Meehl, P. E. (1959). Q-technique, pros and cons. In B. C. Glueck (Ed.),
Report on Conference on Social Adjustment Rating Scales (pp. 11-18).
University of Minnesota Center for Continuation Study.

Report of Conference

on

Social Adjustment Rating Scales

held at

Center for Continuation Study

University of Minnesota

October 10, 1958

Under the auspices of

Department of Psychiatry,

The Medical School, U. of M.

and

Psychopharmacology Service Center, National Institute

of Mental Health, Bethesda, Md.

Edited by

Bernard C. Glueck, Jr., M.D.

Q-technique, Pros & Cons

Paul E. Meehl, Ph.D.

University of Minnesota

I'm supposed to talk about Q-technique in the assessment of personality. As some of you know, this method has become sort of a fad among psychologists, and I am not, let me say right off, passionately identified with the "Q-technique movement," nor do I believe that the theoretical account that has been given by Dr. Stephenson [1953] is correct in all respects. But I don't believe that his view of the method is necessary in order for one to use it in assessment. I know that some of you are very familiar with the method, but others of you aren't, so that I'll have to say a little as to what Q-technique is about.

Differences between R and Q Correlation Matrices

TYPE R

Patient Name	X (dominance)	Y (aggressiveness)
Jones	X_j	Y_j
Smith	X_s	Y_s
Brown	X_b	Y_b
⋮		
Fisbee	X_f	Y_f

TYPE Q

Trait	X (Jones)	Y (Smith)
Dominance	X_d	Y_d
Aggression	X_a	Y_a
Sweetness	X_s	Y_s
Fluency	X_f	Y_f
⋮		
Passivity	X_p	Y_p

The traditional correlation coefficient is computed by considering similarity or association between two columns of numbers. We have a variable X (say, "dominance") and a variable Y (say, "aggressiveness") and then we have persons Jones, Smith, Brown, etc., i.e., our range of values is over a collection of individuals. In computing the traditional coefficient of correlation we associate, in pairs, the dominance and aggression scores (or ratings) of Jones, Smith, Brown and so forth, and the coefficient is then an expression of the degree to which these arrays of numbers "go together." Now in the Q-type of correlation, each person is treated statistically as a variable. X, [one] entire column, is Jones; and Y, [another] column, is Smith. The first pair of ratings or scores are on "dominance," the second pair are on "aggression," the third pair are "sweetness," the fourth "fluency," and so on. What the correlation coefficient tells us when thus calculated is the extent to which Jones' pattern on this series of traits is like or unlike Smith's pattern on the same series of traits. Now of course there are certain problems that arise here immediately in regard to the metric. Even if we happen to be rating all of the traits there are difficulties; and suppose that we happen to have mixtures of ratings and measures. Some of these might be in I.Q. units, whereas the dominance variable is in terms of a seven-step rating scale, the blood turpentine titer is in yet another kind of unit, and so forth. You can see that the arbitrary nature of the size of these units, to say nothing of their distribution properties, will lead to absurdities in interpreting the coefficient unless some way is taken to standardize them. But the essential idea is that when we speak of a Q-correlation, we mean that we are correlating one person with another person. We are asking to what extent the "pattern" or "profile" or "configuration" of the one individual is like (or unlike) that of another individual.

This idea actually goes back a long time. I believe that the first use of correlation of persons (rather than variables in the usual sense) was about 1912, a study in England on types of imagery, But the man who put this method on the map, and who should get the credit for it, is an English psychologist now in this country, William Stephenson, formerly at the University of Chicago. He is the person who, beginning in the middle 1930s, began to write in a systematic way about this method, to present data about its use, and convinced the profession of its value. It took us a while to get around to realizing how important this method was.

There is a variation in the way the terminology is employed. Some of you may feel that I am not using the word “Q-technique” properly because I haven’t said all the things that are sometimes said to define it. I would prefer myself just to talk about person-person correlation” and avoid the controversial phrase “Q-technique,” because this term commits us, some people think, to accept the whole family of ideas associated with the historical defense of person-person correlation. For example, one doctrine that Stephenson has associated with the technique is that to make the best use of this you have to employ what Cattell [1944] calls an “ipsative” frame of reference in making the ratings. There are three kinds of frame of reference or norms that the psychologist may use. Using Cattell’s language, first we have what is called the interactive kind of unit, which is essentially a physicalistic unit, denoting the organism’s interaction with its environment. Thus, how many degrees a person moves an arm is an interactive unit; or, in industrial psychology, how many pounds of butter somebody can pack in a certain period of time is an interactive unit. The customary thing in psychology is the so-called normative frame of reference, where we express a person’s position in terms of where he stands in relationship to some population of persons. We say he’s “at the 83rd percentile,” or that he’s “up one standard deviation.” It is customary to transform interactive scores into normative scores, because interactive scores by themselves (in most situations) don’t tell us very much.

