

Methodological and Statistical Issues in Psychological Measurement

Carlo Chiorri, PhD

University of Genoa, DiSA - Department of Anthropological Sciences
Psychology Unit
Corso Podestà, 2
16128 Genova – Italy
tel. 01020953726/7 - FAX: 01020953728
tel. (direct): 01020953709
E-mail: carlo.chiorri@unige.it

1. Basic Concepts

1.1 Measurement and Variables

Psychological measurement has ever been an intriguing challenge for psychologists and social scientists. The problem of measuring variables is common to all sciences, and each discipline has devised methods to cope with it. Psychology is not an exception, though the task is quite hard.

The usual definition for “variable” is *any characteristic that can assume multiple values or can vary in units of analysis*. This is a very wide and comprehensive definition that aims towards a more specific one, that is, *any factor which has the potential to influence another factor in a research study*. When dealing with humans, measurement takes place at different levels: physical (e.g., age, height, weight, etc.), behavioural (e.g. reaction time, number of aggressive behaviours, total sleep time, etc.) and, of course, psychological. Usually, the measurement of physical and behavioural variables does not pose serious difficulties, once we are provided with the proper measuring instruments. It is then possible to assign numbers to objects in order to operate on the former as we operated on the latter, that is, to find a link between the empirical and the numerical relational system. The same holds for the measurement of psychological variables, but the problem of the measuring instrument becomes more complex when we attempt to measure variables such as intelligence, sociability or introversion since we are not dealing with something ‘real’ in the sense of ‘manifest’, but rather with something which is ‘latent’, and thus not directly observable.

In social sciences many phenomena are not objects or events, therefore the problem of abstractness must be overcome showing that some operations for linking psychological objects to numbers *do exist* in terms of psychological units of measure. In other words, we need to state a relationship between a set of observable responses (*manifest variables*) to an underlying and not observable theoretical concept (*latent variable*, or *construct*). The keyword for this task is *operational*.

To operationalize means defining terms and concepts in ways that are precise, measurable and concrete and that anybody can observe and/or perform. At this stage, a fundamental role is played by theory, since the process of building a psychological measuring instrument begins with the definition of the construct. Say you want to devise a method for detecting tax dodgers. Of course, you do not rely upon their declaration that they have no income. So you must find a way to measure their income, which you do not know, by means of something you know or may know. At first, you may think that a person who has no income cannot afford to buy luxury goods: he or she simply has no money at all. Right or not, this is your definition of the concept ‘wealth’, and it stems from a theory. You may observe that such a person drives a shiny Ferrari, owns a yacht and wears a

diamond-studded Rolex watch. Some doubts about his or her *actual* income might arise. In other words, you are trying to measure the ‘wealth’ of that person, which is not observable, by means of some observable indicators (kind of car, owning a yacht, watch brand) that are conceptually related to that latent variable. The same happens when dealing with psychological variables. Say we intend to measure intelligence. The first step is defining *what is* intelligence. This is crucial, since all the observable indicators that we will use to measure the construct depend on such definition. For example: if we define ‘intelligence’ as ‘being good at maths’, then it is quite easy to measure it: we only need to know the person’s marks in mathematics at school, or we can administer a mathematical aptitude test. But if we define ‘intelligence’ as ‘capability of solving problems and fitting to the environment’, then the question becomes more complex. We will need several indicators to get a valid and reliable measure, and we will face the problem of *convergence*, that is, any new operational event of the construct that must bring towards a better measurement of that construct and exclude other possible explanations.

1.2 Reliability and Validity

The *reliability* of a measuring instrument can be defined as the *degree to which an instrument measures the same way each time it is used under the same condition with the same subjects and the testing procedure is free from errors of measurement*. In other words, it answers to the question: “Did we measure accurately?”. The classical theory of measurement defines a measure as the sum of the *true score* (T) plus an *error* (E):

$$X = T + E$$

Reliability of a measure is thus defined as the ratio between the true score and the measure, T/X , that is, the proportion of a measure that is not affected by the error measurement.

