

## Reliability: A Dissenting View

by Alberto Marradi

Universities of Florence and Bologna

1. The concept of reliability originates in the physical sciences, as a property of an instrument and an observer operating it, which make repeated observations of the state of an object on a property. It may be considered the inverse of the variance of all instrument readings of the same state by the operator. The lesser that variance, the more reliable the instrument-plus-operator.

This conceptual building rests on three assumptions:

- a) there is a true state of the observed object on the property of interest;
- b) that true state does not change spontaneously during the interval between the first and the last observation;
- c) that true state is not changed by the observations themselves.

A fourth assumption needs to be mentioned, although it is so general as to characterize the whole conceptual and theoretical building of the physical sciences:

- d) differences between individual objects of the same kind (e.g. atoms of the same element) are — if they exist — irrelevant for the purpose at hand (in our case, assessing the reliability of the instrument-plus-operator).

Given these assumptions, for any observation the difference between the observed and the (unknown) true state is attributed to an error (it is well known how Gauss noticed that the distribution of a sufficiently large series of observations of the position of a star took a bell-shaped form — an essential branch of inductive statistics, the theory of errors, originating from that remark).

2. In the social sciences, assumption (d) is untenable — this being the most fundamental difference with the physical sciences. Assumption (b) is hazardous to say the least: according to Irwin Deutscher, “the assumption that human thought and behaviour is static... is simply antithetical to social science” (1966b, pp. 240).

The psychological literature is full of evidence on recall, practice effects, etc., showing that assumption (c) is also untenable (see a bibliography in Anastasi, 1953, pp. 190-1).

Nevertheless, early psychometricians have adopted a concept of reliability which implied assumptions as close as possible to the physical scientists'. I am referring to the concept of 'test-retest reliability', which is operationalized through the following procedure:

- i) a given instrument (usually a multi-item test: let us call it test X) is administered at time  $t'$  in place  $p$  to a population of subjects (usually a “captive” population of young people attending courses in psychology);
- ii) answers are gathered, scored, and transformed into a vector of figures;
- iii) steps (i) and (ii) are repeated at time  $t''$ .
- iv) a correlation is computed between the two vectors obtained by steps (ii) and (iii).
- v) such correlation is christened “the reliability coefficient of test X” and is considered as a permanent and definitive attribute of that test, following it wherever, whenever and to whomever it happens to be administered.

This procedure introduces a crucial (though un-noticed) new aspect into the concept of reliability. In the physical sciences, repeated observations are made of (the state of) *a single* object on a property; data are arranged in a vector. In psychometrics, repeated observations are made of (the states of) *a plurality* of subjects on a property; data are arranged in a matrix:  $t$  vectors of  $n$  elements crossing  $n$  vectors of  $t$  elements (where  $n$  is the number of subjects and  $t$  is the number of time points).

As a consequence, whereas in the physical sciences the reliability concept is scalar in nature, in psychology test-retest reliability should be vectorial (one value per each subject). Psychologists, however, after having introduced a plurality of subjects as a concession to the epistemological peculiarity of their object (man), reap the computational advantages of this plurality while rejecting its undesirable consequences in terms of limitations in the generality of the reliability figure thus computed.

Let us give that gambit a closer look.

While physical scientists, having one vector of observations only, operationalize reliability as the variance of a distribution, psychometricians, having (at least) two vectors, may operationalize it as a correlation coefficient

between two vectors. As two vectors of figures are enough for a correlation, psychometricians may (and in fact they do) spare all observations after the second one (at time  $t''$ ), while physical scientists need many more than two in order to build an eligible distribution); moreover, though starting with a matrix, psychometricians do end up with a nice scalar concept too (one single reliability value for the whole matrix of observations). Having obtained those advantages from their recognition of the relevance of inter-subject differences, psychometricians proceed to deny that same relevance when (step v) they permanently attach to a given test a “reliability coefficient” computed on a given population in place  $p$  at times  $t'$  and  $t''$ .

Let us now list the assumptions behind a concept of reliability operationalized through the procedure described:

- a') there is a true state of each of a set of observed subjects on the property of interest;
- b') such true states do not change spontaneously during the interval between  $t'$  and  $t''$  ;
- c') such true states are not changed by the observation process taking place at time  $t'$ ;
- d') given the relevance of inter-subject differences, the reliability of an instrument cannot be ascertained if one subject only is observed. At least  $n$  subjects are needed. However, they can be taken from among the subjects at hand (captive students);
- e) notwithstanding the relevance of inter-subject differences, the reliability of an instrument is the same for all the observed subjects, and can be expressed by a single figure;
- f) notwithstanding the relevance of inter-subject differences, the reliability coefficient, once computed for a group of subjects, is extendable to any other possible group of subjects in the world, and need not be re-computed;
- g) notwithstanding the relevance of inter-subject differences, the reliability coefficient computed when the test was administered by mister Soandso in such and such conditions need not be re-computed when whoever else in whatever conditions administers the test;
- h) the reliability coefficient computed at times  $t'$  and  $t''$  need not be re-computed when the test is administered at any other times;
- i) the reliability coefficient computed in place  $p$  need not be re-computed when the test is administered in any other place in the world.