Thirdly, we speak of ipsative measurement. Here we express a person’s standing on some variable not in physicalistic units, and not in relationship to what other people do, but somehow in relationship to the other things that he does. Stephenson has argued that the essence of the Q-technique, what makes it a good method, is that the frame of reference be taken in terms of the individual or the population of characteristics possessed by the individual. Many psychologists doubt that this can really be done. I think myself, for instance, that in usage of trait names there is usually an implicit covert reference to the normative, person-population frame. I have read Stephenson on the subject, and I’ve argued with him about it, but I still don’t understand his belief that you eliminate the normative frame of reference by giving some particular kind of instructions to the Q-sorter. If I rate a human being (or if he rates himself) on a trait like “shy,” and we ask ourselves to defend a given rating by taking episodes of behavior that would support that rating, I think it almost always turns out that this defense involves some kind of a reference to what is the normal social expectancy with regard to behavior, and that’s what leads us to classify the concrete episode as supporting a rating of high or low on the variable “shy.”

The question, how much difference does it make for the technique whether you give essentially normative or ipsative instructions to the rater, or whether you leave it deliberately ambiguous, is an empirical question not yet resolved by evidence. There is some evidence in the literature (such as Jack Block's [1956]) to indicate that we don't have to get too excited about these theoretical issues, because in practice the inter-person correlations produced by the two kinds of instructions to raters may approximate that permitted by their reliabilities, and if that's true we don't have to worry too much about these theoretical points.

Another problem which is associated with this technique is the problem of constructing a domain of traits to be sampled. In, our own local work on the Ford Project, we have begun by constructing a domain that says practically everything that anybody can think of saying about a person; that means we have, after screening, something like twenty-five hundred items, still to be further reduced. We are checking for the adequacy of the "coverage" by several methods, such as taking episodes out of novels, out of therapy protocols, out of social work histories, and out of the Old Testament. Using random numbers on these various sources, we've pulled out short behavioral "episodes," which episodes are presented singly to people (some psychologists, some psychiatrists, and some bright laymen) and they're told, "Tell us all of the things that come to mind about a person who would do this kind of thing. Never mind whether you could prove it, or whether it's completely out in left field. Just write down what occurs to you." We are studying the convergence to an asymptote as you add more and more of these associations of readers of episodes; i.e., how fast do you get complete coverage? This is one way to check on whether we have in our pool "everything that anybody can think of" when presented with a single behavior episode.

Now what's the point of person-person correlation anyway? Why did it make such a splash? What are its special advantages and disadvantages? Even if you set aside Stephenson's particular philosophy of the method, it provides one possible statistical approach to the concept of "type" or "syndrome," which has been needed for a long time. It is not, in my view, clearly the optimal approach to this problem. Personally, I think that the mathematical model of what "type" means in taxonomy, or in medical diagnosis, has yet to be satisfactorily formulated by the statisticians. But Q is at least better than R (the traditional correlation over a range of persons) for this purpose. The word "type" has several meanings not all equally related to person-person correlation methods. But one of the legitimate meanings of "type" is this: A type (or syndrome) exists whenever the incidence in a population of persons who deviate by more than a specified amount on all of the variables defining the syndrome is much higher than you could predict if you just looked at the first-order R-correlations of the variables by pairs. Some psychologists have been a little naive, it seems to me, in the application of statistics to typological concepts. For instance, take such a concept as the "anal character" as described by Freud, Abraham, Jones and others; it is not really appropriate to check out Freud on this idea by R-correlating the traits neatness, parsimony and obstinacy over the person population. If you go back to the original paper, Freud [1948] did not say, or imply, that neatness, parsimony and obstinacy would have high pairwise Pearson *r*'s over the person population. What he said was, that in analytic work one runs across, from time to time, individuals who show extreme degrees of all three of these characteristics in combination; and that is a very different kind of claim statistically from the kind of claim you're investigating if you calculate three Pearson *r*'s over people in general.

