

## **La validité des indicateurs et la fidélité des données**

*par Alberto Marradi*  
Université de Florence

Les concepts de validité et de fidélité sont très souvent rapprochés et voire confondus. On présente ici la thèse d'une profonde différence entre les deux concepts.

Parmi les diverses interprétations de la nature de la validité, on va soutenir la vision que la validité est une propriété du lien sémantique entre un concept et son ou ses indicateurs. Un tel lien n'existe pas "dans les faits"; il est établi par un chercheur ou par une équipe de chercheurs. C'est aussi la communauté des chercheurs qui juge de la validité, c. à d. du fait que le lien est plus ou moins sémantiquement approprié. Ce jugement peut être illuminé, mais jamais remplacé, par le niveau des "coefficients de validité" qui mesurent la force des corrélations statistiques entre vecteurs de chiffres référés à des indicateurs du même concept ou de concepts différents.

Eux aussi, les "coefficients de fiabilité" mesurent la corrélation statistique entre vecteurs — d'où la présente confusion entre les concepts de fiabilité et de validité. Mais ils sont exposés à des critiques plus radicales que les coefficients de validité.

Si l'on définit la fidélité comme le degré de correspondance entre les états sur une propriété et les chiffres qui les représentent, étant donnée une certaine définition opérationnelle, sur un vecteur, deux conséquences s'ensuivent:

a) la fidélité ne peut point être mesurée, puisque les états sur une propriété ne peuvent être connus directement, mais seulement à travers des définitions opérationnelles dont aucune ne peut certifier la fidélité de l'autre. La fidélité peut seulement être appréciée à travers un jugement plus ou moins illuminé par des comparaisons entre les informations recueillies au moyen de différentes définitions opérationnelles de la même propriété.

b) Etant données les différences entre individus et entre situations d'observation, dans les sciences sociales la fidélité ne peut être l'attribut d'un instrument, ou d'une variable, ou de n'importe quoi de relatif à plusieurs individus et/ou plusieurs situations. Elle est l'attribut d'une donnée, c. à d. du code qui a été affecté à l'état du sujet X après l'avoir observé dans une certaine situation. Le même instrument peut enregistrer avec des degrés radicalement différents de fidélité les états d'individus différents, ou du même individu dans des situations différentes.

Bien sûr on peut évaluer, avant ou après la collecte des données, et sur la base de toute sorte d'information, la probabilité qu'une certaine définition opérationnelle produise ou ait produit une proportion jugée suffisante d'enregistrements suffisamment fidèles. Mais un concept si vague devrait être séparé du concept de fidélité, et identifié par un terme différent — fiabilité — peut être en distinguant fiabilité à priori (avant la collecte) et à posteriori (après une évaluation des résultats de la collecte).

## 1. *Les indicateurs*

Le terme ‘indicateur’ dans son acception actuelle est déjà présent au siècle dix-neuvième dans la technologie (indicateur de niveau, de pression) et dans plusieurs sciences (chimie, économie). A ma connaissance, soit Quételet (1869) soit Villermé (1840) emploient fréquemment le concept, mais pas le terme. Durkheim parle (1893; 1896) d’“indices extérieurs” qui “symbolisent” des “faits intérieurs” et des “phénomènes moraux”. Le positiviste italien Niceforo écrit un ouvrage (1921) sur les “indices numériques de la civilisation et du progrès”. Nous n’avons trouvé aucun sociologue qui utilise le terme ‘indicator’ avant Dodd (1942a). Mais c’est Lazarsfeld (1958) qui codifie l’usage aujourd’hui prédominant parmi les sociologues, en établissant la distinction entre ‘indicator’ (simple) et ‘index’ (combinaison d’indicateurs).

Sur la raison du recours aux indicateurs l’accord parmi les méthodologues est remarquable: les concepts théoriquement intéressants, et/ou de haut niveau de généralité, ne sont point (ou pas aisément) observables; il faut se servir de leurs indicateurs qui le sont (Merton 1948, 514; Lazarsfeld et Barton 1951, 181; Galtung 1967, 122 et 292; Blalock 1968, 6; Blalock 1969a, 151; Abell 1969, 398; Przeworski et Teune 1970, 102; McKennell 1973, 218; Nowak 1976, 52; Carmines et Zeller 1979, 10).

**1.1.** L’accent sur l’observabilité semble une marque sans conséquence de l’héritage épistémologique des sciences naturelles. Il entraîne pourtant une certaine confusion entre indicateurs et définitions opérationnelles, témoignée par des propositions comme: “Theoretical concepts are considered as more abstract than operational definitions, and several operational definitions may be indicators of a single theoretical concept” (Reynolds 1971, 50), ou encore comme: “The concepts that go into the model have to be converted into operational indicators, so called because they make explicit the operations or procedures by which the phenomena of the referent world are expressed in sets of scientifically useful data” (Singer 1982, 182), ou comme: “Multiple operational definitions of a given concept are not always clearly distinguishable from indicators of a multiplicity of distinct concepts” (Macrae 1970, xii).

Une telle confusion ne serait pas possible si l’on voyait clairement que l’indicateur est un concept tandis que la définition opérationnelle est une série de règles, explicites et implicites, qui permettent de transformer l’état d’un cas sur une propriété en une donnée. Identifier indicateurs et définitions opérationnelles est donc une “erreur catégorielle” (Ryle 1938). Si l’on comprend cela, on verra mieux la nature du lien entre les deux, c. à d. le fait que le rôle des indicateurs dans la recherche est une conséquence du rôle des définitions opérationnelles.

Beaucoup d’épistémologues et de méthodologues ont défini ce dernier rôle comme essentiel pour la science empirique (Rapoport 1958, 978; Campbell et Fiske 1959, 101; Nowak 1976, 296) ou pour le contrôle des propositions (Chapin 1939, 155; Dodd 1942b, 484; Hempel 1952, 54; Frank 1961, ix). Il serait plus prudent de le

définir essentiel pour tel contrôle s'il est exécuté moyennant des opérations logico-mathématiques (y compris toutes les opérations de la statistique) sur un vecteur ou une matrice de données. En établissant les règles pour transformer l'état de chaque cas sur une propriété en une donnée, la définition opérationnelle permet de transformer telle propriété en une variable, c. à d. en un vecteur de chiffres interprétés à l'aide d'un code.

On n'aurait pas besoin d'indicateurs si pour chaque propriété que l'on peut vouloir transformer en variable l'on disposait d'une définition opérationnelle qui tienne compte de tous les aspects de la propriété jugés importants. Lorsqu'on n'en dispose pas, et quelque aspect important est négligé, automatiquement la connotation de la propriété est modifiée, et il ne s'agit plus de telle propriété mais d'une autre.

Par exemple, si la définition opérationnelle de la propriété 'liberté politique' d'un Etat prévoyait une série d'opérations pour évaluer les conséquences négatives subies par les citoyens à cause de leurs opinions politiques, ce serait seulement la propriété 'étendue des restrictions légales à la liberté d'opinion' à être définie. Cette propriété-ci est sans doute un aspect important de la liberté politique, mais il y a d'autres aspects qui sont négligés. On aurait donc défini une propriété à un plus bas niveau de généralité.

Il est aussi des cas où la définition opérationnelle ne semble pas réduire la généralité d'une propriété, mais nos connaissances psychologiques et sociologiques autorisent à croire que tout autre propriété soit en effet en jeu. Un exemple serait une définition opérationnelle de 'degré d'autoritarisme' prescrivant que les états individuels soient représentés par les choix entre les réponses "beaucoup/assez/un peu/pas de tout" à la question "Etes-vous autoritaire?". On conviendra que la disposition à s'avouer publiquement autoritaire ne compte pas parmi les aspects caractérisant la syndrôme autoritaire. On fera donc recours à une définition opérationnelle moins directe, par exemple en demandant aux sujets leurs réactions à un propos tel que: "les subordonnés ont une vision des problèmes plus réaliste que leurs supérieurs".

Bien sûr, ladite réaction n'est qu'un aspect de la syndrôme autoritaire, tout comme l'étendue des restrictions légales à la liberté d'opinion est un aspect de la liberté politique. En les utilisant pour suggérer une définition opérationnelle des propriétés plus générales, on les choisit automatiquement, et même si l'on n'en a pas conscience, comme indicateurs desdites propriétés. "The use of indicators is called for whenever the researcher has definite theoretical concepts... for which he is unable to obtain or defend simple, unambiguous, direct operational definitions" (Curtis et Jackson 1968, 195).

On voit donc qu'il n'est pas parfaitement exact de lier — comme le fait la majorité des méthodologues — la nécessité des indicateurs à l'intérêt théorique et/ou au niveau de généralité des concepts de propriété. Comme Blalock a remarqué (1961, 163), "certain theoretically defined variables can be readily associated with operations for measurement purposes... age, sex, race, and religious membership". On ne voudra pas disputer l'intérêt théorique ou le niveau de généralité des propriétés ci-dessus, ainsi que de la région de naissance, du degré scolaire, du statut professionnel et, en ce qui concerne les Etats, du territoire, population, forme constitutionnelle, etc.

Toutefois, elles n'ont pas besoin d'indicateurs, car il n'y a pas de difficulté à établir une définition opérationnelle qui, d'après le jugement des compétents, ne néglige pas d'aspects importants.

Le fait que telles difficultés existent est donc la condition nécessaire et suffisante du recours à un autre concept, qui à son tour pourra être choisi comme indicateur seulement si tous les aspects jugés importants de sa connotation sont couverts par sa définition opérationnelle.

**1.2.** Cela étant dit, on peut considérer la question de la *nature* du lien entre un concept et son ou ses indicateurs.

La majorité des méthodologues y voient une relation de cause à effet (Stevens 1951, 47; Blalock 1961, 145; Jacobson et Lulu 1974, 215; Sullivan 1974, 250; Smelser 1976, 164 et 194), même si quelques-uns soulignent que la relation est probabiliste plutôt que déterministe (Lazarsfeld 1958, 103; 1961, 307; et 1966, 162; Verba 1969, 64; Nowak 1976, 48). De façon cohérente avec cette conception-là, Siegel et Hodge (1968), Costner (1969), Blalock (1969b) et son école (Blalock 1974) ont proposé d'estimer la force du lien entre concept et indicateur (nommé "corrélation épistémique" par Costner) par des modèles d'équations structurelles dérivés de la *path analysis*.

D'autres méthodologues conçoivent la relation selon les modèles familiers de la statistique inductive, en parlant d'inférence (Galtung 1967, 294; McKennell 1973, 218; Singer 1982, 205), d'hypothèse (Teune 1968, 128), d'échantillonnage d'un "univers de contenu" (Guttman 1950; Cronbach et Meehl 1955, 282). Une troisième possibilité est parfois envisagée, c. à d. une relation logique (Nowak 1976, 48), analytique (Sullivan et Feldman 1979).

**1.2.1.** A notre avis, l'interprétation analytique n'est pas soutenable, car, entre les dénnotations d'un concept et de son indicateur, il n'existe jamais une relation d'inclusion complète comme celle qui existe entre le tout et la partie, le genre et l'espèce. On ne dispute pas que tous les chiens gris sont chiens, mais l'on peut toujours trouver quelqu'un qui dispute que le concept I est un bon indicateur du concept C.

L'analogie avec l'échantillonnage est mal choisie, car personne ne saurait définir un "univers de contenu" (Campbell et Kerckhoff 1968; Carmines et Zeller 1979, 21), ni des procédés d'extraction (McKennell 1970, 242). Le terme 'inférence' n'est pas approprié non plus, car il se réfère aux liens entre propositions, et non pas aux liens entre concepts.

On peut penser à une relation causale entre un concept et son indicateur (pour reprendre l'un de nos exemples, on peut juger que Mr. X nie que les subordonnés soient plus réalistes des supérieurs puisqu'il est autoritaire), ce qui justifie une relation causale à direction renversée dans l'esprit de l'observateur (on juge que Mr. X est autoritaire puisqu'il nie...).

