pretends to measure; or, more specifically, to the uses to which an instrument is put. In the case of predictive validity, one investigates the relationship with another variable (mostly an outside criterion). In the case of content validity, one investigates whether the instrument can be reasonably said to represent the area of phenomena concerned. In the case of construct validity, one investigates whether the instrument produces results which are generalizable to a theoretically meaningful domain of observables. The latter sort of validity has been explained beautifully by Nunnally (1978, p. 109):

> ...one could rightly argue that all this fuss and bother about construct validity really boils down to something rather homespun -- namely, circumstantial evidence for the usefulness of a new measurement method. New measurement methods, like most new ways of doing things, should not be trusted until they have proved themselves in many applications. If over the course of numerous investigations a measuring instrument produces interesting findings and tends to fit the construct name applied to the instrument, then investigators are encouraged to continue using the instrument in research and to use the name to refer to the instrument. On the other hand, if the evidence is dismal in this regard, it discourages scientists from

> investing in additional research with the instrument, and it makes them wonder if the instrument really fits the trait name that had been employed to describe it. From the standpoint of the work-a-day world of the behavioral scientist, essentially this is what construct validity is about.

Even when taking exception to the word "trait" here, one cannot but agree with the general tenor of this explanation. It is in a similar vein that the present project has been envisaged. It does not try to compare Q data with some sort of external criterion. Instead, it remains completely within the area of subjective utterances: persons' own points of view therefore. At issue is the validity of Q methodology in analyzing subjective phenomena versus a comparable R approach to the very same phenomena, the latter being much more en vogue (see Turner & Martin 1984).

Some authors on Q have already ventured into the problem area of validity. Fairweather (1981) has found two separate Q studies to produce similar results, and from that concludes that they must be representative of the population and therefore have external validity, a clear example of a *non sequitur* (if anything at all, this might be a matter of reliability). He adds the finding that his Q data do to a certain extent match demographic data, and then concludes to some construct validity: this is not very convincing either (rather a simple case of correlation with a background variable). Dennis (1988) also provides some comments on the validity of Q methodology, yet does not advance much beyond stressing the necessity of qualitative checks.

The most interesting study on validity and reliability (or, more generally, generalizability) may well be the one by Thomas and Baas (1991). They take issue with the usual approach of Q researchers in the following way:

> That respondents whose Q-sorts define different factors take a different view of the facts, however, does not *ipso facto* render the data so provided as artifactual. On the contrary, these data are scientifically compelling precisely *because of their status as facts*. Hence the difficulty of consigning matters of subjective understanding to a domain "conceptually in-

> dependent" of matters of fact. Viewed thus, the concern with generalization in Q-methodology is, we believe, neither misplaced nor unwarranted. To hold that findings from Q-studies should not be held accountable to *any* standards of replicability appears to us, at the very least, as an unproductive course of action. Granted, the quest for replicative rigor should not serve as the *sole* criterion for assessing methodological acuity; certainly, to the extent that such preoccupations preclude or interfere with the discovery of heretofore unimagined possibilities, such concerns may indeed impede scientific progress. On the other hand, there is little to be gained by dismissing such questions as simply inappropriate. (p. 4)

The present author could not agree more. Unfortunately, the research design of Thomas and Baas in fact takes a step backwards (at least in terms of Expositor's [1987] principle of "reliable schematics") by focusing on the comparison of two differently conceived Q studies. By doing so, they can only compare factor-score composites, which means that "... there is no simple way to assay their 'degree of fit' by means of a summary statistic on the order of a reliability coefficient" (p. 24). For many Q researchers, that is already going too far; for the present author, however, it is not going far enough.

It should be feasible to establish some aspects of the validity (and reliability) of Q methodology vs. R methodology and vice versa without having recourse to external facts and without reliance on further checks. Or, to be more specific: instead of comparing different sets of raw data (subjective or otherwise), one may use one and the same set of subjective data and then proceed to analyze these data, both according to usual R procedures and according to usual Q procedures. Validity would then apply to the comparability of the results of different procedures of analysis, not different sets of data.