The Q-technique can sometimes enable you to find types or taxonomies, even without any fancy statistics like cluster or factor analysis, simply by inspection of the matrix of inter-person coefficients. In such a group as ten manics and ten depressives rated or measured psychometrically on a set of dimensions relevant to this disorder, and intercorrelated by pairs so that each patient is correlated with each other patient, the two subsets will be obvious at a glance. There are ten people who correlate high with each other pairwise, and there are ten other people who correlate high with each other pairwise, but these second ten could correlate low or negatively with the first ten (i.e., these cross-group coefficients would be low or negative).

Another important thing about the technique is that it can be shown theoretically (and it has been shown empirically by Block [1955]) that there are situations in which the traditional R-technique will not give you the information that you want unless it is preceded by Q-technique. The interesting situation is that in which sub-populations organize their trait structure in very different ways. This is an old point, made by Gordon Allport [1937] back in the 1930s. Allport didn't, perhaps, make the point in the most rigorous fashion, operating within the mathematical frame of argument, so he didn't get much agreement. But he was right. You may be dealing with two sub-populations in a clinical population, for one of which the Pearson r's over persons have a certain pattern, whereas within the other sub-population a very different pattern of correlation prevails, including for example, even reversals in sign from +.70 to –.65. If we were to consider these two sub-populations as one, and calculate a traditional Pearson r over the range of persons, what we would get would depend simply upon the proportions of the two sub-populations, and might be near zero. Eysenck, studying political attitudes, showed that among Communists direct and indirect aggression correlate –.94, whereas for Fascists they correlate +.6l. If you take a sample of Communists and Fascists mixed, and you don't begin with Q-technique, but perform a traditional R-analysis, you would get a correlation that depends upon the proportions of the two. For a neutral group, Eysenck reported a correlation of –.64.

The same thing was shown in personality traits of Air Force officers by Block working at IPAR. If you cluster analyze by Q-technique 100 Air Force captains, you get a big dimension running through the sample which is "degree of control." On the one end you have the high-controlled, inhibited, constricted, spastic, anal sort of character, and at the other end you have the flamboyant, dilated, freewheeling, under-controlled kind of individual. What you find if you take the individuals at these extremes is that, while they differ greatly on a trait like "rigidity," they both rate very high on the trait “sarcasm." Now how does this work psychologically? Presumably for the constricted characters, sarcasm is a sort of aim-inhibited, more or less acceptable form of hostility, where they handle their own inadequacies and get out their aggressions in a way that is experienced as not too dangerous by them; for the other, free-wheeling characters, sarcasm is like any other open expression of the impulse life, i.e., they are sarcastic, like they are anything else, chiefly because they are not restrained. But the psychological organization is different, and this difference is reflected statistically. If we R-correlate "rigidity" or "control"—the big variable in this group—with "sarcasm" over the entire officer population, we get an insignificant correlation of –.11. The traditional tactics of computing the Pearson r over a range of persons without first isolating the sub-syndromes who organize their R-correlations differently would lead you not to discover this important piece of truth. I think that one of the big selling points of Q-technique is that it is somehow closer to the way clinicians think; it gives us a better lead on the distinctive way in which each person organizes his traits. Another example Block reports is that "impulsivity" correlates +.54 with “suggestibility" among Air Force captains, whereas among male graduate students at the University of California these two traits correlate –.50. Here there happens to be an "extrinsic," obvious (demographic) basis for separating the two populations. But the point is that often there are personality variables (rather than some crude, extrinsic, or demographic basis of classification) which can be discovered by the use of a technique like Q-correlation.