Mais déjà dans l'autre exemple cette interprétation semble plus artificielle (la liberté politique *cause* le manque d'entraves à la liberté d'opinion?), et parfois elle semble terriblement artificielle; par exemple si la population du chef-lieu est choisie

comme indicateur de “centralité” du département (Cartocci 1985), ou si le niveau de pollution de l’atmosphère mesuré dans telle place est choisi comme indicateur de qualité de la vie dans le canton.

Nous pensons que l’interprétation causale trahit l’intention, due à un résidu d’objectivisme, d’exorciser la nature stipulative du choix des indicateurs. Le rapport d’indication n’existe pas en nature; il est établi par un chercheur en raison de l’idée qu’il se fait des étendues sémantiques du concept C et de l’indicateur I et de leur partie commune. Au moment du choix, il ne lui arrive pas de se demander si cette partie commune est due à des processus causales ou d’autre nature; plutôt il se demande si la partie commune a une étendue suffisante pour considérer I comme un indicateur valide de C (sur le concept de validité, voire la section 2). Il ne se demande pas si la liberté politique cause ou ne cause pas le manque de restrictions juridiques à la liberté d’opinion; il se demande quels rapports existent entre les connotations des deux concepts (le manque d’entraves s’accorde bien avec les autres aspects de la liberté politique?) et entre leurs dénnotations (combien de pays sont jugés libres malgré les entraves...?).

Le rapport d’indication est donc un rapport de représentation sémantique que le chercheur stipule entre deux concepts.

**1.3.** Par définition (voir § 1.1), l’étendue sémantique commune entre un concept et un indicateur n’épuise pas la connotation du concept — qui dans ce cas-là n’aurait pas besoin d’indicateurs. La représentation sémantique est par sa nature partielle.

Il est donc convenable de choisir plusieurs indicateurs du même concept, comme le recommande la majorité des méthodologues (des avis différents sont exprimés par Abell 1969; Pawson 1980; Saris 1981, auxquels on renvoie). La raison généralement avancée est que plusieurs indicateurs garantissent une meilleure représentation sémantique de la connotation d’un concept (Cronbach 1949, 620; Lazarsfeld 1958, 110; Curtis et Jackson 1968; Cook et Selltiz 1964, 37; Etzioni et Lehmann 1967, 4; Scott 1968, 211 et 248; Teune 1968, 128; McKennell 1973, 255; Jacobson et Lulu 1974, 217). Il faut pourtant remarquer que la connotation ne saurait être entièrement représentée par un nombre quelconque d’indicateurs, comme elle ne saurait être entièrement saisie par n’importe quelle définition; cela est dû à ce que les gnoséologues appellent la nature “ouverte” des concepts.

Les avantages techniques d’une pluralité d’indicateurs sont aussi soulignés (on réduit le poids des lacunes et des erreurs pendant la récolte des données: Zetterberg 1954, 118; Scott 1968, 211; McKennell 1970, 231), tout comme un important avantage épistémologique (l’emploi de plusieurs indicateurs “is the surest cure for overinterpretation of any single measure taken alone” — Campbell et Converse 1972, 5).

Finalement, on remarque “that any single item necessarily reflects attributes other than the one in which the investigator is interested” (Scott 1968, 211) et que leur emploi combiné permet “to obtain a degree of insurance against confounding influences” (Blalock 1961, 167; des propos semblables en Lazarsfeld 1958, 110; Cook et Selltiz 1964, 37; Galtung 1967, 121). Une telle remarque nous permet

d'introduire une autre caractéristique du rapport d'indication: le fait que l'étendue sémantique commune entre un concept et son indicateur, si par définition n'épuise pas la connotation du concept, généralement n'est pas censée épuiser la connotation de l'indicateur non plus. "Each indicator contains a combination of truth (is indicating the thing being assessed) and error (is indicating something else)" (Teune 1968, 128).

Il est même possible que "a particular observational concept be taken as an instance of more than one theoretical concept" (Doreian 1970, 5); cela peut arriver à la question d'un questionnaire "because it is relevant to more than one question of the investigator" (Lutynski 1979, 51). Tessler prend l'exemple d'une question qu'il avait introduite dans un questionnaire distribué en Tunisie: "les femmes musulmanes devraient avoir le même droit que les hommes de se marier à un étranger". Ni l'analyse des données, ni ses collègues tunisiens ont su l'aider à établir si la réaction à un tel propos devait être considérée un indicateur d'orthodoxie coranique, ou de faveur pour l'émancipation de la femme, ou bien de nationalisme (1973, 35).

A fortiori il peut arriver que le même concept (ou mieux, un concept désigné par le même terme) est considéré un indicateur de concepts différents par des chercheurs différents. Zetterberg déclare que "In a review of small groups research we were struck by a peculiar circumstance: the indicator that members of a group evaluate each other favorably is in one research tradition linked to the definition of 'sociometric popularity', in another to the definition of 'cohesiveness', in a third to a definition of 'sentiment', and in a fourth school of thought it is linked to 'morale'" (1954, 119-120).

Zetterberg se déclare choqué par la découverte; nous pensons qu'elle est la conséquence inévitable de la nature stipulative du rapport d'indication, et qu'elle pourrait sans doute être constatée à propos de nombre d'autres termes et domaines. Il se peut qu'une meilleure institutionnalisation de la communauté des sciences sociales conduise à une certaine standardisation des indicateurs des concepts les plus importants, de façon que la stipulation personnelle ou d'école soit dans une certaine mesure remplacée par la convention disciplinaire. Toutefois, la nécessité d'adapter les indicateurs que l'on choisit au contexte (Nowak 1976, 50), dans ses variations spatiales et diachroniques, pose de bien précises limites au degré de standardisation que l'on peut songer d'achever.

## 2. *La validité*

Avec un remarquable accord, la validité est définie par la question "est-ce que l'indicateur/la mesure/l'échelle/le test indique/mesure ce qu'il est cru indiquer/mesurer?" (May 1932, 136; Dodd 1942, 488; Kerlinger 1965, 457; Ebel 1968, 39; Fleishman 1968, 373; Frey 1970, 244; Merritt 1970, 44; Smelser 1976, 185; Clark 1977, 109; Ingram 1977, 18; Carmines et Zeller 1979, 12; Singer 1982, 192). Elle est vue comme une relation entre concept et indicateur (Nowak 1976, 58), entre définition nominale et indicateur (Zetterberg 1954, 114), entre concept et définition opérationnelle (Soukup et Charvat 1968, 47; McKennell 1973, 209 et 214) entre définition nominale et définition opérationnelle (Anderson 1957, 206).

**2.1.** Les psychométriciens, auxquels on doit la formation du concept, ont distingué maints types différents de validité. Pour codifier ces différences, la American Psychological Association a nommé un comité *ad hoc*, qui après quatre ans de travail a défini (1954) quatre types de validité: content, predictive, concurrent, construct.

La validité “de contenu”, aussi dite *intrinsic* ou *face* “depends on the extent to which an empirical measurement reflects a specific domain of content” (Carmines et Zeller 1979, 20); elle “rests mainly on appeal to reason regarding the adequacy with which important content has been cast in the form of test items” (Nunnally 1978, 93; des propos semblables en Deutscher 1966a, 35; Singer 1982, 192). Maints auteurs préfèrent distinguer la validité *face*, qui est “relevant to the layman” (Cronbach 1949, 183), “has to do with the surface credibility or public acceptability of a test” (Ingram 1977, 18), de la validité de contenu, qui est “established by an expert appraisal of the test content” (Davies 1977, 62). Selon Nunnally, pour cette raison-là “face validity is concerned with only one aspect of content validity” (1978, 111).

La validité prédictive est définie comme la mesure du succès avec lequel “test scores predict some important future performance” (Cronbach 1949, 106; voire aussi Kaplan 1964, 199; Carmines et Zeller 1979, 18). Il s’agit du sens “le plus classique” (Scott 1968, 251) du terme ‘validité’, qui a prévalu au moins jusqu’à la deuxième guerre mondiale (par exemple, dans *The American Soldier* la validité est conçue exclusivement comme prédictive; cela arrive encore dans un mot de la dernière encyclopédie internationale des sciences sociales: Fleishman 1968); mais à partir de McNemar (1946) elle a reçu des critiques, que l’on va voir.

La validité “concurrente” est fondée sur la corrélation avec d’autres indicateurs du même concept (Cronbach 1949, 106; Cronbach et Meehl 1955, 282; Carmines et Zeller 1979, 18). A partir de Lentz *et al.* (1932), Zubin (1934) et Richardson (1936), on compute des “coefficients de validité” qui servent à sélectionner les indicateurs.

Souvent la forme prédictive et la concurrente sont considérées ensemble comme validité “par critère” (Kerlinger 1965, 459; Carmines et Zeller 1979, 18; Singer 1982, 193; Ammassari 1984, 142) ou “pragmatique” (Ingram 1977, 18-19). C’est un autre indicateur, considéré sûrement valide, qui sert comme critère de validité. D’après Dodd, ce serait la seule forme légitime de validité: “Validity always involves a criterion. Without an accepted criterion, validity in the technical sense accepted in psychology and statistics has no meaning” (1942b, 488).

Si aucun indicateur n’est considéré un critère parmi ceux qu’on vient de corréler, on parle de validité “convergente” (Osgood 1952, 221; Campbell et Fiske 1959; Scott 1968, 254).

Une technique de validation qui semble avoir été négligée par le comité de l’A.P.A. est la *known groups validation*: “the instrument is administered to two groups of subjects, one of which can be confidently assumed to possess the attribute to a greater degree than the other”; l’indicateur est considéré valide s’il parvient à discriminer clairement les deux groupes (Scott 1968, 253). L’une des premières applications de la technique est Thurstone et Chave (1929); voire aussi Kendall et Lazarsfeld (1950); De Fleur et Westie (1958).

Dans un essai très important et connu, sur lequel on reviendra, Campbell et Fiske (1959) ont proposé une validation “discriminante”, qui s’obtient à condition que, dans une matrice, les indicateurs du concept C soient plus étroitement corrélés entre eux qu’avec les indicateurs des concepts D, E, F.

On pourrait considérer la validité discriminante comme une forme minimale de *construct validity*. D’après Cronbach et Meehl (1955, 281), qui étaient membres du comité, le concept de *construct validity* a été formulée par le comité de l’ A.P.A. (1954). Pour être *construct valid*, l’indicateur d’un concept doit montrer, avec d’autres variables définissant opérationnellement d’autres concepts ou leurs indicateurs, des corrélations empiriques de la grandeur et, bien sûr, du signe attendus sur la base de considérations théoriques (Scott 1968, 253-4; McKennell 1970, 235; Carmines et Zeller 1979, 23-26). McKennell soutient la supériorité de la *construct validation* sur les autres formes: “it is the network of relations with other variables which gives the scale its meaning”, tout en prévenant contre une vision raide et mécanique du procès de contrôle des attentes théoriques (1973, 236).

**2.2.** La forme de validation qui a soulevé plus de critiques est la prédictive. Parry et Crossly (1950) signalent qu’une prédiction incorrecte n’implique pas nécessairement invalidité; ils allèguent la récente expérience des sondages qui avaient prévu la victoire de Dewey en 1948, et qui auraient bien pu décrire fidèlement les intentions de la majorité des électeurs à 15 jours du vote.

Beaucoup des critiques adressées à la validation prédictive sont ou peuvent être généralisées à toutes les formes de validation par critère. Une critique immédiate pourrait être : s’il y a déjà un indicateur que l’on croit valide, pourquoi en chercher d’autres? A cela on a répliqué que “le nouveau instrument peut être d’emploi plus facile, ou moins conditionné culturellement, ou avoir une plus grande portée cognitive, ou bien viser seulement un aspect de ce que mesure l’instrument témoignant de sa validité” (Ammassari 1984, 142).

Une remarque très fréquente est que d’acceptables critères existent seulement pour des propriétés spécifiques, comme les “intentions to behave” (Scott 1968, 252); en dehors de cela, et donc “for many if not most measures in the social sciences, there simply do not exist any relevant criterion variables... the more abstract the concept, the less likely one is to discover an appropriate criterion for assessing a measure of it” (Carmines et Zeller 1979, 19-20; du même avis Galtung 1967, 293; Ebel 1968, 39; Kahn et Cannell 1968, 151; et même le comité de l’A.P.A., 1954, 14-15). En psychologie la position privilégiée d’un critère est souvent une création artificielle: “the asymmetry between the ‘test’ and the so-designated ‘criterion’ arises only because the terminology of predictive validity has become a commonplace in test analysis” (Cronbach et Meehl 1955, 285).