Most users of Q may be quite aware of the fact that Stephenson himself has already argued in 1936 that such double factorisation of one and the same matrix may be done only under the very special condition of universality of measuring unit (Stephenson, 1936). Brown (1980, pp. 13-14) has provided a striking example of the results of doing so with the measurements of human body segments in terms of inches: R then

provides an almost unintelligible factor matrix; whereas Q provides a one-factor solution, pertaining to the obvious observation that the relative proportions of various body segments amongst human beings are roughly similar, and that most such measurements in inches reflect mainly differences between persons as to total size (height and arm length becoming the main factor scores). Nunnally (1978, p. 428) states in this same respect, "Even if it were true that transforming one type of analysis to the other usually would be possible, this would be beside the point. To think otherwise would be analogous to thinking that because the same machine could be used to measure heart rate and brain waves, it would make no difference which was measured." Warnings all around.

Nevertheless this paper will try to provide some evidence *re* validity in an analysis which is somewhat analogous to Brown's example. The data were collected in the form of Q sorts, with subjective statements only. These subjective statements, however, were culled from the well-researched domain of political ideology in terms of R research, the main difference being that in the present case they are not administered in the usual absolute form of R ("to what extent do you agree with such-and-such?"), but in the form of Q sorts. It is clear that such data should nevertheless be amenable to R analysis as well as Q analysis. In doing so, the focus will be on the comparison of systems of analysis instead of comparing sets of data.

One final introductory remark: this comparison does not aim for a one-sided interpretation of whatever will be found. The extent to which the findings concur will be considered to constitute a mutual validation of the two systems of analysis. To the extent the findings differ, the next step will be to try and explore the merits of each system separately.

Description of the Data

The 80 items for the present study originated in Middendorp's long standing research efforts with respect to political ideology (see *inter alia,* Middendorp, 1978, 1991). They range from

statements like "Differences between high and low incomes are still too large" through "It should be free to demonstrate for or against something" to "It should not be possible for a woman to have an abortion only because she so wishes." For the present project (see also Middendorp & Brouwer, 1992), Middendorp himself divided the items into two presumably equivalent stacks of 40 items each, to be called the white stack and the yellow stack. The subjects then were asked to Q sort each of these stacks separately (with a random assignment per person as to which of the two stacks would be first).

Fieldwork was carried out as a cross-sectional survey by the Netherlands Institute for Public Opinion (NIPO) in the spring of 1992. The target number (with a maximum set at $3 \times 27 = 81$ interviews) was $3 \times 20 = 60$ interviews. Each interviewer was instructed to produce three interviews. The resulting 70 interviews could not all be used: in some cases an interviewer had produced only two instead of three interviews; in some other cases the recording of the answers was flawed. All in all, rigorous scrutiny left 54 cases for analysis, produced according to instructions by 18 interviewers -- therefore $3 \times 18 = 54$ interviews. Thus, the data matrix in this project consists of $54 \times 80 = 4320$ personal item scores. Item scores range from 1 to 9 and are distributed normally for each individual person, as determined by the Q sort instructions, which asked for degree of subjective agreement.

It might have been possible simply to perform two factor analyses on this matrix, one horizontally and one vertically. Two objections may be raised against such a procedure, however. First, there are theoretical reasons not to factorize a matrix which is rather close to being square, although Arrindell and Van der Ende (1985) have shown that even N-to-n ratios of 1.3 to 1 may produce recognizable factor solutions. Second, and at least as important: the usual Q procedure turns out to employ a large number of items in relation to the number of persons, whereas the usual R procedure turns out to employ a large number of persons in relation to the number of items. In order not to deviate too much from usual proce-

dures in both Q and R, the data matrix was split up in the following ways.

For the purposes of *Q analysis,* the sample of 54 persons was divided (on the basis of clustering by the interviewer) into three equivalent subsamples of 18 persons each -- respectively, P(1), P(2) and P(3). For each of these three P samples, a factor analysis was performed with all 80 items. This led, in each separate Q analysis, to a factor matrix with 18 person loadings per factor, plus 80 item z-scores.

For the purposes of *R analysis,* the sample of 80 items was subdivided into four subsamples of 20 items each. The first step was simply to split the data according to the two stacks in which the items were administered (white and yellow); the second step was to divide each set randomly (in fact the odd vs. even numbered items) into two subsets. For each of these four I samples -- respectively, I(wo) (=white odd), I(we) (=white even), I(yo) (=yellow odd) and I(ye) (=yellow even) -- a factor analysis was performed with all 54 persons. This led, in each separate R analysis, to a factor matrix with 20 item loadings per factor, plus 54 person z-scores.