Another nice feature of the technique is that it enables you to do such things as showing the extent to which an individual moves toward or away from a certain standard pattern, such as the "ideal." In the outpatient study we did here last year with a couple of new tranquilizers, we wouldn't have had much by way of results if we had not used the Q-sort. That was the technique that gave us our best positive findings. One of the methods used there was to have several “experts," who are presumed to know what it is to be "well-adjusted," Q-sort "the Ideal 'Cured' Out-patient." Then we averaged these ratings. It was reassuring to find out that a half-dozen clinicians at least agreed among themselves as to what it was to be the Ideal "Cured" Out-patient (the disagreements were also revealing!). That stereotype is then used as a basis for evaluating the change in each patient, before the drug and after the drug, how well he is Q-correlated with "the Ideal ‘Cured’ Out-patient." One of the advantages of the method is that sometimes a rater (interviewer) is not able to make reliable or valid global judgments as to how much somebody is improving. This is partly because a therapist or interviewer doesn't even remember (as you find out when you listen to earlier statements from a series of interviews) exactly what the person's complaint was, or what, and how extreme, the high and low points were. However, if you make him record a set of judgments on a more atomistic, restricted piece of behavior repeatedly, and then treat these more restricted judgments by a statistical technique such as Q-correlation, you are able to show that there is a change in the pattern toward or away from the ideal.

The same thing can be done if you have an interest in studying diagnostic error. Let us suppose, for example, that one reason why a certain drug doesn't "work,” or why the MMPI or Rorschach is "missing" cases, is that the diagnoses themselves are wrong, so that we believe that a sizable proportion of patients in the

Out-patient Department that are officially called "psychoneuroses, mixed" or whatever, are really ambulatory schizophrenias. Well, do we want to rely upon the diagnostic judgment of the clinician for this purpose? Not necessarily, since it is partly his diagnostic errors that are under study. What you can do is to rely upon his judgment for these smaller, segmental aspects of behavior, and you can then have somebody else, who is interested in pseudoneurotic schizophrenia and has devoted his attention to it, do a Q-sort on "What is the typical pseudoneurotic schizophrenic, who gets mis-called neurotic, like?" Then you can use raters, some of whom may not even believe in the concept of pseudoneurotic schizophrenia, but who can nevertheless discriminate the particular pieces of behavior that are involved; they can Q-sort the patients and then you can find out whether those individuals on whom the Rorschach or the Multiphasic "missed," or where the drug didn't work in the expected way, are the individuals tending to be more highly Q-correlated with one another, and with the idealized pseudoneurotic schizophrenia Q-sort. Hence you can use the ratings of clinicians who might not "believe in" what you are up to at all, or who lack the needed diagnostic skills, but who are able to make discriminations at a more atomistic level.

Well, I have said some of the good things, so now we'll take a couple of minutes to talk about some of the bad things and unsolved problems. One of the worst problems, which has not been solved, nor much progress made in solving it, is the problem of correlated errors. That is, we cannot assume that just because two "independent" judges show high reliability in terms of high Q-correlations in rating the same patient, therefore they are getting at the truth. There are several possible reasons for such high correlations. For instance, suppose that you are using a Q-pool of 100 items and only one item is discriminable. Only one—let's make it very bad for the investigator. Placement of the other 99 items is all error. Now suppose we attempt to show that people who cluster together in their Rorschach profiles are also clustered together clinically as the therapist Q-describes them. There Q-correlations might be quite impressive, but they might occur on the basis of this one variable—such as "depression"—that is being validly discriminated. That is, the sorter makes the (valid) discrimination that the patient is depressed to degree X; and then, because he went to medical school or studied psychology, he thinks "Well, we know about depressed people, they have this other trait, and they have that, and this, and that…," and so he sorts out the non-discriminable items accordingly , Similarly the other clinician, who also went to school, does the same thing, and so they correlate with each other on a whole array of items, but the only thing they're actually discriminating validly is that the patient is depressed to degree X.

Another kind of correlated error arises because of the fact that, if you can't make a reliable distinction, you still have to place the item on some basis or other, so you place it where you do because you think it's improbable in general in the population, or because you don't like to say nasty things about people, or on the basis of all the forces that determine the "threshold" of an item for a given rater. There are, say, two items which I cannot discriminate beyond a 7-step scale; if you require me as a judge to do an 11-step sort, how then do I decide where to put them? Well, maybe I don't like to say that people are "schizoid," but I don't mind saying they're a little "hysteroid," so that when I have to spread the items out finer than I can validly discriminate between patients, I give them a lower rating on "schizoid" and a higher rating on "hysteroid," and that means that a correlation is produced between my assessment and that of any clinician with similar rater psychology. This produces an "agreement" that is an artifact.