Si la validation “convergente” est montrée par des indicateurs qui ont été définis opérationnellement de façon semblable (par exemple, les *items* d’une batterie), la corrélation étroite des indicateurs peut dépendre de leur similarité structurale, comme il a été suggéré par Campbell et Fiske (1959) et confirmé par nombre d’applications de leur *multitrait-multimethod matrix*.

De plus, il faut rappeler que même s'il était démontré que tous les concepts en question sont des indicateurs valides du même concept, à la rigueur "this has nothing to do with validity, for an index may have maximum internal consistency, yet measure something quite different from what the researcher believes it measures" (Galtung 1967, 300; des propos semblables en Sletto 1936; Macfarlane 1942; Sargent 1945, 276-7; Verba 1972, 320).

En effet, on a oublié souvent que "one validates not a test, but an interpretation of the data arising from a specified procedure" (Cronbach 1971, 447); "what we have to validate is the correctness of our interpretation of the measure in terms of the concept" (McKennell 1973, 235 et 1977, 212. De remarques semblables par Kaplan 1964, 198; Hyman 1972, 292; Carmines et Zeller 1979, 12 et 17). Comme l'a remarqué Pawson (1980, 660), il est curieux de penser que la relation entre un vecteur de codes et un concept puisse être exprimée par un nombre. "Validity is an inference rather than a number" (Scott 1968, 253). C'est probablement pour cette raison que Cronbach et Meehl, deux membres importants du comité lequel proposait la *construct validity*, se montrent très prudents sur la plausibilité d'un *construct validity coefficient* (1955, 290). Mais comme la prudence n'est pas universelle, on peut encore dire, comme il y a cinquante ans, que "much validation in attitude research seems to consist of correlating unknown with unknown, and then demanding consistency" (Kirkpatrick 1936, 78).

**2.2.1.** La recherche et l'emploi de coefficients de validité s'explique par un souci d'objectivité qui pousse à cacher derrière des chiffres solides et impersonnelles toute intervention du jugement "subjectif" du chercheur. En effet, le niveau des coefficients de corrélation entre les indicateurs présumés du même concept ou de concepts différents peuvent être des éléments précieux pour le chercheur qui est en train d'évaluer la validité, car "the apparent meaning of item batteries is an insufficient basis on which to establish validity" (McKennell 1973, 235).

Mais dans le meilleur cas ces informations-là pourront arriver à illuminer le jugement du chercheur, jamais à l'éliminer. Bien au contraire, c'est le niveau même des coefficients qui, loin d'être vénéré comme la quintessence de l'objectivité scientifique, devra être évalué pour y distinguer la composante informative (le signal) de la composante-rumeur produite par des facteurs techniques (comme par exemple les ressemblances formelles entre indicateurs) mis en évidence par Campbell et Fiske (1959).

Nous pouvons croire avoir validé notre indicateur en montrant qu'il est étroitement corrélé avec un critère, ou qu'il a des corrélations du niveau attendu (étroites ou faibles) avec l'indicateur d'un autre concept. Mais par cela le problème de la validation n'a que reculé d'un pas, car qui a validé le critère ou les autres indicateurs? Comme S. Turner (1979) l'a montré si bien, le désir d'exorciser le recours à un jugement sur la connotation de l'indicateur (sur sa validité *content* ou *face*) n'ouvre qu'une *regressio ad infinitum* (A qui valide B qui valide C qui valide D...) ou un cercle vicieux (A qui peut valider B en ayant été validé par B).

Cela a été bien compris par Frey à propos de l'équivalence transculturelle des indicateurs: "Our demonstrations of equivalence always are approximations which

rest increasingly on faith and subjective judgement as one probes further and further back into their empirical support” (1970, 243); pourtant, il s’applique également à la *construct validation*, malgré le contraire avis de Cronbach et Meehl (1955, 282: “construct validity... is the only way to avoid the ‘infinite frustration’ of relating every criterion to some more ultimate standard”). Comme Carmines et Zeller ont remarqué, “whenever one assesses the construct validity of the measure of interest, one is also evaluating simultaneously the construct validity of measures of the other theoretical concepts... Unfortunately, there is no foolproof procedure for determining which one... of these interpretations... is correct” (1979, 25).

La validation par jugement sémantique, pourtant, n’est pas seulement un pas “préliminaire” (Mosier 1947; Armor 1974, 24; McKennell 1977, 212); elle est le tissu connectif qui encadre, amalgame et organise tout autre élément d’évaluation, numérique ou non.

### 3. La fidélité

**3.1.** Les concepts de fidélité et d’erreur d’observation naissent en astronomie, lorsqu’il devient habituel d’observer plusieurs fois le même phénomène (McKenzie 1981, 56). La fidélité est conçue comme une propriété de l’instrument et de l’observateur qui l’utilise pour observer maintes fois l’état d’un objet. Elle peut être considérée comme l’inverse de la variance de toutes les observations relatives au même état. Plus cette variance est grande, moins fidèle est le couple observateur-instrument.

Selon Cronbach (1947, 1-2), cette conception repose sur deux suppositions:

- (a) l’état de l’objet ne change pas spontanément pendant l’intervalle entre la première et la dernière observation;
- (b) l’état de l’objet n’est pas changé par les observations elles-mêmes.

Il serait peut-être le cas d’ajouter deux suppositions plus générales:

- (c) il y a un “état vrai” de l’objet sur la propriété en question;
- (d) les différences entre objets du même type (par exemple, les atomes du même élément) sont, si elles existent, négligeables à toutes fins utiles (dans notre cas, à la fin d’apprécier la fidélité d’un couple observateur-instrument).

Le système de ces quatre suppositions permet de considérer comme erreur d’observation toute différence entre une valeur observée et l’état vrai de l’objet sur la propriété. Comme l’état vrai est inconnu, il est putativement remplacé par la moyenne des valeurs observées, qui est aussi leur mode et leur médiane, puisque les valeurs observées ont tendance à une distribution normale.

**3.2.** Dans les sciences sociales, la supposition (d) n’est pas à sa place: les différences inter-individuelles ne peuvent point être négligées — ce qui est la source d’un écart épistémologique de la majorité des sciences naturelles.

La supposition (a) est au moins “highly unlikely” (Cronbach 1947, 4); selon Deutscher, elle aussi n’est pas à sa place: “the assumption that human thought and behaviour is static... is simply antithetical to social science” (1966b, 240).

Même la supposition (b) est fort douteuse pour la majorité des processus d'observation, comme l'on verra bientôt.

Néanmoins, beaucoup de psychologues et sociologues ont adopté une définition de fidélité qui est la plus proche possible à la définition qu'on en donne en physique. "Reliability may be briefly described as any index measuring the degree of agreement... among reobservations of the same phenomenon" (Dodd 1942b, 484). "The precision with which subjects' locations are determined may, in principle, be established by repeated measurements. To the extent that identical scores are obtained on both occasions, subjects' locations are precise and the instrument is reliable" (Scott 1968, 256). On peut lire des définitions semblables en Jahoda, Deutsch et Cook (1951, 100), Zetterberg (1954, 124), Kaplan (1964, 195), Oppenheim (1966, 79), Fleishman (1968, 372), Siegel et Hodge (1968, 56), Upshaw (1968, 65), Frey (1970, 244), Macrae (1970, 275), Clark (1977, 109), Ingram (1977, 16), Hill (1980, 407).

Même les exemples sont souvent pris de la physique (élémentaire): un règle qui mesure plusieurs fois la longueur d'une table (Spector 1981, 13-14) ou la hauteur d'une personne (Carmines et Zeller 1979, 13), une balance qui pèse un objet (Phillips 1966, 85), etc.

Dans le premier quart de ce siècle, les psychométriciens ont développé un procédé de mesure de la fidélité qui est le plus cohérent possible avec le concept qu'ils avaient. Le procédé — dit test-retest — se déroule à travers les pas suivant (Guttman 1946; Carmines et Zeller 1979, 37):

- (i) un certain instrument (d'habitude un test à plusieurs *items*) est administré au temps  $t$  dans la place  $p$  à une population de sujets (d'habitude une population "captive" de jeunes étudiants en psychologie);
- (ii) les réponses sont codées et ainsi transformées en un vecteur de chiffres;
- (iii) les pas (i) et (ii) sont répétés au temps  $t'$ ;
- (iv) on calcule un coefficient de corrélation entre les vecteurs obtenus aux pas (ii) et (iii);
- (v) le coefficient (ou son carré: McKennell 1970: 228-9) est appelé le "coefficient de fiabilité" ou de "stabilité" du test X, et il est considéré un attribut permanent et définitif du test X, le suivant n'importe où, quand et à qui lui arrive d'être administré.

Bien qu'étroitement inspiré au concept de fidélité hérité de la physique, le procédé qu'on vient de décrire introduit un nouvel aspect dans ce concept-là, un aspect qui (à notre connaissance) seulement Carmines et Zeller ont remarqué en passant (1979, 31). Dans les sciences physiques, en conséquence de la supposition (d), la fidélité d'un couple observateur-instrument peut être parfaitement bien mesurée sur la base d'observations répétées d'un seul objet; les données sont arrangées en forme de vecteur. Dans les sciences sociales, puisque la supposition (d) n'est pas acceptable, les observations sont faites sur une *pluralité* de sujets; les données sont arrangées en forme de matrice.

Par conséquent, tandis que dans les sciences physiques la fidélité d'un couple observateur-instrument est une grandeur scalaire, la fidélité calculée par le procédé *test-retest* devrait être une grandeur vectorielle (un chiffre pour chaque sujet; on

reviendra mieux sur cela au § 3.7). Mais les psychométriciens, après avoir été contraints d'introduire une pluralité de sujets par la particularité épistémologique dudit sujet, jouissent des avantages computationnels de la situation tout en rejetant ses conséquences indésirables, c. à d. les limites à la généralité du coefficient de fiabilité ainsi computed.

On va expliquer tout de suite ce qu'on vient de dire.

Pour les raisons vues plus haut, les physiciens peuvent définir opérationnellement la fidélité comme la variance d'une distribution d'observations conduites sur un seul objet. Afin que la distribution ne soit pas faible, ils ont besoin d'un certain nombre d'observations. Chaque observation produit un chiffre, et les chiffres sont rangés dans un vecteur référé à un seul cas et plusieurs moments. Les psychométriciens ne peuvent pas opérer avec un seul cas: chaque opération d'observation doit porter sur un certain nombre de cas, et ainsi produire un vecteur de chiffres qui sont référés à plusieurs cas et un seul moment. La deuxième opération produit un vecteur semblable. Puisque deux vecteurs suffisent pour calculer une corrélation, les psychométriciens peuvent définir opérationnellement la fidélité comme un coefficient de corrélation entre deux vecteurs. Le désavantage de devoir observer plusieurs cas est ainsi compensé par l'avantage de pouvoir les observer deux fois seulement au lieu des plusieurs fois auxquelles sont obligés les physiciens — un avantage que le procédé *test-retest* ne manque pas d'exploiter. L'autre avantage de la définition opérationnelle de la fidélité au moyen d'un coefficient de corrélation est que tout en partant d'une matrice de chiffres, le procédé produit une grandeur scalaire, tout comme le procédé des physiciens qui part d'un seul vecteur.

En affectant ladite grandeur scalaire comme "coefficient de fiabilité" unique, les psychométriciens dénie les mêmes différences parmi les individus de la population observée dont ils avaient tenu compte en calculant un coefficient de corrélation; de plus, en conférant un caractère définitif et universel (le "coefficient de fiabilité" du test X) à un chiffre relatif à une population donnée observée dans un moment donné, ils dénie toute différence entre celle population-là et toute autre population possible.