Table 1
Person Data Matrix

	Q1 loadings	Q2 loadings	Q3 loadings
Rwo z-scores	n=18	n=18	n=18
Rwe z-scores	n=18	n=18	n=18
Ryo z-scores	n=18	n=18	n=18
Rye z-scores	n=18	n=18	n=18

The combination of the seven factor analyses (three Q analyses and four R analyses) produces the 12-cell abstract table for data on persons as shown in Table 1. Any comparison (correlation) between the findings per person from Q and the findings per person from R can be made in 12 (semi-)inde-

pendent ways. Similarly for the findings per item, as shown in Table 2.

Table 2
Item Data Matrix

	Q1 z-scores	Q2 z-scores	Q3 z-scores
Rwo loadings	n = 20	n = 20	n = 20
Rwe loadings	n = 20	n = 20	n = 20
Ryo loadings	n = 20	n = 20	n = 20
Rye loadings	n = 20	n = 20	n = 20

One can now calculate, for each cell, a meta-correlation between Q findings and R findings. This can be done relatively independently for 12 different cells -- independent in the sense that it will always concern a different combination; only relatively independent, however, since the factor analyses concerned will always have been based on a full dataset of 54 or of 80 entries.

Hypotheses of This Validity Study

In formulating expectations as to the results of the calculations mentioned, we try to start out on the conservative side. If there should be a central phenomenon of subjective political ideology, which is being tapped both by Q analysis and by R analysis, then we expect similarity between the outcomes for persons as well as for items. In fact, the research findings of Middendorp (*v supra*) would indicate that, at least in R, there are two distinct factors to be expected. There is no reason why this should not already be true for the nonrotated factor matrices for at least one factor. We may therefore formulate:

Nonrotated First Factors Only

Hypothesis 1(P): the combined probability for all 12 cells of the meta-correlations between the 18 loadings from Q and the 18 z-scores from R will fall below .001.

Hypothesis 1(I): the combined probability for all 12 cells of the meta-correlations between the 20 z-scores from Q and the 20 loadings from R will fall below .001.

If these two hypotheses prove to be acceptable, we may proceed to study the results of nonrotated second factors; of rotated factor structures; etc. etc. In the original version of the present design, hypotheses were formulated as to such further testing. In the light of the findings of the next paragraph, however, they have been deleted, both for the sake of brevity and in order to evade superfluous pedantry.

Findings re Validity

With regard to *Hypothesis 1(P),* the correlation coefficients in Table 3 were obtained. The hypothesis has to be discarded in favor of the null hypothesis (p being far above .001).

Table 3
Metacorrelations for Hypothesis 1(P)

	Q1 loadings	Q2 loadings	Q3 loadings
Rwo z-scores	.77	.16	.05
Rwe z-scores	.09	.63	.63
Ryo z-scores	-.31	-.39	.16
Rye z-scores	.07	.11	-.36

Similarly for *Hypothesis 1(I)* (see Table 4): The hypothesis has to be discarded in favor of the null hypothesis (p being above .001).

Table 4
Metacorrelations for Hypothesis 1(I)

	Q1 z-scores	Q2 z-scores	Q3 z-scores
Rwo loadings	.61	.49	.41
Rwe loadings	.59	.68	.64
Ryo loadings	-.01	.24	.01
Rye loadings	.03	-.05	.06

Exploration re Validity

The findings are disappointing, yet they may allow for some exploration. One may first observe that, generally speaking, the results for the cells belonging to the white stack (both odd and even) are much better than those for the cells belonging to the yellow stack. In fact, if the analysis were to be carried out with only the data belonging to the white stack, both hypotheses would have been acceptable at the pre-established level of significance. Maybe the two stacks are not as equivalent as they were meant to be.

One way of exploring this possibility is to repeat the Q factor analysis splitting the data into the items from the white stack and the yellow stack, respectively (which is actually much closer to the Q way of administering stimuli as performed). This leads to the meta-correlation coefficients (for P) shown in Table 5.