There is also the problem arising from the fact that when you compute a Q-correlation you lose all the information about absolute placement, because every individual's profile is first reduced to standard form, that being the nature of the Pearson correlation coefficient. Do we want to say that two people who scatter very little, speaking now interactively or normatively rather than ipsatively, but whose profile hills and valleys are of the same relative order of magnitude, are as much "alike" for clinical purposes, as two people who have a pronounced agreement on both elevation, and hill and valley pattern? According to Q-technique they appear as much alike. According to some critics of the method, they aren't actually that much alike.

There is also the unsolved problem of free vs. forced sorts. People argue about this, and empirical studies are being done in an effort to settle it. Do you tell somebody who makes the sorts how many items he must put in each position, or do you let him sort freely? This is still an open question empirically. The available evidence suggests that there isn't very much difference, but it also suggests that if you do force them, it's better to force them in an approximately rectangular distribution than in a normal distribution. If people are allowed to sort freely, they are more likely to sort closer to rectangular than they are to normal. There are three or four published studies and we have unpublished data which indicate that when you let a person sort freely he prefers something closer to rectangular than to normal. A rectangular distribution also maximizes the total number of discriminations among items by pairs, which presumably is desirable; it increases the standard deviation and in general will tend to increase the correlations, so if you like bigger correlations, you have more range to work with (and therefore you have more variation in the pairwise resemblances of persons) [if you] use a semi-forced or forced rectangular distribution. My own preference currently is to semi-force; and if you do force, force rectangularity rather than normality. But a good deal more research on this matter is needed.

There is also the problem of social desirability, and this is a terrible problem, because these ratings are heavily infected with how "sick" or "well" the rater thinks an item is. Just because it is correlated with desirability doesn't mean it's invalid, of course. I don't think we know which direction the causal arrow should be drawn there. There are some "sick" characteristics that some people do have, and they're "undesirable," and they're relatively rarer than less sick traits are. This is just a fact about these people and these items. If you eliminate all the content that's loaded with desirability and undesirability, you won't have much of any "psychiatric" content left. It would be easy to demonstrate statistically that non-psychological symptoms differ widely in their loading on "social desirability." For example, any physician would rate Cheyne-Stokes respiration or the Hippocractic facies as "worse" than nausea or a fever of 101°. Furthermore, people manifesting the former would be (globally) judged as "sicker." Thirdly—since there are in fact many more sick than moribund patients—the incidence of attribution of Cheyne-Stokes respiration or Hippocractic facies in almost any sample of patients will be low. When you screen a pool for social desirability, or modify the items so that they don't reflect this in any way, it's questionable to me whether or not you are also eliminating some intrinsic validity from the item content.

Finally, there is a serious problem with regard to the assessment of statistical significance in Q-technique. When we do a significance test between two correlation coefficients, we have a model in mind, involving the idea of sampling from a population. Now just what population are you sampling from when you do a Q-sort? In the original view it was actually a finite, but large, population. Literally, that's what Stephenson did. He took, for example, all the statements out of one of Jung's books so he had something like 1500 cards with these statements on them. And then he would actually sample, at random, 100 items from this huge item pool; so that he could employ some kind of a suitable model. But most people today don't do this. If I ask the question, "Is this Q-resemblance of Jones and Smith significantly greater than the Q-resemblance between Robinson and Fisbee?", in order to make sense of that question I need some mathematical model and a defined population from which I am sampling. What the population is here is quite hard, if not impossible, to say. One approach that has been taken is to deal with the pattern of coefficients themselves in a non-parametric manner, doing sign tests and various other kinds of counting procedures on them. E.g., we ask how many times is this kind of inter-person coefficient bigger than that kind, rather than treating the correlation coefficient itself as an estimate of any hypothetical population value.

References

1. Stephenson, William (1953) *The study of behavior*. Chicago: Univ. of Chicago Press.

2. Cattell, Raymond B. (1944) Psychological measurement: ipsative, normative, and interactive. *Psychol. Rev.*, *51*, 292-303.

3. Block, Jack (1956) A comparison of forced and unforced Q-sorting procedures. *Educ. Psychol. Measurement*, *16*, 481-493.

4. Freud, Sigmund (1948) Character and anal erotism. *Collected Papers*, II, 45-50. London: Hogarth Press.

5. Block, Jack (1955) The difference between Q and R. *Psychol. Rev.*, *62*, 356-358.

6. Allport, Gordon (1937) *Personality*. New York: Henry Holt and Co.

pdf by LJY, November 2015