Des suppositions énumérées au § 3.1, (a) (b) et (c) restent inchangées; la (d) est remplacée par la:

(d') l'importance des différences inter-individuelles étant donnée, la fidélité d'un instrument ne peut pas être établie si l'on observe un sujet seulement; on a besoin de  $n$  sujets au moins. Toutefois, ils peuvent être recrutés parmi les sujets "à disposition" (d'habitude, les étudiants d'un cours de psychologie).

D'autres suppositions s'ajoutent:

(e) la fidélité d'un instrument est la même pour tous les sujets observés, et peut être exprimée par une seule chiffre;

(f) une fois calculé pour une population de sujets, le coefficient de fiabilité peut être étendu à n'importe quelle autre population;

(g) une fois calculé sur des observations faites dans la place  $p$  et aux moments  $t$  et  $t'$ , le coefficient de fiabilité du test X ne doit pas être calculé à nouveau quand le test X est administré dans n'importe quelle autre place et quel autre moment;

(h) une fois calculé quand le test X a été administré par M. Dupont dans les conditions  $x, y, z$ , le coefficient de fiabilité ne doit pas être recalculé quand le test X est administré par M. Suzuki dans les conditions  $x', y', z'$ .

Les suppositions (e), (f), (g), (h), leur plausibilité mise à part, contredisent la supposition (d'). Les physiciens ne sont pas coupables d'une telle contradiction car ils supposent (d) que les différences entre objets du même type sont négligeables. D'autre part, ils tiennent compte des différences entre les observateurs, et donc ils évitent la supposition (h), en se gardant d'attacher un coefficient de fidélité à n'importe lequel de leurs instruments en faisant abstraction de l'opérateur. Là aussi il n'y a pas de contradiction, car les observateurs n'ont pas la même nature que leurs objets, et donc ils ne sont pas affectés par la supposition (d).

**3.3.** Les suppositions (e), (f), (g), (h) ont été critiquées dans le cadre du général débat épistémologique — auquel on renvoie — sur les sciences sociales, car elles ne sont que l'application au coefficient de fiabilité d'une position épistémologique plus générale.

Les suppositions (a) et (b) ont été critiquées avec égard au procédé *test-retest*.

On a remarqué que l'état d'un individu change spontanément entre la première et la seconde collecte, et l'on a apporté des données de *panel* confirmant (ce que l'on pouvait attendre) que la grandeur du changement est fonction de la longueur de l'intervalle (Rozelle et Campbell 1969; Converse 1970).

Mais plus cet intervalle-là est réduit, plus est fort le changement qui peut être provoqué par la première observation. Dans toutes les propriétés impliquant une quelque forme d'habileté ou connaissance spécifique, la première collecte peut comporter une croissance d'habileté (*learning* ou *practice effects*: Cronbach 1949, 619; Anastasi 1953, 190-1; Webb *et al.* 1966, 19; Siegel et Hodge 1968, 56) ou d'intérêt (Carmines et Zeller 1979, 39). La réaction générale d'un individu à une seconde collecte sera différente (plus grande confiance et donc sincérité; plus grand contrôle de la situation et donc opportunisme; plus grand ennui, hostilité, etc.: voire Scott 1968, 239; Kahn et Cannell 1968, 162), et cela pourra influencer son état observé sur n'importe quelle propriété.

On ne peut pas exclure des changements non-accidentels dans l'habileté, les attitudes, les attentes de l'observateur non plus (Webb *et al.* 1966, 22; Kahn et Cannell 1968, 162).

Tous les changements, naturels ou induits, produisent une réduction dans le niveau de la fiabilité *test-retest*. D'autre part, le niveau peut être artificiellement haussé si les réponses données pendant la première collecte sont rappelées et répétées par insouciance ou par souci de cohérence: Kuder et Richardson 1937, 151; Baughman 1958, 134; Ingram 1977, 17).

Galtung (1959) remarque qu'une hausse artificielle peut être produite parce que les changements naturels sont balancés par ceux que produit la collecte. Kirkpatrick observe que "retest reliability would be increased by the presence of irrelevant propositions which evoke consistent answers" (1936, 86). Personne ne semble mentionner à ce propos la fausse homogénéité, diachronique aussi bien que

synchronique, produite par certaines définitions opérationnelles (par exemple, l'organisation en batterie des *items*, si fréquente dans les tests et les questionnaires).

La suggestion de Cronbach (mentionner la longueur de l'intervalle entre test et retest: 1947, 14-5) pourrait apporter une amélioration importante si le changement naturel, qui est une fonction monotone du temps, était le seul facteur troublant — ce qui est bien loin d'être vrai, comme on vient de constater.

Le procédé *test-retest* n'est plus si populaire chez les psychométriciens comme auparavant; mais l'on peut présumer que c'est moins les critiques vues plus haut qui ont réduit sa popularité plutôt que de banales raisons de (relative) incommodité. "The test-retest method has one great practical disadvantage: it is often quite impossible to get permission to test a group of students twice" (Ingram 1977, 17-18); il est aussi "unduly expensive" (Carmines et Zeller 1979, 79).

**3.4.** D'autres concepts de fidélité ont été formulés ensuite. Quand plusieurs chercheurs doivent codifier le même matériel (visuel, oral, écrit), on a développé un concept lié au degré d'accord manifesté par leurs affectations (Hempel 1961, 142; Bowers 1964, 587; Kaplan 1964, 195). Ce concept-là a été désigné par les termes "interindividual reliability" (Zetterberg 1954, 124), "interindividual constancy" (Galtung 1959, 113), "intersubjectivity" (Galtung 1967, 28), "intercoder reliability" (Singer 1982, 194), "multi-judge reliability" (Cartwright 1956).

Les psychométriciens ont pensé que, comme la fidélité pouvait être définie par l'accord entre observateurs différents, elle pouvait aussi être définie par l'accord entre instruments différents. Ainsi est née la technique connue comme "parallel forms" (Thurstone 1928, 552) ou "equivalent forms" (Fleishman 1968, 373), qui consiste à corrélérer deux vecteurs de chiffres obtenus en administrant aux sujets deux tests différents mais jugés parallèles ou équivalents. Le coefficient ainsi produit a été nommé "coefficient of equivalence" (Amer. Psych. Ass. 1954, 29; Ebel 1968, 38), et la fidélité a été définie "the degree to which measurement instruments of the same type give the same results" (D'Andrade 1974, 160; voir aussi Ingram 1977, 16).

Une variante, qui a été la technique préférée entre 1935 et 1955, consiste à diviser un test en deux (*split-half*), en assignant les *items* à l'une ou l'autre moitié et corrélant les vecteurs résultant de l'administration (Brownell 1933; Kuder et Richardson 1937, 151-2). Le coefficient qu'on obtient a été appelé *internal consistency* (Cronbach 1947, 6-8; Amer. Psych. Ass. 1954, 29; Ebel 1968, 38).

Le nouveau concept de fidélité comme congruence a permis aux psychométriciens d'éviter les critiques portées à la supposition (a) — absence de changement spontané — mais non les critiques à la supposition (b), car l'observation à travers un long test produit sans doute des changements pendant l'administration même du test.

Toutefois, la préoccupation principale pour nombre de psychométriciens n'étaient pas les critiques méthodologiques, mais plutôt la baisse des coefficients de fiabilité produite par la longueur réduite des tests *split-half* ou *parallel forms* par rapport aux tests du procédé *test-retest*. Comme remède, on appliquait une formule calculée par Spearman (1910) et Brown (1910), qui permettait de manipuler les coefficients en fonction de la longueur du test.

Un remède plus satisfaisant fut proposé par deux membres du laboratoire psychométrique de Thurstone (Kuder et Richardson 1937), avec une formule calculant un coefficient de fiabilité sur la base d'une matrice des corrélations entre tous les *items* d'un seul test plutôt que — comme il arrivait pour *test-retest*, *split-half* et *parallel forms* — sur deux seuls vecteurs, chacun portant les scores individuels sur un test (ou demi-test) globalement considéré.

Ce passage-ci a eu de graves conséquences méthodologiques, car il éliminait ce qu'il y avait de bon dans les vieux procédés, c. à d. le fait que les scores globaux d'un individu sur un test sont une base bien meilleure que les scores sur chaque *item* pour calculer la fiabilité d'un test. On franchissait ainsi le dernier seuil sur la route de la transformation de la fidélité en un problème exclusif de manipulation mathématique de chiffres.

On doit ajouter que la formule de Kuder et Richardson contemplait seulement des *items* dichotomiques, c. à d. du format le plus sensible aux erreurs (dans une dichotomie, toute erreur change le blanc en noir et le noir en blanc) et aux distorsions que les distributions déséquilibrées introduisent dans le coefficient de corrélation.

Plus tard, on a proposé le coefficient *alpha*, valable pour *items* de n'importe quel format (Cronbach 1951); on a trouvé une formule fort simplifiée qui approximait *alpha* (McKinnell 1970); on en a dérivé d'autres coefficients pour des situations particulières (Raju 1977). On a aussi développé le coefficient *omega* fondé sur l'analyse factorielle (Heise et Bohrnstedt 1970), et le coefficient *theta* basé sur l'analyse des composantes principales (Armor 1974).

**3.5.** Quels qu'ils soient les mérites et les limites techniques des coefficients qu'on a mentionnés (et pour lesquels on renvoie à Giampaglia 1986), ils n'ont apporté aucun changement important au concept de fidélité des psychométriciens, qui a été adopté par la majorité des sociologues et politologues. Le seul changement a été, comme on a dit plus haut, dans le sens du renforcement de la tendance à s'en remettre exclusivement à des manipulations mathématiques de vecteurs de chiffres, sans aucune considération pour l'effective correspondance de ces chiffres aux états individuels sur la propriété. "The cumulative experience of these studies has taught us to have confidence in certain measurement techniques... Persons using these techniques, therefore, no longer feel obliged to go through the tedious process of checking their reliability" (Zetterberg 1954, 124).

On demande aux coefficients d'épargner des travaux ennuyeux tout en rassurant sur la qualité "scientifique" de ce qu'on a fait. Ce sont des techniques auxquelles on veut se fier. Dorénavant on appellera donc 'fiabilité' le concept des psychométriciens, comme on l'a déjà fait pour les coefficients relatifs, afin de souligner sa différence du concept ordinaire de fidélité, sur laquelle on reviendra. On va aussi soutenir que le concept de fiabilité et les procédés relatifs sont parties d'une ceinture protectrice de concepts, termes et techniques qui minimisent le contact du sociologue avec son objet, la société, en le renfermant dans une tour d'ivoire de certitudes "scientifiques".

Dans le paragraphe qui suit on va exposer quelques conséquences négatives de l'adoption du concept de fiabilité.

**3.5.1.** Dans le § 2.1.1 on a vu que, afin de cacher l'intervention "subjective" du jugement sémantique du chercheur, on a réduit le problème de la validité à une corrélation entre vecteurs de chiffres. On vient de voir que le problème de la fidélité a été réduit aux mêmes termes par l'adoption du concept et du cadre de suppositions des physiciens, et d'un système de suppositions supplémentaires qui permettent de méconnaître les conséquences méthodologiques des différences interindividuelles, c. à d. l'inadmissibilité d'un concept vectoriel de fidélité, et la nécessité d'évaluer séparément la fidélité de chaque acte d'observation.

La conséquence inévitable de tout cela a été l'obscurcissement des différences non seulement techniques, mais aussi conceptuelles, entre fidélité et validité. Une telle confusion est déjà manifeste à partir des années trente dans les oscillations terminologiques (par exemple en Horst 1934; 1936) et dans le fait que la même technique est parfois proposée pour mesurer l'une et l'autre; par exemple, Likert (1932, 93) propose, sous le nom de *internal consistency*, une technique de mesure de la fiabilité qui n'est autre chose que la *known groups*, considérée comme une technique de mesure de la validité (voir § 2.1; voir aussi la liste de techniques de validation dressée par Frey 1970, 248-9).

Campbell et Fiske ont perçu (correctement si l'on se borne à des corrélations de vecteurs) que fiabilité et validité sont seulement les extrêmes d'un 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, 83).