This seems to indicate very clearly that the white and the yellow stacks are indeed rather different, at least in their relationship to the pertinent z-scores. Again, if the project had been carried out with the white stack only, the null hypothesis would have been rejected at the specified level (.001). Within the limitations of exploratory analysis (and within the limitations of nonrotated first factors), we may therefore conclude

Table 5
Metacorrelations in P

	Loadings		
white stack	Qw1	Qw2	Qw3
Rwo z-scores	.68	.18	.06
Rwe z-scores	.50	.49	.65
yellow stack	Qy1	Qy2	Qy3
Ryo z-scores	-.19	-.12	.12
Rye z-scores	.46	.24	-.09

that there is, after all, a good deal of evidence for the mutual validation of Q and R.

Two more observations are pertinent here. First, the white stack is upon inspection much closer to the time-honored investigations of Middendorp than the yellow stack is. Second, at least one of the subjects cooperating with the present author in an additional small study spontaneously commented on the difficulties she encountered in Q sorting the yellow stack as compared to her Q sorting the white stack (mentioning, *inter alia*, that the white stack did not contain statements starting with "I think," etc., whereas the yellow stack had several such statements).

Exploration with Rotations

Another exploratory way of looking at the findings is to check on the amount of variance explained by the first nonrotated factor in all seven cases of factor analysis. For Q, the relevant percentages are 24%, 28% and 22%, respectively. For R, the relevant percentages are 9%, 6%, 7% and 8%, respectively. It should be clear that, in this stage of the analysis, Q is superior to R.

Rotation of factors does not basically change the picture in this respect. Varimax rotation on the basis of five selected factors out of seven, for example, still leaves us with a situation where the three Q analyses each show one considerable first factor, whereas the four R analyses show modest amounts of explained variance spread out over many different factors.

Exploration re (Meta)Reliability

One of the most compelling findings in this exploratory analysis refers to the issue of reliability (or rather: meta-reliability). The focus here is not on the direct comparison of answers in a test-retest situation or in a split-half construction. Rather, the first nonrotated factor findings from the seven factor analyses are to be compared as to their z-scores. Within R, the four analyses were carried out separately for the four item groups wo, we, yo and ye; this resulted in four different estimates of the z-scores of the subjects (Table 6). These z-scores per person may be correlated to estimate meta-reliability. The findings are called "meta-reliability" because they do not pertain to direct comparisons between raw test and retest data, or between the two parts of a split-half procedure, but to the comparison of higher order results stemming from comparable data.

The obvious conclusion is that no minimum degree of reliability can be observed here. In view of the fact that many studies in R are actually carried out with some 20 items as the basis for a condensation by factor analysis, this may provide food for thought.

Is it different for the analogous check on meta-reliability in Q? Let us have a look (see Table 7). The pattern here is quite different from the one for meta-reliability in R: all coefficients are positive, high and significant. Even if we drop three interviews, carried out by one and the same interviewer, where the raw data turned out to be uncannily similar, the same conclusion obtains. For ease of representation, the coefficients of meta-correlation for the latter data set, with

Table 6
Meta-Reliability in R (Person z-scores)

	P1	P2	P3
wo vs we	.18	-.44	-.24
wo vs yo	-.26	-.18	.16
wo vs ye	.04	.57	-.01
we vs yo	-.14	-.26	.17
we vs ye	-.16	-.14	-.31
yo vs ye	-.30	-.16	-.43

Table 7
Meta-Reliability in Q: 18 (or 17) Clusters (Item z-scores)

	P1 vs P2	P1 vs P3	P2 vs P3
wo	.83 (.81)	.77 (.78)	.83 (.85)
we	.95 (.95)	.92 (.91)	.93 (.91)
yo	.71 (.72)	.84 (.84)	.83 (.84)
ye	.87 (.86)	.81 (.78)	.88 (.87)

$17 \times 3 = 51$ subjects (instead of the basic data set with $18 \times 3 = 54$ subjects), have been added within brackets.

Q researchers may rejoice in this. Yet, some notes of caution may be in order. What we have here is a data set in which the four sets of items, constructed to be equivalent, do

not turn out to be so. The equivalence of the three sets of persons, however, seems to be quite present. Also, the Q-sort technique itself may have forced the respondents into a pattern of answering which turns out to be more congenial for a Q factor analysis than for an R factor analysis.