Assumptions (e) through (i), and more specifically (e) and (f), besides being hardly tenable, contradict assumption (d'). Notice that physical scientists do not fall in that contradiction because they assume (d), the irrelevance of inter-object differences. On the other hand, they take well into account inter-subject differences between observers, and therefore they abstain from attaching a single reliability figure to their instruments. Again there is no contradiction on their part, because their observers are not of the same kind as their objects, and therefore they are not considered under assumption (d).

**3.** Assessment of reliability through test-retest correlation has been criticized by some psychologists and social scientists, mainly — as we have seen — on grounds of amenability of assumptions (b') and (c'). Galtung (1959) has observed that a high test-retest reliability coefficient may be an artefact of the balancing off of spontaneous changes in true states with contrary changes induced by the process of testing. More convincingly, Converse, Rozelle and Campbell used data from many-wave panels to show evidence of diachronic changes in true states: vectors of observations of states on the same property were the less strictly correlated the farther apart in time (Converse, 1970; Rozelle and Campbell, 1969). Accordingly, Cronbach (1947) pointed to the relevance of the time interval between observations, claiming that it should be mentioned as a necessary complement to the reliability figure

However, if test-retest reliability has lost its monopoly position within psychometrics, this is probably due less to the impact of such critiques as reported above than to a desire of avoiding its “great practical disadvantage... it is often quite impossible to get permission to test a group of students twice” — as a widely used textbook candidly confesses (Ingram, 1977, pp. 17-18; also see Carmines and Zeller, 1980, p. 39)

While the test-retest procedure (and the associated concept) has never been abandoned, new procedures have been devised, still based on mathematical operations on vectors of figures.

The manuals presently available to me do not report who first had the idea permitting the development of these new procedures (manuals typically neglect the origins and developments of the crafts they expose — as Kuhn has remarked).

The idea is: given that reliability may be assessed by correlating two vectors of figures, why should the

difference between vectors necessarily be diachronic?

In other words, instead of correlating scores taken by the same  $n$  subjects on the same test  $T$  at times  $t'$  and  $t''$ , let us correlate scores taken by the same subjects on two very similar tests,  $T_a$  and  $T_b$ .

The two similar tests can be obtained by tearing a previous test in two parts ("split half") either by random assignment of items or by assigning odd-numbered items to  $T_a$  and even-numbered to  $T_b$ ; alternatively, two shorter tests are built so as to be maximally similar ("parallel forms").

At any rate, besides avoiding the practical disadvantages associated with a concept of reliability based on diachronic stability, the new concept of reliability based on synchronic consistency allowed psychometricians to drop as no more needed the two most criticized assumptions: (b') invariance of true states and (c') un-obtrusivity of the observation process.

As a matter of fact, the technique of administering parallel forms at *two* different points in time has been also frequently employed ("alternative forms": Carmines and Zeller 1980). This seems a way of combining the problems associated with the diachronic and with the synchronic assessment of reliability, with the exception of memory effects, that are avoided by a different content of the items.

However, even though the two most criticized assumptions could be dropped, all other, no less criticisable though in fact not criticized, assumptions remained. [Cronbach (1947) in fact challenged part of the assumption (g) by warning that reliability coefficients are generalizable only to similar testing situations]. An instrument continued to be judged (sufficiently) reliable for all times and places if in a particular administration to a bunch of sophomores a sufficient statistical agreement had been found between two distributions of scores.

The main preoccupation of psychometricians was with the lowering of reliability figures due to the reduced length of split halves or parallel forms vis-a-vis the good old un-split test to be administered twice. But this shortcoming was soon put right by the so-called "prophecy formula" (a mathematical manipulation of coefficients aimed at compensating for reduced test length) independently arrived at by Spearman (1910) and by Brown (1910) in the same issue of the "British Journal of Psychology".

The next step on the road to mere mathematical manipulation of figures came in 1937 when two members of Thurstone's psychometric laboratory proposed a formula to compute reliability from a correlation matrix of scores taken *on each single item* of a test rather than from the correlation of two vectors of scores each taken on *an entire* test (Kuder and Richardson, 1937).