Basically, we have to face two kinds of error: *random error* and *bias*. Random error is any factor that can potentially affect a measurement of a phenomenon. This error is ubiquitous but **it is** usually small in size. It represents the oscillation of a number of measures around the true score. It has some properties that make it tractable:

- Average = 0
- Correlation between true score and error = 0
- Correlation between true score at time T_n and true score at time T_{n+1} = 0
- Correlation between errors of different measurements = 0

Random error can be caused, e.g., by careless recording of behaviour, variability in testing conditions, variability in participant behaviour due to other factors (e.g., alertness), etc. It makes individual scores less trustworthy (i.e., reduces reliability) and may obscure group differences.

Bias is more tricky, since it is a systematic and constant distortion of a measure. It may be of importance in size but, when detected, it can be fully eliminated. Say your clock is not working properly, so you are always late for your appointments. Once you know that your clock is ten minutes slow, you can correct for it. A bias can be caused by observers recording what they expect participants will do (the so-called *observer bias*), by researchers treating groups differently or by participants responding to demand characteristics or social desirability. As the random error, it reduces the validity of scores.

When we deal with the reliability of a psychological test, there are some statistical methods to assess it. The easiest one is re-administering the test after some time and figure out the correlation between the scores: the higher the correlation, the higher the reliability of the test. This

method is called *test-retest*, which is not very different from another method, *parallel/alternative forms*, in which different but equivalent forms of the same test are administered. These methods require at least two administrations, while methods like *split-half* and *internal consistency* need only one. In the former, the test is split into two equal parts and a correlation coefficient between the scores of the halves is obtained. The latter is instead a statistical index that tells the degree to which the items of the test measure the same construct.

The *validity* of a measure has a number of aspects. *Content validity* of a psychological test is the degree to which test items are a representative sample of the universe of behaviours under investigation. This means that we are not suppose to measure a construct with all its possible indicators. There may be thousands of them, but a valid measure can be accomplished choosing the most representative ones.

Face validity is the degree to which test items *seem* to measure what is intended. This is not a trivial issue. Participants will provide more reliable answers if they feel that the test they are administered with is measuring what they have been said, thus reducing measurement error. For this reason, *reliability* is another aspect of validity.

Construct validity is the degree to which a test measures precisely what it is intended to measure. If we want to measure depression, and *only* that, we must make sure that our instrument is not measuring something related or confused with depression. Then we will provide our research with test about *convergent* and *discriminate* validity. The former concerns the relationship between different measures of the same construct, the latter tests the absence of relationship between the construct and measures of other constructs. In other words, we must show that our test on depression is strongly related to at least another test which has been proven to measure depression, and weakly or not at all related to tests which measure something different, like anxiety.

Criterion validity is the degree to which an item is useful in predicting a real-world outcome. Psychological tests are not an end in themselves, since the scores they produce must have a practical significance. We define *predictive criterion validity* as the degree to which a measure accurately forecasts how a person will think, act, and feel in the future, while *concurrent criterion validity* is the degree to which a test yields the same results as other measures of the same phenomenon. For instance, an aptitude test which intends to select a restricted number of students that applied to enter a university course should have both concurrent and predictive validity, since it must select students that *are* actually good students and that *will probably have* a brilliant academic career.

Nomological validity is a less known aspect of validity that deals with the network of relationships of the construct under investigation with other constructs in the same theoretical framework. That is, the definition of the construct must be consistent with the underlying theory, which is based on some other constructs.