As far as the latter is concerned: the distribution of answers per item was of course normalized in order to calculate product-moment correlation coefficients. Even though that is the usual procedure in R studies employing factor analysis, it is at the same time crystal clear that it is doing away with part of the information in the data set. One would like to repeat the present study without the constraints of "literal" Q-sort technique. According to Brown (1980), that should not materially influence the findings within Q; it might change the results within R, however.

Concluding Remarks

Tentative conclusions from this project might be:

1. It is quite feasible to use the same data set for both Q and R analysis.
2. To a limited extent, Q analysis and R analysis do show similarity, which may be interpreted as validation both ways.
3. Given the Q-sort instruction, Q proves to be superior to R with respect to meta-reliability.

Several points of doubt remain, however. Would not the excellent results for meta-reliability in Q be due to a sort of combined interviewer plus location effect (although the exclusion of the one cluster of three interviews with highest mutual correlations [see Table 7] indicates that this produces only negligible changes)? Would the present results hold true with a less strictly Q-oriented instruction? Would a different way of rotation (e.g., theoretical) lead to a different perspective?

As always: this calls for further research. To finish on one last quote:

> It is to the credit of contemporary social science that it has developed to a fine point the science of asking important questions; what is lacking is an equal development of the art of listening to the answers rather than transforming them, through a kind of behavioral alchemy, into something else. (Brown, 1980, p. 3)

Acknowledgements

The author acknowledges the help of the following persons: Steven Brown, for his bibliographical help and for his gracious comments; Bram Cornelisse, for his clever data file manipulation and general assistance; Peter Korsuize, for his computer programming and for his many comments; Cees Middendorp, for his theoretical input and for securing funds for the project; Jan Stapel (and the NIPO survey institute), for his generous attention to the project.

References

Arrindel, W.A. & J. Van der Ende (1985) An empirical test of the utility of the observations-to-variables ratio in factor analysis. *Applied Psychological Measurement, 9,* 165-178.

Brown, S.R. (1980) *Political subjectivity: Applications of Q methodology in political science.* New Haven, CT: Yale University Press.

Brown, S.R. (1992, October) Q methodology and quantum theory: Analogies and realities. International Society for the Scientific Study of Subjectivity, Columbia, MO.

Dennis, K.E. (1988) Q-Methodology: New perspectives on estimating reliability and validity. In O.L. Strickland & C.F. Waltz (Eds.), *Measurement of nursing outcomes: Vol. 2. Measuring nursing performance: Practice, education and research* (pp. 409-419). New York: Springer.

Expositor. (1987) Reliable schematics. *Operant Subjectivity 10,* 81-86.

Fairweather, J.R. (1981) Reliability and validity of Q-method results: Some empirical evidence. *Operant Subjectivity, 5,* 2-16.

Kimble, G.A., M. Wertheimer, & C.L. White (Eds) (1991). *Portraits of pioneers in psychology.* Hillsdale, NJ: Lawrence Erlbaum.

Middendorp, C.P. (1978) *Progressiveness and conservatism: The fundamental dimensions of ideological controversy and their relationship to social class.* Berlin: Mouton.

Middendorp, C.P. (1991) *Ideology in Dutch politics: The democratic system reconsidered, 1970-1985.* Assen: Van Gorcum.

Middendorp, C.P. & M. Brouwer (1992, June) Quantitative "subjective" individual data analysis in political psychology: An application of Q-methodology in assessing individual positions in an ideological space. International Conference on Social Science Methodology, Trento, Italy.

Nunnally, J.C. (1978) *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Stephenson, W. (1936) The foundations of psychometry: Four factor systems. *Psychometrika, 1,* 195-209.

Stephenson, W. (1983) Quantum theory and Q-methodology: Fictionalistic and probabilistic theories conjoined. *Psychological Record, 33,* 213-230.

Thomas, D. & L. Baas (1991, November) The issue of generalization in Q-methodology: "Reliable Schematics" revisited. Southern Political Science Association, Tampa.

Turner, C.F. & E. Martin (Eds.) (1984) *Surveying subjective phenomena.* New York: Russell Sage.

Turpin, J-M. (1978) *Sol ou Jules Lequier.* Paris: Hallier.