This step was very serious because it entailed the loss of the only justification for computing reliability in the previous way, viz. the fact that scores on an entire test are an inherently better basis for that computation than scores on individual items.

It must be added that Kuder and Richardson's formula assumed dichotomous items, i.e. the format more sensitive to errors (in a dichotomy, every error turns black into white and white into black) and to distortions in the correlation coefficients due to skewed distributions.

Later, the formula has been generalized to multi-category items (Cronbach, 1951); even more general formulas have been proposed (Tryon, 1957); other coefficients based on factor analysis have been developed (Heise and Bohrnstedt, 1970; Armor, 1974).

**4.** While some of these coefficients are technically interesting, we will not deal with them here, because they introduce no relevant change in the concept of reliability as it has been conceived by psychometricians and, by derivation, by most empirically oriented social scientists, i.e. — we repeat — the property of an instrument, to be computed once and for ever by mathematical manipulation of vectors of figures in a data matrix, without any consideration for the degree of correspondence between such figures and the actual states of individuals on whatever property is thus "reliably measured".

In the next section we will argue that this concept is the parody of an answer to the actual problem of reliability; it is just one of the many comfortable devices helping the social scientists to revolve in a vicious circle of technicalities while minimizing contacts with his object, the social world.

In this section we point to another negative consequence of what we call for brevity's sake the psychometric concept of reliability. In order to do this we need a brief parenthesis on the development of the concept of validity.

Unlike reliability, this concept originated within psychology, in connection with problems typical of a science dealing with a symbolic universe: will this question/statement be interpreted by subjects in the same way as by the researcher? will the answer actually be given in function of the trait associated with the question in the researcher's mind? etc.

Accordingly, the first meaning of the concept recognized that validity was a judgement by the researcher on the likely semantic content of an item for the population investigated. However, this concept could not satisfy a

discipline searching to enhance (recognition of) its scientific status through expulsion or concealment of “subjective” interventions. Therefore “content” validity (often verbally down-graded to “face” validity) came to be considered as a poor, unscientific alternative to “criterion” validity, which could be measured by “objective” means. Such means were the correlation of a vector of scores on the indicator to be validated with another vector of scores on a variable being assumed as valid by definition, and therefore able to act as a “criterion” for the validity of the other.

As it has recently been remarked (Turner, 1979) the concept of “criterion” validity just opened a potential infinite regress, in that in its turn the “criterion” needed being validated through a judgment on its semantic content; the elimination of “subjectivity” was just a self-delusion. However, the new concept of validity was more comfortable in that it allowed psychometricians to reduce the problem to the familiar dimension of a mathematical manipulation of vectors of figures.

By reducing the privileged role of the “criterion”, the concept of concurrent validity was formed, where the degree of agreement in the distribution of scores on several independent indicators of the same concepts was emphasized.

So, at the end of our parenthesis we find that the concept of validity has been rapidly homogenized to meet the requirements of “science”: i.e., to a form liable to “measurement” through mathematical manipulation of figures, with no apparent “subjective” intervention and no need to bother about relations of meaning with whatever is outside the matrix.

As a consequence of this homogenisation, the conceptual distinction between reliability and validity becomes blurred, as has been lamented by symbolic interactionists (Blumer, 1954; Deutscher, 1966a) and viewed as a step in the right direction by some psychometricians (Lumsden, 1976; Davies, 1977).

Mathematical formulas connecting the two concepts have been devised (see e.g. Cronbach, 1949/1970, p. 171; Carmines and Zeller, 1980, p.34). Campbell and Fiske have correctly perceived that, within the framework of a matrix approach, reliability and validity are just the two extremes of a continuum: “Both reliability and validity concepts require that agreement between measures be demonstrated. A common denominator which most validity concepts share in contradistinction to reliability is that this agreement represents the convergence of independent approaches. Independence is, of course, a matter of degrees and, in this sense, reliability and validity can be seen as regions on a continuum. Reliability is the agreement between two efforts to measure the same trait through maximally similar methods. Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods. A split-half reliability is a little more like a validity coefficient than is a test-retest reliability, for the items are not quite identical. A correlation between dissimilar subtests is probably a reliability measure, but is still closer to the region called validity” (1959, p. 83).

Campbell and Fiske’s statement has been quoted at length because it shows as clearly as one could desire the consequences of reducing both concepts to a degree of mathematical agreement between distributions. Further steps along this path have been taken by Siegel and Hodge (1968) who advocate “models for simultaneously handling both questions of reliability... and validity” and Blalock (1968), who conceives of a single system of equations simultaneously assessing validity, reliability, and causal relationships between variables.