L'idée que la fiabilité consiste en l'accord mathématique de vecteurs représentant "maximally similar methods", tandis que la validité consiste en l'accord de vecteurs représentant "maximally different methods" est assez répandue (après Campbell et Fiske 1959, Siegel et Hodge 1968; Bentler 1972, 343). Une variante, qui ébauche une sortie du cadre de la matrice des données, est l'idée que "what better distinguishes questions of reliability from validity is the assumption that only underlying concepts and random error affect measures. In contrast, matters of validity arise when other factors... are seen to affect the measures in addition to one underlying concept and random error" (Althausen et Heberlein 1970, 152; voir aussi Werts et Linn 1969).

D'autres auteurs se limitent à constater que "reliability and validity are inextricably intertwined" (Davies 1977, 51) or "closely related" (Carmines et Zeller 1979, 34), ce qui est critiqué par les interactionnistes symboliques (Blumer 1954, 7; Deutscher 1966a, 34), mais jugé tout à fait naturel, par exemple, par Lumsden (1976).

On établit de simples relations mathématiques entre les formules des coefficients de validité et de fiabilité (voir par exemple Cronbach 1949, 171), et plus tard on conçoit des modèles “for simultaneously handling both questions of reliability... and validity” (Siegel et Hodge 1968, 56), ainsi que les relations causales entre les variables, au moyen de la *path analysis* (voir § 1.2).

**3.5.2.** Une autre conséquence de la réduction de la fidélité à un coefficient est que maximiser le niveau de ce coefficient-là devient une fin à elle-même. De nouvelles formules sont proposées avec ce but déclaré (Horst 1936; Richardson 1936; Callender et Osburn 1977). Dans la psychologie de laboratoire, les tests “objectifs” sont préférés aux tests non-directifs (projectifs et autres) car “the scores obtained from objective tests tend to be more reliable [c. à d. à produire de plus hauts coefficients de fiabilité] than those obtained from essay tests” (Ebel 1968, 34).

A cause de la structure des formules, plus nombreux sont les items dans le test, plus haut sera — *ceteris paribus* — le coefficient de fiabilité (Davies 1977, 58); par conséquent, on tend à produire des tests de longueur exagérée. Comme l’on remarquait plus haut, “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” (McKennell 1970, 236).

Le phénomène qu’on vient de décrire peut être considéré caractéristique: afin de grossir le coefficient fatal, on néglige soit les effets très bien connus (Festinger 1947; Converse 1970, 178-9; Noelle-Neumann 1970, 194; Ferrand et Martel 1986) de la fatigue et de l’ennui sur la fidélité des réponses (à ne pas confondre avec la fiabilité mesurée par les formules — la différence entre les deux concepts ne saurait être plus claire), soit le droit moral de chaque sujet à être “observé” au moyen d’instruments conçus afin d’évaluer *ses* capacités, attitudes, opinions, valeurs, plutôt que conçus afin de gonfler un coefficient (global) de fiabilité.

D’autres raisons éloignent le concept technique de fiabilité du concept ordinaire de fidélité. “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 *et al.* 1960, 703). Un exemple de ce mécanisme-là est apporté, sans aucun signe d’ironie, par un petit manuel très lu et entièrement dédié à des questions de fiabilité: “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 et Zeller 1979, 13).

Dans les sciences sociales, une fiabilité tout à fait artificielle peut être produite en réduisant l’étendue du domaine sémantique touché par les items (McKennell 1977, 211), en homogénéisant la forme structurale et même verbale des questions (Hyman *et al.* 1954, 30; McKennell 1970, 234). Souvent on réduit le nombre des réponses prévues pour chaque question, parce que cela tend à produire des coefficients plus élevés (spécialement dans le *test-retest*) en empêchant que les réponses soient

éparpillées sur des catégories trop détaillées — qui pourtant reflèteraient les nuances des attitudes et des opinions (ce que déplore Galtung 1967, 28). Une batterie qui n'obtient que des *response sets* aura une fiabilité parfaite selon n'importe laquelle des formules courantes; l'on pourrait allonger à plaisir la liste des artifices et des accidents techniques qui peuvent hausser un coefficient de fiabilité.

**3.6.** La différence entre le concept des psychométriciens (appelé ici fiabilité) et le concept ordinaire de fidélité n'a pas manqué d'être perçue: "Use of the term 'reliability' in its technical sense of a test of consistency of response is unfortunate. The nontechnical, layman's sense of reliability is trustworthiness. Behavioral 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, 265-66).

Comme ils sont désormais entrés d'une manière durable dans les habitudes des psychologues, sociologues et politologues, le concept de fiabilité et les techniques relatives barrent la route au développement d'un concept et de techniques plus adéquats à la situation épistémologique des sciences sociales. "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, 399-400).

Même si l'on partage les derniers propos par Frank, on voudrait faire des réserves sur son opinion que les psychométriciens ont adopté leurs techniques de mesure de la fiabilité faute d'une voie alternative.

Comme on le verra au § 3.8, il y a des voies alternatives et des chercheurs qui les parcourent, occasionnellement voire habituellement. Mais parcourir ces voies-là exige beaucoup plus de temps, de patience et de créativité de la part du chercheur et — surtout — n'est pas compatible avec l'image du savant qui, habillé d'une blouse blanche et hautain dans sa tour d'ivoire, manie des hypothèses et des théories au moyen de la logique formelle, et/ou matrices des données à l'aide d'un ordinateur.

Le concept de fiabilité et les techniques relatives, tout comme les techniques pour "mesurer" la validité, font partie de plein droit d'une ceinture protectrice de concepts et techniques qui ont la fonction de permettre au sociologue de minimiser ses contacts avec la société tout en gardant la prétention de parler d'elle, voire d'en chercher les "lois".

Le modèle d'organisation de la recherche aujourd'hui prédominant aux sciences sociales garde un seul point de contact avec son objet, au moment de la collecte des données. Mais les définitions opérationnelles qui permettent la transformation en données des états individuels (voir § 1.1) prévoient des opérations aussi standardisées et mécaniques que possible, afin d'éliminer la "subjectivité" (et,

avec elle, toute compréhension interprétative de la réalité). Ces opérations mécaniques et standardisées sont confiées à de véritables chaînes de collaborateurs (experts d'échantillonnage, agences de sondages, établissements publics de collecte des données, administrateurs de tests, enquêteurs, codificateurs, opérateurs et officiers du recensement, compteurs de mots), ce qui délivre le sociologue de toute tâche ancillaire et menue.

Bien que des négligences et des fautes commises par chaque maillon de ces chaînes aient été dénoncées et prouvées plusieurs fois, la majorité des chefs de recherche — ceux qui manient les fonds et signent les publications — n'exercent aucun contrôle sur les opérations de collecte, soit pour épargner leur temps précieux soit par souci de neutralité. Ainsi, les innombrables décisions, menues et non, qui sont nécessaires à chaque pas du procédé sont prises par des exécutants qui dans la plupart des cas n'ont aucun intérêt scientifique à la recherche et suivent le principe de l'effort minimum; même s'ils sont scrupuleux et animés des meilleures intentions, ils manquent des connaissances et des informations nécessaires pour encadrer le problème dans une vision générale de la recherche et du domaine de la discipline.

Toutes ces décisions, qui collectivement déterminent les résultats de n'importe quelle recherche, sont d'habitudes prises par chaque exécutant à l'insu de tout autre participant à l'entreprise. Elles sont rarement communiquées aux autres exécutant la même tâche, et encore plus rarement rapportées aux niveaux supérieurs dans l'organisation de la recherche. Même si elles le sont, elles finissent pour se perdre le long de la chaîne qui arrive au directeur de la recherche. On en trouve encore quelque trace dans les rapports de recherche, mais absolument aucune trace dans les essais et les manuels d'épistémologie.

Bien sûr, le nombre même de ces décisions empêcherait d'en rapporter plus qu'une petite part; mais au fur et à mesure qu'on passe des rapports de recherche aux publications qui en présentent les résultats, aux élaborations théoriques et épistémologiques, le fait qu'il est des décisions à prendre est censuré bien plus radicalement qu'il serait nécessaire pour des raisons pratiques. Ce fait est réfoulé de la conscience collective puisqu'il troublerait la vision épistémologique prédominant du processus d'observation comme une photographie de la réalité, qui comme toute photographie n'aurait besoin d'aucune décision sauf celle de photographeur — ce qui, en parenthèses, est incorrect même pour les photographies.

L'observation doit être photographique puisque son résultat — la matrice des données — doit être une photographie de la réalité, voire la réalité même. Quand le chercheur (ou plutôt, encore une fois, ses collaborateurs: cette fois les experts de statistique et d'informatique) traite une matrice des données par des techniques d'analyse, dans son esprit on ne trouve d'habitude aucune trace du fait que la matrice représente (un morceau de) la réalité seulement grâce à un véritable échafaudage de conventions et une myriade de décisions "subjectives" et fort peu coordonnées que chaque exécutant a prises.

Ces conventions et décisions (de même que les conventions et décisions nécessaires à tout traitement statistique, et encore plus à toute interprétation de ses résultats) doivent être réfoulées afin qu'on puisse penser et proclamer que les résultats se réfèrent directement à la réalité.

La ceinture protectrice dont on parlait au § 3.5 a la fonction de supporter cette mystification. Plus en détail, certains concepts-termes de la ceinture ont la fonction de cacher la distinction entre un aspect de la réalité et son correspondant conventionnel dans la matrice des données. Très important à ce propos est le rôle du concept de variable, dont les référents sont simultanément un vecteur-colonne de chiffres dans la matrice, un concept dans l'esprit du chercheur, et un ensemble d'états individuels qui sont — d'une façon plus ou moins conventionnelle — considérés comme relatifs à la même propriété. Un 'cas' désigne simultanément un vecteur-ligne de chiffres dans la matrice et un individu réel. Une 'donnée' désigne simultanément un état individuel sur une propriété, l'information non-traitée à propos de cet état (comme dans "collecte des données"), et la même information passée à travers toutes les manipulations prévues par la définition opérationnelle, et placée dans une position univoquement définie dans la matrice.

D'autres concepts-termes de la ceinture remplissent la fonction de réduire (en légitimant telle réduction) à des manipulations mathématiques (donc "objectives") de vecteurs celles qui sont — et encore plus devraient être — des opérations intellectuelles qui demandent des jugements et décisions "subjectifs" du chercheur, soit comme possesseur de connaissances sur l'objet et le domaine de sa recherche, soit comme connaisseur des ficelles de son métier (par exemple, comme l'on doit interpréter certains résultats statistiques, combien de confiance on peut avoir aux résultats des définitions opérationnelles de certaines propriétés, etc.).

Dans le § 2.1.1 on a vu comme une telle fonction réductrice est remplie par les concepts de validité prédictive, concurrente, convergente et même par certaines interprétations du concept de *construct validity*. Dans le § 3.5 on a vu comme la même fonction est remplie par le concept de fiabilité. Un autre cas important est le concept d'explication, que dans bien d'écrits des sciences sociales l'on trouve réduit à la corrélation statistique entre un vecteur de scores sur une variable dépendante et un ou plus vecteurs de scores sur une ou plus variables indépendantes.

Il nous semble évident que l'effet immédiat de la ceinture protectrice est de minimiser le contact du sociologue avec son objet et l'engagement de son temps, connaissances, créativité dans la recherche. Toutefois, nous pensons que la fonction centrale de la ceinture est de préserver l'image de la science comme une entreprise "objective".

Un tel but est poursuivi en standardisant et objectivant tout procédé de collecte et d'analyse de l'information, en détournant l'attention, par les astuces terminologiques que l'on vient de voir, des larges espaces qui restent à remplir par les décisions du chercheur ou de ses collaborateurs, etc. Mais le résultat vraiment essentiel qu'on obtient en interprétant toute opération cognitive comme une opération sur et dans les bornes de la matrice des données est l'établissement d'une situation de monisme gnoséologique.

Le monisme est la condition gnoséologique de toute prétention à l'objectivité: lorsqu'on admet une pluralité de sources de la vérité, un conflit peut se produire entre ces sources, d'où la nécessité de décider entre les prétentions contrastantes à la vérité — ce qui est évidemment incompatible avec l'idée d'une objectivité photographique et impersonnelle.