2. Levels of measurement

As it has been stated by S.S. Stevens, variables can be measured on four different scales:

- Nominal
- Ordinal
- Interval
- Ratio

Nominal Scale. Variables that are measured at a nominal level are said to be *categorical*. This means that their values are categories, different to each other. Not all scientists agree that a

nominal scale is actually a measurement scale even though we can assign numbers to each value of the variables, numbers are just tags, that is, they are not used to quantify, but only to differentiate. For example, the numbers on football players' shirts are just some kind of label that helps the referee to distinguish one player from another. The use of numbers for this purpose is purely arbitrary: we could use letters or player's name just as well. Thus, knowing that Totti wears the number 10 shirt and Buffon wears number 1 tells us only that Totti is different from Buffon, or, at the most, that Totti is a forward while Buffon is the goalkeeper. Nothing more can be said, such as, for instance, that Totti is 10 times better than Buffon, since numbers do not represent quantities of something. Likewise, if we label Male=1 and Female=2, this does not mean that females are better than males because $2 > 1$, or that males are more important than females because 1 means that they are in first position. Statistical operations that are available at this level of measurement are examining if a datum is equal/different to some particular value or counting the number of occurrences of each value (frequencies).

Ordinal Scale. When data are measured at an ordinal level it is possible to say if a datum is less than or greater than another value. In other words, we can *rank* data on the basis of some criterion. Say you are seeing athletes on a podium. All the information conveyed by the image is that the athlete on the highest step has run faster, or jumped higher, or thrown a javelin further than the other two. The athlete on the second step has performed worse than the athlete on the first but better than the athlete on the third step. Even though it is possible to rank athletes, we cannot say *how much* the winner is better than the athlete that has come in second position. The gap may be negligible or very wide, but we do not know. In other words, when variables are measured on an ordinal level, we can not quantify differences between two ordinal values. A broadly used ordinal scale in psychology is the Likert scale, i.e., a rating of preference, or frequency, or agreement. Therefore, in a psychological test a question may be put about how much he or she has felt depressed during the last seven days, or how much he or she agree with a given statement, or how often he or she uses the car. In any case, the participant has to indicate on which point of the scale he or she reckons to be. For example, there may be a question on how much he or she feels satisfied with his or her work from 1 to 5 where 1 = not at all and 5 = very much. Values 2, 3 and 4 indicate intermediate levels of satisfaction. Strictly speaking, we can not quantify the difference in satisfaction between a subject that answered 5 and an other subject that answered 4. All that we can say is that the first subject is more satisfied than the second. Nevertheless, this kind of scale is often treated as if it were an interval scale, i.e., a scale where the difference between two consecutive steps is exactly defined. This is the case of Thurstone's *equally appearing intervals*. It is supposed that when the participant is asked to rate his or her satisfaction about his or her work on a scale from 1 to 5, he or she will divide the so-considered psychological dimension into four equal spaced parts. The outcome is that the difference between two job satisfaction scores like 4 and 3 is the same that separates 2 from 1.

Interval Scale. An interval scale allows to quantify the difference between two interval scale values, even though there is no natural zero. For instance, temperature scales provide interval data with 10°C warmer than 5°C and such 5°C difference has the same physical meaning as the difference between 200°C and 195° or 3°C and -2°C . Anyway, one must not fall into the temptation to state that a day with a temperature of 10°C is twice as hot as a day with a temperature of 5°C , since there is no absolute zero. *Absolute zero* means that a zero score or value stands for "absence of the quality or the quantity being measured". In the Celsius scale, zero is the temperature at which water freezes into ice. If we used a Fahrenheit scale, this value would be 32°F , while 10°C and 5°C corresponding to 50°F and 41°F , respectively. That is, the same temperature, under another measuring scale, is not twice as high as the other. This is true only if temperatures are measured on the Kelvin scale, where zero actually stands for "absence of heat". Psychological tests are the most

obvious examples of interval scale: scoring 0 on an intelligence test does not mean ‘absence of intelligence’.

Ratio Scale. This is the highest level of measurement. It allows to take ratios among ratio scaled variables. Typical ratio scale variables are those that are measured by count, e.g., number of children, number of anti-social behaviours, number of errors, etc. It is now meaningful to say that 10 m is twice as long as 5 m. This ratio hold true regardless of which scale the object is being measured in (e.g. metres or yards), since zero, in any case, means “absence of length”.