**5.** Another consequence of reducing reliability to a figure is that, by an obvious psychological mechanism, maximizing that figure becomes an end in itself. In laboratory psychology, “objective” tests are preferred to “essay” tests because “the scores obtained from objective tests tend to be more reliable <i.e. to produce higher reliability coefficients> than those obtained from essay tests” (Ebel, 1968).

In social psychology and other social sciences, reliability figures are maximized through standardization of questions (see Hyman et al., 1954, p. 30); the number of answer categories is reduced, so that respondents are forced to choose the least remote item rather than disperse into detailed categories reflecting the nuances of their actual attitude or point of view (see criticism by Galtung, 1967).

Since, due to the structural mechanism of the most used formulas, the more items in the test, the higher tends to be the reliability figure, test length is often increased beyond reasonableness. As we have remarked, “the subjects for these tests are usually drawn from 'captive' populations — schoolchildren, students, job applicants, members of the armed forces, mental patients and so on — who can be subjected to test batteries of the enormous length necessary to reach the required standards of reliability” (McKinnell, 1970, p. 236).

The above described phenomenon may be taken as paradigmatic: in order to pump up the fatal figure, the well known effects of test fatigue on the *actual* reliability of answers (as distinguished from the reliability as measured by questionable formulas) are light-heartedly neglected, as well as the ethical right of each subject to

be exposed to instruments devised in order to correctly assess his own abilities, attitudes, opinions rather than to instruments devised in order to pump up a global reliability figure.

However, there are other reasons why the psychometric concept of reliability is at odds with the layman's concept. "Reliability of a rating scale tells us very little about its value, since the apparent reliability may be due to bias rather than true score" (Bartlett, Quay, and Wrightsman, 1960, p. 703).

This mechanism is exemplified, with no apparent signs of irony, by a current textbook on reliability: "Let us assume that a particular yardstick does not equal 36 inches: instead, the yardstick is 40 inches long. Thus, this yardstick systematically underestimates height by 4 inches... this error of 4 inches per yard will not affect the reliability of the yardstick since it does not lead to inconsistent results on repeated measurements" (Carmines and Zeller, 1980, p. 13).

Transferring their favourite physical science examples into a social science situation, systematic bias obtains due to response sets (a test or battery eliciting response sets and nothing else from every respondent will have a perfect reliability according to any of the current formulas), to "the presence of irrelevant propositions which evoke consistent answers" (Kirkpatrick, 1936, p. 36), to items that "have tapped a narrow range of content, being slight rewordings of the same statement" (McKinnell, 1977, p. 211), and by a host of other sources of "irrelevant methods variance" (Scott, 1968, 2nd ed., vol.II, p. 257).

Through the ingenious multitrait-multimethod matrix, Campbell and Fiske (1959) have mustered compelling evidence that a large part of the correlation between items of the same "scale" is due to "irrelevant methods" factors, i.e. to the common format, similar wording, etc. of items rather than to their semantic content.

6. A few authors have perceived the gap between the psychometricians' and the layman's concepts: "Use of the term 'reliability' in its technical sense of a test of consistency of response is unfortunate. The non-technical, layman's sense of reliability is trustworthiness. Behavioural scientists may be able to quote accurately the technical definition of the term, but I fear that they are nevertheless more likely to feel confidence in a test of reliability than they ought, simply because of the halo effect of the layman's meaning of the term. It would be much better to call such tests tests of consistency, which they are, than to call them tests of reliability, which in the everyday sense of the word they most certainly are not" (Naroll, 1968, p. 265-66).

Being entrenched in the linguistic and practical habits of social scientists, the psychometric concept blocks the road to the development of a concept closer to common sense and more germane to the epistemological situation of the social sciences. "Since no other data were available on the prior history and development of the subjects, reliability had to be determined by statistical manipulation of the test data themselves. It would appear that these tests devised to meet the difficulty presented by the absence of other data now act as barriers to the use of any other procedures... the accepted methods... offer indices only for the group, not for any individual subject on that group" (Frank, 1939, pp. 399-400).

While we obviously share Frank's last remark, we don't share his opinion that psychometricians turned to their present techniques of reliability assessment due to the lack of viable alternatives. As we shall see in section 8, alternatives exist, and a minority of researchers explore them, either occasionally or systematically. However, these alternatives are much more demanding in terms of time and commitment, and — what is even worse — are not compatible with the dominant image of a scientist who, dressing white overall and aloof in his ivory tower, manipulates hypotheses and theories by means of formal logic and/or data matrices by means of computers.