Dans notre cas, on ne peut avoir monisme si l'on reconnaît que le sociologue a des sources d'information sur la réalité autres que la matrice des données; en particulier, si l'on reconnaît qu'un patrimoine de connaissances non formalisées et d'expériences dans la recherche est nécessaire pour décider quelles opérations sont à effectuer sur la matrice, pour savoir comment les effectuer, et pour en interpréter les résultats.

**3.7.** Si l'on veut abandonner le concept de fiabilité pour les raisons que l'on vient de voir, comment former un concept de fidélité qui puisse orienter les opérations de recherche en psychologie, sociologie et science politique?

On doit d'abord remplacer la longue liste des suppositions nécessaires au concept vectoriel de fiabilité (voir § 3.2) par une liste bien plus courte de suppositions mieux adaptées à la situation épistémologique des sciences sociales. La supposition (c) est remplacé par la:

(c') même en supposant qu'il y a, à chaque moment, un "état vrai" de chaque sujet sur la propriété qui nous intéresse, dans la plupart des cas on ne peut être sûrs qu'on le connaît. Toutefois, plus l'on cherche soigneusement, plus de chances l'on a d'approcher de sa connaissance, ou de vérifier son inexistence. Au contraire, plus superficielle, standardisée, mécanique est la recherche, plus de chances l'on a d'enregistrer une donnée qui altère gravement l'état vrai, ou qui ne correspond à aucun état effectif.

Les suppositions (d'), (e), (f), (g), (h) sont remplacées par la:

(d'') étant donné que non seulement les différences inter-individuelles sont importantes, mais aussi l'est toute différence entre les situations d'observation, la fidélité est une propriété de chaque acte particulier d'observation d'un état individuel, donc de chaque donnée qui en est le résultat.

Le jugement sur la fidélité doit porter sur la correspondance entre l'idée que l'on s'est faite de l'état d'un individu sur une certaine propriété, et le résultat d'un certain acte d'observation guidé par une certaine définition opérationnelle de la propriété. Chaque jugement particulier ne peut être étendu à d'autres actes d'observation, par le même ou d'autres instruments et observateurs, de l'état du même ou d'autres individus.

**3.7.1.** A notre connaissance, les suppositions nécessaires à un concept non-vectoriel de fidélité sont à peine ébauchées par Frank (1939, 400) et Gostkowski (1974, 22), et formulées moins vaguement par Cicourel (1964, 80 et 109-11). Mais plus que des élaborations méthodologiques, le concept non-vectoriel de fidélité a inspiré des recherches sur la "qualité des données" (que l'on n'a pas reconnu comme des recherches sur la fidélité/fiabilité parce que ces termes étaient identifiés avec les procédés des psychométriciens).

Avant d'analyser ces recherches, on voudrait souligner que la littérature abonde en témoignages et considérations sur l'importance de l'un ou l'autre aspect de la situation particulière pour le résultat de l'observation. On peut bien s'étonner que toutes ces considérations n'aient pas produit une critique généralisée du concept vectoriel.

La performance dans un test d'habileté sera influencé par les conditions physiques et mentales des sujets à ce moment-là (Thorndike 1949, 73). Les mêmes conditions, ainsi que l'humeur du moment et les petites fluctuations de l'attention pourront influencer la réponse à une question (Kendall 1954; Converse 1970; Noelle-Neumann 1970).

Bien de recherches (Schanck 1932; Gorden 1952; Rose 1961; Bindman 1965; Lutynska 1978; Przybilowska et Kistelski 1981) ont montré que les réponses du même sujet à la même question peuvent changer radicalement en fonction de la perception qu'il a de celui qui lui soumet le test ou le questionnaire, de ceux qui assistent, et surtout de son rôle, soit dans la situation spécifique soit plus en général (on lui demande ses idées sur les opinions prédominantes dans sa culture, ses opinions comme citoyen, père, etc., ou ses opinions plus intimes et privées?) Pour les sujets, l'interview est "an occasion for an act of self-presentation" (Harré 1981, 153; voir aussi Cook et Selltiz 1964, 39-44) et ils vont répondre selon leur "perception of the appropriate mode of taking part in a research exercise... and of... the appropriate form of response" (Pawson 1982, 45). "A given individual may hold a number of attitudes toward the same object, perhaps applicable to different segments, but logically incompatible if they should be confronted with each other in the same setting" (Rose 1961, 266. Sur la différence entre opinions publiques et privées, voir aussi Rose 1950, 206-8; Cicourel 1964, 56-7; Deutscher 1972, 324-5; Gostkowski 1974, 16).

Comme le langage des questionnaires et des tests est peu familier à la majorité des sujets ("C'est difficile, pour un chercheur, d'imaginer à quel point les enquêtés peuvent être étrangers à son propre univers et souvent même à son langage": Pinto 1964, 700. Voir aussi Cicourel 1964, 110; Bourdieu *et al.* 1968, 63), des sujets vont interpréter toute question et toute tâche moins qu'élémentaire sur la base des indications que les circonstances (tout ce qui s'est passé auparavant dans l'interview) lui fournissent (Thurstone 1922; Sargent 1945, 269; Cicourel 1964, 80; Belson 1981).

S'il est questionné sur un thème qui lui est étranger — ce qui arrive très souvent aux individus marginaux par leur instruction, leur occupation, leur âge — le sujet répondra au hasard (Converse 1964 et 1970; Gergen et Back 1966) et pourra être influencé par n'importe quel accident de la formulation verbale, des questions et réponses passées, des événements personnels récents (Haynes 1964; Bishop *et al.* 1982). Cela peut arriver même si le sujet est intéressé au thème mais la situation de l'interview le presse: "l'enquêté, pris par surprise, obligé de répondre rapidement, même s'il connaît le sujet, n'a pas la possibilité d'organiser sa réponse. Inévitablement il donnera l'information la plus facile à exprimer, à faire comprendre, c'est à dire la plus superficielle" (Pinto 1964, 674; voir aussi les remarques des interviewées rapportées par Hyman *et al.* 1954). En général, "le sociologue... s'expose à... prendre pour expression d'une attitude profonde des jugements superficiels suscités par la nécessité de répondre à des questions sans nécessité" (Bourdieu *et al.* 1968, 57; un propos semblable en Lutynski 1979, 45).

Si l'on ajoute les fautes accidentelles commises par l'enquêteur, le codificateur, etc., il y en a assez pour conclure que chaque donnée dans un vecteur représentant

une propriété peut être plus ou moins fidèle pour des raisons indépendantes de chaque autre donnée.

- ABELL, Peter (1969) *Measurement in Sociology. II: Measurement, Structure and Sociological Theory*, "Sociology" III, 3 (septembre): 397-411.
- ALTHAUSER, Robert P. et HEBERLEIN, Thomas A. (1970) *Validity and the Multitrait-Multimethod Matrix*, pp. 151-69 dans Edgar F. Borgatta et George W. Bohrnstedt (eds.), *Sociological Methodology 1970*. San Francisco: Jossey-Bass.
- AMERICAN PSYCHOLOGICAL ASSOCIATION (1954) *Technical Recommendations for Psychological Tests and Diagnostic Techniques*, "Psychological Bulletin" LI, Supplément, 2.me partie : 1-38.
- AMMASSARI, Paolo (1984) *Validità e costruzione delle variabili: elementi per una riflessione*, "Sociologia e Ricerca Sociale" V, 13 (avril): 141-156.
- ANASTASI, Anne (1953) *Differential Psychology*. London: Macmillan.
- ANDERSON, Bo (1957) *Some Notes on Operationism and the Concept of Validity*, "Acta Sociologica" II, 4: 202-13.
- ARMOR, David J. (1974) *Theta Reliability and Factor Scaling*, pp. 17-50 dans Herbert L. Costner (ed.), *Sociological Methodology 1973-1974*. San Francisco: Jossey-Bass.
- BARTLETT, Claude J., QUAY, Lorence Childs et WRIGHTSMAN, Lawrence S. (1960) *A Comparison of Two Methods of Attitude Measurement: Likert Type and Forced Choice*, "Educational and Psychological Measurement" XX, 4 (hiver): 699-704.
- BAUGHMAN, E. Earl (1958) *The Role of the Stimulus in Rorschach Responses*, "Psychological Bulletin" LV, 3 (mai): 121-47.
- BELSON, William A. (1981) *The Design and Understanding of Questions in the Survey Interview*. London: Gower.
- BELSON, William A., MILLERSON, B. L. et DIDCOTT, P. J. (1968) *The Development of a Procedure for Eliciting Information from Boys about the Nature and Extent of their Stealing*. London School of Economics.
- BENTLER, P. M. (1972) *A Lower-Bound Method for the Dimension-Free Measurement of Internal Consistency*, "Social Science Research" I, 4 (décembre): 343-57.
- BINDMAN, Aaron M. (1965) *Interviewing in the Search for "Truth"*, "Sociological Quarterly" VI (été): 281-88.
- BISHOP, George et al. (1982) *Effects of Presenting One versus Two Sides of an Issue in Survey Questions*, "Public Opinion Quarterly" XLI, 1 (printemps): 69-85.
- BLALOCK, Hubert M. (1961) *Causal Inferences in Nonexperimental Research*. Chapel Hill: University of North Carolina Press.
- BLALOCK, Hubert M. (1968) *The Measurement Problem: A Gap Between the Languages of Theory and Research*, pp. 5-27 dans Hubert M. Blalock et Ann B. Blalock (eds.), *Methodology in Social Research*. New York: McGraw-Hill.

- BLALOCK, Hubert M. (1969a) *Theory Construction. From Verbal to Mathematical Formulations*. Englewood Cliffs: Prentice-Hall.
- BLALOCK, Hubert M. (1969b) *Multiple Indicators and the Causal Approach to Measurement Error*, "American Journal of Sociology" LXXV, 2 (septembre): 264-72.
- BLALOCK, Hubert M. (ed., 1974) *Measurement in the Social Sciences. Theories and Strategies*. Chicago: Aldine.
- BLUMER, Herbert (1954) *What Is Wrong With Social Theory?*, "American Sociological Review" XIX, 1 (février): 3-10
- BOURDIEU, Pierre, CHAMBOREDON, Jean-Claude et PASSERON, Jean-Claude (1968) *Le métier de sociologue. Problèmes épistémologiques*. Paris: Mouton.
- BOWERS, Raymond V. (1964) *Reliability*, pp. 587-8 dans Julius Gould et William L. Kolb (eds.), *A Dictionary of the Social Sciences*. Glencoe: Free Press.
- BROWN, William (1910) *Some Experimental Results in the Correlation of Mental Abilities*, "British Journal of Psychology" III (octobre): 296-322.
- BROWNELL, William A. (1933) *On the Accuracy with which Reliability May Be Measured by Correlating Test Halves*, "Journal of Experimental Education" I: 204-15.
- CALLENDER, John C. et OSBURN, H. G. (1977) *A Method for Maximizing Split-Half Reliability* (hiver): 819-25.
- CAMPBELL, Angus et CONVERSE, Philip E. (1972) *Social Change and Human Change*, pp. 1-16 dans A. Campbell et P. E. Converse (eds.), *The Human Meaning of Social Change*. New York: Russell Sage.
- CAMPBELL, Donald T. et FISKE, Donald W. (1959) *Convergent and Discrimination Validation by the Multitrait Multimethod Matrix*, "Psychological Bulletin" LVI, 2 (mars): 81-105
- CAMPBELL, E. Q. et KERCHOFF, A. C. (1972) *A Critique of the Concept 'Universe of Attributes'*, "Public Opinion Quarterly" XXI: 295-303.
- CANNELL, Charles F. et ROBINSON, John P. (1971) *Analysis of Individual Questions*. Ann Arbor: Survey Research Center.
- CARMINES, Edward G. et ZELLER, Richard A. (1979) *Reliability and Validity Assessment*. London: Sage.
- CARTOCCI, Roberto (1985) *Differenze territoriali e tipi di voto: le consultazioni del maggio-giugno 1985*, "Rivista Italiana di Scienza Politica" XV, 3 (décembre): 421-54.
- CARTWRIGHT, Desmond S. (1956) *A Rapid Non-Parametric Estimate of Multi-judge Reliability*, "Psychometrika" XXI, 1: 17-29.
- CHAPIN, F. Stuart (1939) *Definition of Definitions of Concepts*, "Social Forces" XVIII, 2 (décembre): 153-60.
- CICOUREL, Aaron Victor (1964) *Method and Measurement in Sociology*. New York: Free Press.
- CLARK, Ruth (1977) *The Design and Interpretation of Experiments*, pp. 105-45 dans J. P. B. Allen et Alan Davies (eds.), *Testing and Experimental Methods*. London: Oxford University Press.