The concept of reliability has been developed, in a perfectly *zweck*-rational way, so as to be part of a whole protective belt of concepts allowing the social scientist to pretend to talk about reality while minimizing his contacts with it.

Practically the only such contact left is the data-gathering stage, when individuals' states on a series of properties are transformed into data in a matrix via a set of operations controlled by operational definitions. Such operational definitions are, in most research, as standardized and mechanical as possible, so as to reduce "subjectivity" — i.e. interventions of meaningful interpretations of reality — and to be entrustable to a host of collaborators (polling agencies, interviewers, coders, word-frequency counters, test administrators, etc.).

The degree of researcher control on these operations is often minimal, again for reasons of both time-saving and "objectivity" preserving (e.g., response sets are coded at face value even when they are detected). The controlling idea is that the data matrix is an objective photography of reality and, as such, hardly needs any interventions besides the click. The innumerable interventions actually needed are therefore performed in the most impersonal, routine, unreflexive way, and removed from individual and collective conscience. They leave almost no trace in research reports as well as in epistemological writings celebrating the objectivity of science.

Once the data matrix has been filled, the next controlling idea is that the array of figures does not stand for reality thanks to a many-layered building of conventions; rather, the array of figures simply *is* reality. What the researcher has to do is just to manipulate the figures by means of pre-canned statistical techniques (often entrustable to computer operators, and at any rate presented so as to minimize and conceal “subjective” decisions and interventions), and whatever he (or rather the computer man) finds will automatically be considered as a finding about reality.

The protective belt of *ad hoc* concepts is most instrumental in supporting that mystification. Some concepts in the belt perform the function of concealing the distinction between aspects of reality and the (conventionally corresponding) features in the matrix. Most conspicuous in this role is the concept of 'variable', whose referents are simultaneously a column of figures in the matrix, a concept in the researcher's mind, and a range of individual states (which have been construed as states) on a given property. A 'case' simultaneously designates a row of figures in the matrix and a whole individual in the real world. A 'datum' stands simultaneously for a single state on a property, the un-processed bit of information about it (as in 'data gathering') and the same bit as processed through all the conventions of an operational definition (as in 'data matrix'). And so on.

Other concepts in the protective belt perform the function of construing and legitimising as mere mathematical manipulations of vectors in the matrix some intellectual operations that do — or should — involve “subjective” judgments by the researcher both as a holder of tacit knowledge about reality, subjects, research situations, etc., and as an experienced interpreter of the results of statistical procedures. In the previous sections we have seen how the concept of reliability performs this function; in section 4 we have seen the same for the concept of validity. Another prominent case in point is the concept of explanation, which has been reduced to the degree of agreement between the distribution(s) of scores on a (set of) “independent variable(s)” and the distribution of scores on a “dependent variable”.

While we think it is self-evident how this protective belt of concepts and associated procedures, conventions, etc. acts as to minimize the social scientists' involvement in his object and his need to invest time, commitment, decisional skills in the research process, perhaps a few more comments are in order concerning what may be considered the central function of all these institutions: viz., preserving the self-image of science as an objective enterprise.

True, this result is obtained by standardizing data-gathering and data-analysis procedures, by concealing through terminological tricks whatever interventions are left, etc. However, a subtler and more crucial function is performed by re-interpreting all the cognitive operations as operations on and inside the data matrix; by this reduction and reinterpretation, a situation of *gnoseological monism* is established.

Monism is thegnoseological precondition of claims to objectivity, because whenever a plurality of sources of truth are admitted, a conflict between such sources may set in, with the consequent need to make a decision between contrasting truth claims — a situation obviously incompatible with claims to objectivity. In our case, monism would be hampered by admitting that the social scientist has other sources of information about reality than the data matrix, and that his tacit knowledge and research experience are also necessary in order to decide which manipulations to perform on the data matrix, how to perform them, and how to interpret the results.

**7. How should one proceed in forming an alternative concept of reliability, free from the heavy burden of functions seen above?**

This task leads us to reconsidering the assumptions listed in section 2. We have seen that assumptions (b') and (c') were needed by test-retest reliability only and have been dropped by psychometricians using other concepts. We would reformulate assumption (a') as follows:

a'') there may be at any given moment a true state of any subject on a property of interest. We can never be certain to know what that true state is. However, the more carefully we investigate, the more likely are we to get closer to that true state, or to ascertain its non-existence. On the contrary, the more cursory, standardized, un-reflexive is the investigation, the more likely are we to record a datum significantly distorting such a true state, or corresponding to no state at all.

As a consequence, we would reformulate assumption (d') as follows:

d'') since each state (and the associated datum in the matrix) refers to an individual only, reliability (being a judgment as to the degree of correspondence between the state and the corresponding datum, given the coding conventions established in the operational definition) refers to an individual only.

We will soon argue in favour of assumption (d''). Let us now state its consequences:

— reliability is the property of one datum (more precisely, of the process whereby one state is recorded into one datum), not of one vector or one matrix of data. The degree of reliability of an individual datum cannot be

attributed to any other individual data, even though they are recorded under the same operational definition, and/or by the same operator, and/or within the same piece of research. Assumption (e) is untenable.

—as a further consequence and a fortiori, the degree of reliability of a datum cannot be generalized to data regarding individuals or populations in different times, places, situations, etc. Assumptions (f) through (i) are untenable.

Each time an individual state on a property is transformed into a datum in the matrix, the outcome of the transformation process is influenced by a *unique* set of factors associated with the subject, the property and its operational definition, the data gatherer, the situation, etc. A psychologist of behaviourist sympathies, R. L. Thorndike, has published (1949) a rather comprehensive classification of factors associated with the subject and the property, including the subject's ability to comprehend instructions, his own definition of his tasks as testee/interviewee, his self-confidence, training, attention, fatigue, tendency to lie, to invent, to guess, fortune in guessing, etc. Converse (1964) and Cicourel (1964) have insisted on the different linguistic skills of subjects in different subcultures and on the relevance of the saliency of the topic for their life-worlds.

Harré (1981) and Pawson (1982) remind that any answer must be considered an act of presentation of the self, taking into account both “the subject's perception of the appropriate mode of taking part in a research exercise” and the cues uttered by the data gatherer as to the researchers' purposes and “the appropriate form of response” (Pawson, 1982, p. 45).

Many (Cook and Selltitz, 1944; Deutscher, 1966b; Gostkowski, 1974; and other methodologists of the Lodz school) stress the distinction between private opinions and public statements, and hence the relevance of interviewer-interviewee rapport and of the (physical and human) environment around the interaction. More generally, ethnographers insist (see e.g. the essays edited by Tyler, 1969) that all speech events are highly dependent on the situational context.

Some of the remarks reported above are more appropriate of a test situation, others of an interview situation; many are relevant to both. No dearth of concurring opinions and probably quite a few other suggestions concerning influences on the transformation process (e.g.: errors and deliberate distortions by the data gatherer, the coder, etc.) can be found in the literature.

However, there is enough to conclude that, given a vector of  $n$  data recording the states of  $n$  individuals on a property, any datum can be anything from perfectly reliable to completely unreliable, and any datum is more or less reliable for reasons largely independent from any other's.

This said, one could wonder how a vectorial concept of reliability has managed to survive almost unchallenged among social scientists working with data matrices: I can only quote two remarks by Frank (see section 6 above) and Cicourel: “reliability cannot be achieved by the same procedures for all subjects, but only for each subject taken separately” (1964, p. 80).

#### **8. Two questions are legitimate at this point:**

Is such an idiographic concept of reliability useful, why and what for?

Can one assess reliability, and how?

Whatever the answer to the second question, our answer to the first question is an unconditional yes. A concept of reliability is needed to remind us that our information-gathering activities are not inherently vested with the ability to correctly record whatever aspects of reality we are interested in. Correct recording is just an ideal target that we must try to approximate.

An idiographic concept of reliability must supplant the current concept because it reminds the social science community that a reliable recording must be toiled at afresh in any situation, for any individual and for any of his states. To proceed as if that problem could be solved once and for ever by using such and such pre-canned test, or at any rate in any global or collective way, is plain self-deception.

The answer to the second question depends on what do we mean with 'assess'. If we mean 'to attach a figure', the answer should be negative excepting very special conditions and conventions. If we mean 'to get information largely sufficient to decide which data to throw away and which to consider with suspicion, and highly useful to improve our next data-gathering', the answer is positive — provided the researcher is willing to invest the time and efforts necessary.

It must be added that a positive answer is not conjectural, because idiographic information on reliability is actually being gathered by a number of research teams and centres.

Of course the problems to be confronted vary with the type of property and the type of menacing factor.

Statements about one's age, birthplace, education, family status, position on job, income, etc., may be checked against public records, and in fact they have been and are being so checked by many researches.

Several national census bureaus and polling agencies effect random checks of the reports by their

enumerators/interviewers, including re-interviews where needed.

Cannell proposed in the 1970's a "verbal interaction coding" scheme, designed to classify some easily detectable and codable features of the recorded verbal interaction during an interview (see Morton-Williams, 1979).