- CONVERSE, Philip E. (1964) *The Nature of Belief Systems in Mass Publics*, pp. 202-61 dans 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 dans Edward R. Tufte (ed.), *The Quantitative Analysis of Social Problems*. Reading: Addison-Wesley
- COOK, Stuart W. et SELTZER, Claire (1964) *A Multiple Indicator Approach to Attitude Measurement*, "Psychological Bulletin" LXII, 4 (juillet): 36-55.
- COSTNER, Herbert L. (1969) *Theory, Deduction and Rules of Correspondence*, "American Journal of Sociology" LXXV, 2 (septembre): 245-263.
- CRONBACH, Lee J. (1947) *Test Reliability: Its Meaning and Determination*, "Psychometrika" XII, 1 (mars): 1-16.
- CRONBACH, Lee J. (1949) *Essentials of Psychological Testing*. New York: Harper & Row.
- CRONBACH, Lee J. (1951) *Coefficient Alpha and the Internal Structure of Tests*, "Psychometrika" XVI: 297-334.
- CRONBACH, Lee J. (1971) *Test Validation*, pp. 443-507 dans Robert L. Thorndike (ed.), *Educational Measurement*. Washington: American Council on Education.
- CRONBACH, Lee J. et MEEHL, Paul E. (1955) *Construct Validity in Psychological Tests*, "Psychological Bulletin" LII, 4 (juillet): 281-302.
- CURTIS, Richard F. et JACKSON, Elton F. (1968) *Multiple Indicators in Survey Research*, "American Journal of Sociology" LXXIV, 2 (septembre): 195-204.
- D'ANDRADE, Roy G. (1974) *Memory and the Assessment of Behaviour*, pp. 159-186 dans Hubert M. Blalock (ed.), *Measurement in the Social Sciences. Theories and Strategies*. Chicago: Aldine.
- DAVIES, Alan (1977) *The Construction of Language Tests*, pp. 38-104 dans J. P. B. Allen et Alan Davis (eds.), *Testing and Experimental Methods*. London: Oxford University Press.
- DE FLEUR, Melvin L. et WESTIE, Frank R. (1958) *Verbal Attitudes and Overt Acts: An Experiment on the Saliency of Attitudes*, "American Sociological Review" XXIII: 667-73.
- DEUTSCHER, Irwin (1966a) *Looking Backward: Case Studies in the Progress of Methodology in Sociological Research*, "American Sociologist" IV, 1: 34-42.
- DEUTSCHER, Irwin (1966b) *Words and Deeds: Social Science and Social Policy*, "Social Problems" XIII: 233-54.
- DEUTSCHER, Irwin (1972) *Public and Private Opinions: Social Situations and Multiple Realities*, pp. 323-49 dans Saad Z. Nagi et Ronald G. Corwin, *The Social Contexts of Research*. New York: Wiley.
- DODD, Stuart C. (1942a) *Dimensions of Society*. London & New York: MacMillan.
- DODD, Stuart C. (1942b) *Operational Definitions Operationally Defined*, "American Journal of Sociology" XLVIII, 4 (janvier): 482-9.

- DOREIAN, Patrick (1970) *Mathematics and the Study of Social Relations*. London: Widenfield & Nicolson.
- DURKHEIM, Emile (1893) *De la division du travail social*. Paris: Alcan.
- DURKHEIM, Emile (1896) *Le suicide: Etude de Sociologie*. Paris: Alcan.
- EBEL, Robert I. (1968) *Achievement Testing*, dans *International Encyclopedia of the Social Sciences* I: 33-39.
- ETZIONI, Amitai et LEHMANN, Edward W. (1967) *Some Dangers in "Valid" Social Measurement*, "Annals of the American Academy of Political and Social Science" Vol. 373 (septembre): 1-15.
- FERRAND, D. J. et MARTEL, J. M. (1986) *Le choix multicritère des items d'une échelle de mesure*, "Mathématique et Sciences Humaines" n. 89: 35-59.
- FESTINGER, Leon (1947) *The Treatment of Qualitative Items by Scale Analysis*, "Psychological Bulletin" XLIV: 149-61.
- FLEISHMAN, Edwin A. (1968) *Aptitude Testing*, pp. 369-74 dans *International Encyclopedia of the Social Sciences*, vol. I. London & New York: Macmillan.
- FRANK, Lawrence R. (1939) *Projective Methods for the Study of Personality*, "Journal of Psychology" VIII, 2 (octobre): 389-413.
- FRANK, Phillip (1961) *Introduction*, dans P. Frank (ed.), *The Validation of Scientific Theories*. New York: Collier.
- FREY, Frederick F. (1970) *Cross-Cultural Survey Research in Political Science*, pp. 173-294 dans Robert T. Holt et John E. Turner (eds.), *The Methodology of Comparative Research*. New York: Free Press.
- GALTUNG, Johan (1959) *An Inquiry into the Concepts of 'Reliability', 'Intersubjectivity' and 'Constancy'*, "Inquiry" II, 2 (été): 107-25.
- GALTUNG, Johan (1967) *Theory and Methods of Social Research*. London: Allen & Unwin.
- GERGEN, Kenneth J. et BACK, Kurt W. (1966) *Communication in the Interview and the Disengaged Respondent*, "Public Opinion Quarterly" XXX, 3 (automne): 385-98.
- GIAMPAGLIA, Giuseppe (1986) *Alfa, omega e teta: è attendibile la misura dell'attendibilità?*, "Sociologia e Ricerca Sociale" VII, 21: 75-99.
- GORDEN, Raymond L. (1952) *Interaction between Attitude and the Definition of the Situation in the Expression of Opinion*, "American Sociological Review" XVII, 1 (février): 50-58.
- GOSTKOWSKI, Zygmunt (1974) *Toward Empirical Humanization of Mass Surveys*, "Quality and Quantity" VIII, 1 (mars): 11-26.
- GOSTKOWSKI, Zygmunt (ed., 1978) *Investigations on Survey Methodology*. Warszawa: PAN.
- GUTTMAN, Louis A. (1946) *The Test-Retest Reliability of Qualitative Data*, "Psychometrika" XI, 1: 81-95.
- GUTTMAN, Louis A. (1950) *The Basis for Scalogram Analysis*, pp. 60-90 dans Samuel Stouffer (ed.), *Measurement and Prediction*, Vol. IV. Princeton University Press.

- HARRE', Rom (1981) *Philosophical Aspects of the Macro-Micro Problem*, pp. 139-60 dans Karin D. Knorr-Cetina et Aron Victor Cicourel (eds.), *Advances in Social Theory and Methodology. Toward an Integration of Micro- and Macro-sociologies*. London: Routledge.
- HAYNES, D. P. (1964) *Item Order and Guttman Scales*, "American Journal of Sociology" LXIX, 1: 51-8.
- HEISE, David R. et BOHRNSTEDT, George W. (1970) *Validity, Invalidity, and Reliability*, pp. 104-129 dans Edgar F. Borgatta et George W. Bohrnstedt (eds.), *Sociological Methodology 1970*. S. Francisco: Jossey-Brass.
- HEMPEL, Carl Gustav (1952) *Fundamentals of Concept Formation in Empirical Science*. Chicago University Press.
- HEMPEL, Carl Gustav (1961) *Fundamentals of Taxonomy*, dans J. Zubin (ed.), *Field Studies in Mental Disorders*. New York: Grune & Stratton.
- HILL, Kim Quaile (1980) *Measurement Problems in Cross-National Analysis: Persisting Dilemmas and Alternative Strategies*, "Quality & Quantity" XIV, 3 (mai): 397-413.
- HORST, A. Paul (1934) *Item Analysis by the Method of Successive Residuals*, "Journal of Experimental Education" II: 254-63.
- HORST, A. Paul (1936) *Item Selection by Means of a Maximizing Function*, "Psychometrika" I, 4 (décembre): 229-44.
- HYMAN, Herbert H. (1972) *Secondary Analysis of Sample Surveys*. New York: Wiley.
- HYMAN, Herbert H. et al. (1954) *Interviewing in Social Research*. Chicago University Press.
- INGRAM, Elisabeth (1977) *Basic Concepts in Testing*, pp. 11-37 dans J. P. B. Allen et Alan Davis (eds.), *Testing and Experimental Methods*. London: Oxford University Press.
- JACOBSON, Alvin L. et LALU, N. M. (1974) *An Empirical and Algebraic Analysis of Alternative Techniques for Measuring Unobserved Variables*, pp. 215-42 dans Hubert M. Blalock (ed.), *Measurement in the Social Sciences. Theories and Strategies*. Chicago: Aldine.
- JAHODA, Marie, DEUTSCH, Morton et COOK, Stuart W. (1951) *Research Methods in Social Relations*. New York: Dryden.
- KAHN, Robert L. et CANNELL, Charles F. (1968) *Interviewing: Social Research*, pp. 149-161 in *International Encyclopedia of the Social Sciences*, vol. VIII. London & New York: Macmillan.
- KAPLAN, Abraham (1964) *The Conduct of Inquiry*. San Francisco: Chandler.
- KATONA, George (1951) *Psychological Analysis of Economic Behaviour*. New York: McGraw-Hill.
- KENDALL, Patricia (1954) *Conflict and Mood*. New York: Free Press.
- KENDALL, Patricia et LAZAZSFELD, Paul Felix (1950) *Problems of Survey Analysis*, pp. 133-96 dans Robert King Merton et Paul Felix Lazarsfeld (eds.), *Continuities in Social Research*. Glencoe: Free Press.
- KERLINGER, Fred N. (1965) *Foundations of Behavioral Research: Educational and Psychological Inquiry*. New York: Holt, Rinehart & Winston.

- KIRKPATRIK, Clifford (1936) *Assumptions and Methods in Attitude Measurements*, "American Sociological Review" I, 1 (février): 75-88.
- KUDER, G. Frederick et RICHARDSON, Marion W. (1937) *The Theory of the Estimation of Test Reliability*, in "Psychometrika" II, 3 (septembre): 151-60.
- LAZARSELD, Paul Felix (1958) *Evidence and Inference in Social Research*, "Daedalus" LXXXVII, 3 (automne): 99-130. Tr. fr. en P. F. Lazarsfeld et R. Boudon (eds.), *Méthodes de la sociologie*, vol. I. Paris: Mouton, 1965.
- LAZARSELD, Paul Felix (1961) *Notes on the History of Quantification in Sociology: Trends, Sources and Problems*, "Isis" LII, part 2 (juin): 277-333.
- LAZARSELD, Paul Felix (1966) *Concept Formation and Measurement in the Behavioral Sciences: Some Historical Observations*, dans Gordon J. Drenzo (ed.), *Concepts, Theories and Explanation in the Behavioral Sciences*. New York: Random.
- LAZARSELD, Paul Felix et BARTON, Allen H. (1951) *Qualitative Measurement in the Social Sciences: Classifications, Typologies, and Indices*, pp. 155-92 dans D. Lerner et H. D. Lasswell (eds.), *The Policy Sciences*. Stanford University Press.
- LENTZ, Theodore F., HIRSHTEIN, Bertha et FINCH, J. H. (1932) *Evaluation of Methods of Evaluating Test Items*, "Journal of Educational Psychology" XXIII: 344-50.
- LIKERT, Rensis (1932) *The Method of Constructing an Attitude Scale*, dans R. Likert, *A Technique for the Measurement of Attitudes*, "Archives of Psychology", monogr. n. 140: 44-53.
- LOCANDER, William, SUDMAN, Seymour et BRADBURN, Norman (1976) *An Investigation of Interview Method, Threat and Response Distorsion*, "Journal of the American Statistical Association" LXXI, n. 354 (juin): 269-75.
- LUMSDEN, J. (1976) *Test Theory*, "Annual Review of Psychology" XXVII: 251-80.
- LUTYNSKA, Krystyna (1978) *Ankieterzy i badacze. Z badan nad wplywem ankieterskim*, "Przeglad Socjologiczny" XXX: 143-73.
- LUTYNSKI, Jan (1979) *A Question as a Tool in Social Survey Research*, "Polish Sociological Bulletin", n. 3: 39-58.
- MACFARLANE, Jean W. (1942) *Problems of Validation Inherent in Projective Methods*, "American Journal of Orthopsychiatry" XII: 405-11.
- MACRAE, Duncan (1970) *Issues and Parties in Legislative Voting: Methods of Statistical Analysis*. New York: Harper & Row.
- MAY, Mark (1932) *Problems of Measuring Character and Personality*, "Journal of Social Psychology" III, 2 (mai): 131-43.
- McKENNELL, Aubrey C. (1970) *Attitude Measurement: Use of Coefficient Alpha with Cluster or Factor Analysis*, "Sociology" IV, 2 (mai): 227-45.