Others use the interviewers as active gatherers of information on reliability. Schuman (1966) proposed to assign to each interviewer the task of asking short probing questions about a respondent's understanding of a randomly chosen set of questions. Belson (1981) is devoting "intensive, probing, and challenging" re-interviews to the assessment of respondent's degree of understanding, cultural and psychological difficulties in answering, level of cooperation, etc.

Zygmunt Gostkowski, Jan Lutynski, and Krystyna Lutynska have oriented the main activity of a remarkable group of methodologists in Lodz into a full-fledged research programme in the idiographic assessment of reliability, including various forms of checks on the operation of hired interviewers, and "repeated verification interviews" aimed at reconstructing both the cognitive and the psychological processes concurring in the formation of each answer (Most publications of this group are in Polish. However, see Gostkowski, 1974; Gostkowski (ed.), 1978).

## BIBLIOGRAPHY

- ANASTASI, Anne (1953) #Differential Psychology#. London: Macmillan.
- ARMOR, David J. (1974) #Theta Reliability and Factor Scaling#, pp. 17-50 in Herbert L. Costner (ed.), #Sociological Methodology 1973-1974.# San Francisco: Jossey-Bass.
- BARTLETT, Claude J. (1960) #A Comparison of Two Methods of Attitude Measurement: Likert Type and Forced Choice#, QUAY, Lorence Childs, WRIGHTSMAN, Lawrence S. in "Educational and Psychological Measurement" XX, 4 (winter): 699-704.
- BELSON, William A. (1981) #The Design and Understanding of Questions in the Survey Interview#. London: Gower.
- BLALOCK, Hubert M. (1968) #The Measurement Problem: A Gap Between the Languages of Theory and Research#, pp. 5-27 in Hubert M. Blalock and Ann B. Blalock (eds.), #Methodology in Social Research.# New York: McGraw-Hill.
- BLUMER, Herbert (1954) #What Is Wrong With Social Theory?#, in "American Sociological Review" XIX, 1 (february): 3-10
- BROWN, William (1910) #Some Experimental Results in the Correlation of Mental Abilities#, in "British Journal of Psychology" III (october): 296-322.
- CAMPBELL, Donald T. (1959) #Convergent and Discrimination Validation by the# FISKE, Donald W. #Multitrait Multimethod Matrix#, in "Psychological Bulletin" LVI, 2 (march): 81-105
- CARMINES, Edward G. (1980) #Reliability and Validity Assessment#. London: Sage.
- ZELLER, Richard A.
- CICOUREL, Aaron Victor (1964) #Method and Measurement in Sociology#. New York: Free Press.
- CONVERSE, Philip E. (1964) #The Nature of Belief Systems in Mass Publics#, pp. 202-61 in David E. Apter (ed.) #Ideology and Discontent#. Glencoe: Free Press.
- CONVERSE, Philip E. (1970) #Attitudes and Non Attitudes: Continuation of a Dialogue#, pp. 168-89 in Edward R. Tuftte (ed.) #The Quantitative Analysis of Social Problems#. Reading: Addison-Wesley