- McKENNELL, Aubrey C. (1973) *Surveying Attitude Structures: A Discussion of Principles and Procedures*, "Quality and Quantity" VII, 2 (décembre): 203-94.
- McKENNELL, Aubrey C. (1977) *Attitude Scale Construction*, pp. 183-219 dans Colm O' Muirheartaigh et Clive Payne (eds.), *The Analysis of Survey Data*. New York: Wiley, vol. I.
- McKENZIE, Donald A. (1981) *Statistics in Britain, 1865-1930. The Social Construction of Scientific Knowledge*. Edinburgh: University Press.
- McNEMAR, Quinn (1946) *Opinion-Attitude Methodology*, "Psychological Bulletin" XLIII, 4 (juillet): 289-374.
- MERRITT, Richard L. (1970) *Systematic Approaches to Comparative Politics*. Chicago: Rand-McNally.
- MERTON, Robert King (1948) *The Bearing of Empirical Research Upon the Development of Social Theory*, "American Sociological Review" XIII, 5 (octobre): 505-15.
- MORTON WILLIAMS, Jean (1979) *The Use of "Verbal Interaction Coding" for Evaluating a Questionnaire*, "Quality & Quantity" III, 1 (février): 59-75.
- MORTON-WILLIAMS, Jean et SYKES, Wendy (1984) *A Study of Question Failure through the Use of Interaction Coding*. London: Social and Community Planning Research.
- MOSIER, Charles I. (1947) *A Critical Examination of the Concept of Face Validity*, "Educational and Psychological Measurement" VII: 191-205.
- NAROLL, Raoul (1968) *Some Thoughts on Comparative Method in Cultural Anthropology*, pp. 236-277 dans Hubert M. Blalock et Ann B. Blalock (eds.), *Methodology in Social Research*. New York: McGraw-Hill.
- NICEFORO, Alfredo (1921) *Les indices numériques de la civilisation et du progrès*. Paris: Flammarion.
- NOELLE-NEUMANN, Elisabeth (1970) *Wanted: Rules for Wording Structured Questionnaires*, "Public Opinion Quarterly" XXXIV, 2 (été): 191-201.
- NOWAK, Stefan (1976) *Understanding and Prediction. Essays in the Methodology of Social and Behavioral Theories*. Dordrecht: Reidel.
- NUNNALLY, Jum C. (1978) *Psychometric Theory*. New York: McGraw-Hill.
- OPPENHEIM, A. N. (1966) *Questionnaire Design and Attitude Measurement*. New York: Basic Books.
- OSGOOD, Charles E. (1952) *The Nature and Measurement of Meaning*, "Psychological Bulletin" XLIX, 3 (mai): 197-237.
- PARRY, Hugh et CROSSLY, Helen M. (1950) *Validity of Responses to Survey Questions*, "Public Opinion Quarterly" XIV, 1 (printemps): 61-80.
- PAWSON, Ray (1980) *Empiricist Measurement Strategies: A Critique of the Multiple Indicator Approach to Measurement*, "Quality and Quantity" XIV, 5 (octobre): 651-78.
- PAWSON, Ray (1982) *Desperate Measures*, "British Journal of Sociology" XXXIII, 1 (mars): 35-63.
- PHILLIPS, Bernard S. (1966) *Social Research. Strategy and Tactics*. London: Macmillan.

- PILICHOWSKI, Andrzej et ROSTOCKI, Włodzimierz (1978) *Powtorny wywiad werifikacyjny jako metoda otrzymywania informacji o wartosci odpowiedzi na pytanie kwestionariuszowe*, "Przeglad Socjologiczny" XXX: 69-78.
- PINTO, Roger (1964) *Méthodes des sciences sociales*. Paris: Dalloz.
- POWERS, Edward A., GOUDY, Willis J. et KEITH, Pat (1978) *Congruence Between Panel and Recall Data in Longitudinal Research*, "Public Opinion Quarterly" XLII, 3 (automne): 380-9.
- PRZEWORSKI, Adam et TEUNE, Henry (1970) *The Logic of Comparative Social Inquiry*. New York: Wiley.
- PRZYBYŁOWSKA, Ilona et KISTELSKI, Krzysztof (1981) *The Social Context of Questionnaire Interview*. Lodz: Instit. Sozjologii Univ. Lodzki.
- QUETELET, L.-Adolphe-J. (1869) *Physique sociale, ou essai sur le développement des facultés de l'homme*. Bruxelles: C. Murquaedt.
- RAJU, Nambury (1977) *A Generalization of Coefficient Alpha*, "Psychometrika" XLII, 4 (décembre): 549-65.
- RAPOPORT, Anatol (1958) *Various Meanings of Theory*, "American Political Science Review" LII, 4 (décembre): 972-88.
- REYNOLDS, Paul Davidson (1971) *A Primer in Theory Construction*. Indianapolis: Bobbs-Merrill.
- RICHARDSON, Marion W. (1936) *Notes on the Rationale of Item Analysis*, "Psychometrika" I, 1 (mars): 69-76.
- ROSE, Arnold M. (1950) *Public Opinion Research Techniques Suggested by Sociological Theory*, "Public Opinion Quarterly" XIV, 2 (été): 205-14.
- ROSE, Arnold M. (1961) *Inconsistencies in Attitudes Toward Negro Housing*, "Social Problems" VIII (printemps): 266-82.
- ROZELLE, Richard et CAMPBELL, Donald T. (1969) *More Plausible Rival Hypotheses in the Cross-Lagged Panel Correlation Technique*, "Psychological Bulletin" LXXI (janvier): 74-80.
- RYLE, Gilbert (1938) *Categories*, pp. 189-206 dans AA. VV., *Proceedings of the Aristotelian Society*.
- SARGENT, Helen D. (1945) *Projective Methods: Their Origins, Theory and Application in Personality Research*, "Psychological Bulletin" XLII, 5 (mai): 257-93.
- SARIS, Willem E. (1981) *Different Questions, Different Variables?*, pp. 78-95 dans Claes Fornell (ed.), *A Second Generation of Multivariate Analysis*, Vol. II. New York: Praeger.
- SCHANCK, R. L. (1932) *A Study of a Community and Its Groups and Institutions Conceived as a Behaviour of Individuals*, "Psychological Monographs".
- SCHUMAN, Howard (1966) *The Random Probe. A Technique for Evaluating the Validity of Closed Questions*, "American Sociological Review" XXV, 1 (février): 3-25.
- SCOTT, William A. (1968) *Attitude Measurement*, pp. 204-273 dans Gardner Lindzey et Elliot Aronson (eds.), "Handbook of Social Psychology", Vol. II. Reading: Addison-Wesley, 2nd edition.

- SIEGEL, Paul M. et HODGE, Robert W. (1968) *A Causal Approach to the Study of Measurement Error*, pp. 28-59 dans Hubert M. Blalock et Ann B. Blalock (eds.), *Methodology in Social Research*. New York: McGraw-Hill.
- SINGER, J. David (1982) *Variables, Indicators, and Data*, "Social Science History" VI, 2 (printemps): 181-217.
- SLETTO, R. F. (1936) *Critical Study of the Criterion of Internal Consistency in Personality Scale Construction*, "American Sociological Review" I, 1: 61-8.
- SMELSER, Neil J. (1976) *Comparative Methods in the Social Sciences*. Englewood Cliffs: Prentice-Hall.
- SOUKUP, Miroslav et CHARVAT, Frantisek (1968) *Toward a Theory of Model Principles in Social Sciences: Introduction of the Index of Classification*, "Quality & Quantity" II, 1-2 (janvier): 44-62.
- SPEARMAN, Charles (1910) *Correlation Calculated from Faulty Data*, "British Journal of Psychology" III, (octobre): 271-295.
- SPECTOR, Paul E. (1981) *Research Design*. London & Beverly Hills: Sage.
- STEVENS, Stanley Smith (1951) *Mathematics, Measurement and Psychophysics*#, pp. 1-49 dans S. S. Stevens (ed.), *Handbook of Experimental Psychology*. New York: Wiley.
- SULLIVAN, John L. (1974) *Multiple Indicators: Some Criteria of Selection*, pp. 243-69 dans Hubert M. Blalock (ed.), *Measurement in the Social Sciences*. Chicago: Aldine.
- SULLIVAN, John L. et FELDMAN, Stanley (1979) *Multiple Indicators. An Introduction*. London: Sage.
- SZTABINSKI, Pawel B. (1978) *Metody kontroli pracy ankiet/w w badaniach z zastosowaniem wywiadu kwestionariuszowego*, "Przeglad Socjologiczny" XXX: 197-225.
- TESSLER, Mark A. (1973) *Problems of Measurement in Comparative Research: Perspectives from an African Survey*, "Social Science Information" XII, 4 (août): 29-43.
- TEUNE, Henry (1968) *Measurement in Comparative Research*, "Comparative Political Studies" I, 1 (avril): 123-38.
- THORNDIKE, Robert L. (1949) *Personnel Selection. Test Measurement Techniques*. New York: Wiley.
- THURSTONE, Louis Leon (1922) *The Stimulus-Response Fallacy*, "Psychological Review" XXX, 5.
- THURSTONE, Louis Leon (1928) *Attitudes Can Be Measured*, "American Journal of Sociology" XXXIII, 4 (janvier): 529-54.
- THURSTONE, Louis Leon et CHAVE, E. J. (1929) *The Measurement of Attitude*. University of Chicago Press.
- TURNER, Stephen P. (1979) *The Concept of Face Validity*, "Quality & Quantity" XIII, 1 (février): 85-90.
- UPSHAW, Harry S. (1968) *Attitude Measurement*, 60-111 dans H. M. Blalock et A. B. Blalock (eds.), *Methodology in Social Research*. New York: McGraw-Hill.

- VERBA, Sidney (1969) *The Uses of Survey Research in the Study of Comparative Politics: Issues and Strategies*, pp. 56-105 dans Stein Rokkan *et al.* (ed.), *Comparative Survey Analysis*. Paris: Mouton.
- VERBA, Sidney (1972) *Cross-National Survey Research: The Problem of Credibility*, pp. 309-356 dans Ivan Vallier (ed.), *Comparative Methods in Sociology: Essays on Trends and Applications*. Berkeley: University of California Press.
- VILLERME', Louis R. (1840) *Tableau de l'état physique et moral des ouvriers employés dans les manufactures du coton, de laine et de soie*. Paris: Renouard.
- WEBB, Eugene J., CAMPBELL, Donald T., SCHWARTZ, Richard D. et SECHREST, Lee (1966) *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally.
- WERTS, Charles E. et LINN, Robert L. (1970) *Cautions in Applying Various Procedures for Determining the Reliability and Validity of Multiple Item Scales*, in "American Sociological Review" XXXV,4 (août): 757-9.
- ZETTERBERG, Hans L. (1954) *On Theory and Verification in Sociology*. Totowa: Bedminster Press.
- ZUBIN, Joseph (1934) *The Method of Internal Consistency for Selecting Test Items*, "Journal of Educational Psychology" XXV: 345-56.