- COOK, Stuart W.  
SELLTIZ, Claire (1964) #A Multiple Indicator Approach to Attitude Measurement#, in "Psychological Bulletin" LXII, 4 (july): 36-55.
- CRONBACH, Lee J. (1947) #Test Reliability: Its Meaning and Determination#, in "Psychometrika" XII, 1 (march): 1-16.
- CRONBACH, Lee J. (1949) #Essentials of Psychological Testing#. New York: Harper & Row. Quotations from the 1970 edition.
- CRONBACH, Lee J. (1951) #Coefficient Alpha and the Internal Structure of Tests#, in "Psychometrika" XVI: 297-334.
- DAVIES, Alan (1977) #The Construction of Language Tests#, 38-104 in J. P. B. Allen and Alan Davis (eds.), #Testing and Experimental Methods#. London: Oxford University Press.
- DEUTSCHER, Irwin (1966a) #Looking Backward: Case Studies in the Progress of Methodology in Sociological Research#, in "American Sociologist" IV, 1: 34-42.
- DEUTSCHER, Irwin (1966b) #Words and Deeds: Social Science and Social Policy#, in "Social Problems" XIII: 233-54.
- EBEL, Robert I. (1968) #Achievement Testing#, in #International Encyclopedia of the Social Sciences# I: 33-39.
- FRANK, Lawrence R. (1939) #Projective Methods for the Study of Personality#, in "Journal of Psychology" VIII, 2 (october): 389-413.
- GALTUNG, Johan (1959) #An Inquiry into the Concepts of 'Reliability', 'Intersubjectivity' and 'Constancy'#, "Inquiry" II, 2 (summer): 107-25.
- GALTUNG, Johan (1967) #Theory and Methods of Social Research#. London: Allen & Unwin.
- GOSTKOWSKI, Zygmunt (1974) #Toward Empirical Humanization of Mass Surveys#, in "Quality and Quantity" VIII, 1 (march): 11-26.
- GOSTKOWSKI, Zygmunt (ed.) (1978) #Investigations on Survey Methodology#. Warszawa: PAN.
- HARRE', Rom (1981) #Philosophical Aspects of the Macro-Micro Problem#, pp. 139-60 in Karin D. Knorr-Cetina and Aron Victor Cicourel (eds.), #Advances in Social Theory and Methodology. Toward an Integration of Micro- and Macro-sociologies#. London: Routledge.
- HEISE, David R.  
BOHRNSTEDT, George W. (1970) #Validity, Invalidity, and Reliability#, pp. 104-129 in Edgar F. Borgatta and George W. Bohrnstedt (eds.), #Sociological Methodology 1970#. S. Francisco: Jossey-Brass.
- HYMAN, Herbert H.  
#et al.# (1954) #Interviewing in Social Research#. Chicago University Press.
- INGRAM, Elisabeth (1977) #Basic Concepts in Testing#, pp. 11-37 in J. P. B. Allen and Alan Davis (eds.), #Testing and Experimental Methods#. London: Oxford University Press.
- KIRKPATRIK, Clifford (1936) #Assumptions and Methods in Attitude Measurements#, in "American

Sociological Review" I, 1 (february): 75-88.

- KUDER, G. Frederick  
RICHARDSON, Marion W. (1937) #The Theory of the Estimation of Test Reliability#, in "Psychometrika" II, 3 (september): 151-60.
- LUMSDEN, J. (1976) #Test Theory#, in "Annual Review of Psychology" XXVII: 251-80.
- McKENNELL, Aubrey C. (1970) #Attitude Measurement: Use of Coefficient Alpha with Cluster or Factor Analysis#, in "Sociology" IV, 2 (may): 227-45.
- McKENNELL, Aubrey C. (1977) #Attitude Scale Construction#, pp. 183-219 in Colm O' Muirheartaigh and Clive Payne (eds.), #The Analysis of Survey Data#. New York: Wiley, vol. I.
- MORTON WILLIAMS, Jean (1979) #The Use of "Verbal Interaction Coding" for Evaluating a Questionnaire#, in "Quality & Quantity" III, 1 (february): 59-75.
- NAROLL, Raoul (1968) #Some Thoughts on Comparative Method in Cultural Anthropology#, pp. 236-277 in Hubert M. Blalock and Ann B. Blalock (eds.), #Methodology in Social Research.# New York: McGraw-Hill.
- PAWSON, Ray (1982) #Desperate Measures#, in "British Journal of Sociology" XXXIII, 1 (march): 35-63.
- ROZELLE, Richard  
CAMPBELL, Donald T. (1969) #More Plausible Rival Hypotheses in the Cross-Lagged Panel Correlation Technique#, in "Psychological Bulletin" LXXI (january): 74-80.
- SCHUMAN, Howard (1966) #The Random Probe. A Technique for Evaluating the Validity of Closed Questions#, in "American Sociological Review" XXV, 1 (february): 3-25.
- SCOTT, William A. (1968) #Attitude Measurement#, pp. 204-273 in Gardner Lindzey and Elliot Aronson (eds.), #Handbook of Social Psychology#, Vol. II. Reading: Addison-Wesley, 2nd edition.
- SIEGEL, Paul M.  
HODGE, Robert W. (1968) #A Causal Approach to the Study of Measurement , Error, pp. 28-59 in Hubert M. Blalock and Ann B. Blalock (eds.), #Methodology in Social Research.# New York: McGraw-Hill.
- SPEARMAN, Charles (1910) #Correlation Calculated from Faulty Data#, in "British Journal of Psychology" III, (october): 271-295.
- THORNDIKE, Robert L. (1949) #Personnel Selection. Test Measurement Techniques#. New York: Wiley.
- TRYON, Robert Choate (1957) #Reliability and Behavior Domain Validity: Reformulation and Historical Critique#, in "Psychological Bulletin" LIV, 3 (may): 229-249.
- TURNER, Stephen P. (1979) #The Concept of Face Validity#, in "Quality & Quantity" XIII, 1 (february): 85-90.
- TYLER, Stephen A. (ed.) (1969) #Cognitive Anthropology#. New York: Holt.