Knowing the scale of measurement of variables is essential for applying the proper statistical test. For example, an ANOVA test, which requires at least an interval scale measurement for the dependent variable, can not be applied to a nominal dependent variable. Nevertheless, it is sometimes possible to perform scale transformations. Let’s take into account the variable “Education”. If we know the kind of diploma held by a subject, this is basically a nominal scale level of measurement, since the different values of the variable may be labelled as ‘None’, ‘Primary’, ‘Secondary’, ‘High School’, ‘Degree’ or ‘PhD’. One may argue that this variable can be placed in a specific order, since PhD is a higher diploma than Degree, which, in turn, is higher than High School and so forth. We could label the six different steps from 1 = None to 6 = PhD. Moreover, if we would define ‘Education’ in terms of ‘completed academic years’, then we could re-code the variable into a ratio scale variable, since ‘None’ means ‘zero completed academic years’, while for a PhD diploma, certainly in Italy, 21 completed academic years are needed.

A related issue is that of qualitative or quantitative measurement. Say you want to measure a person’s aggressiveness. Subjects may be categorized as ‘aggressive’ or ‘not aggressive’, that is, on a nominal scale. If we wanted to be more accurate, we could lay down criteria to define a person as ‘highly’, ‘moderately’ or ‘not’ aggressive, thus introducing a ranking, i.e., an ordinal scale. If we were provided with a psychological test measuring aggressiveness we would climb up to an interval scale, while counting the number of aggressive behaviours in a given time window would yield ratio scale scores. In any case, the answer to the question ‘Is a certain variable quantitative or qualitative?’ is not always just one.

3. Factor Analysis as a tool for developing measurement models

A statistical model is an abstract or theoretical representation of a phenomenon. Complex statistical procedures allow to test if the model is consistent or not with observed data, that is, if the model is a plausible explanation of the real world. It must be noted that a statistical outcome consistent with the model does not tell us that the model shows the way in which nature behaves: it just tells that things may go on in this manner.

One of the most popular statistical procedures employed in the development of measurement models is *Factor Analysis*. Factor Analysis is a general term that indicates all those techniques which aim to reduce the number of observed variables to a smaller set of variables called ‘factors’ and to detect structures in the relationship between the observed variables (i.e., *classify* variables). The assumptions of Factor Analysis are that (a) observed (*manifest*) variables are linear combinations of some ‘source’ variables called ‘factor’ or ‘latent variables’ or ‘constructs’; (b) the correlations between the observed variables can be accounted for by a smaller number of theoretical variables, thus matching the *parsimony criterion*; (c) the desired solution is that in which all observed variables have factor complexity equal to 1 (*simplicity criterion*). The term ‘factor complexity of a manifest variable’ indicates the number of factors the variable is substantively related to (see fig. 1).

[Insert fig. 1 about here]

Lets take an example, say we have observed the 13 variables in the boxes in fig. 2. They are supposed to be of some kind of intelligence.

[Insert fig. 2 about here]

If we had to draw an intelligence profile of a person, we could not report thirteen raw scores, since it would not be interpretable. We would need to report a *smaller* amount of information *without* losing the information we are interested in. If we performed a factor analysis, we would find a factor structure like the one depicted in fig. 2, where only four dimensions are considered: Perceptual Organization, Verbal Comprehension, Working Memory and Symbol Comprehension. If we were still not satisfied, yet, we could perform a further Factor Analysis on the scores of these four factors, thus obtaining a further reduction in the number of underlying dimensions.

The relationship that links the factor to the observed variable is supposed to be linear. “Why linear?”, you may ask. Well, linearity is the simplest kind of relationship we can think of, and in many cases, it has proven to be consistent with the real world. The equation is:

$$X = b \cdot F + d \cdot u$$

Where X = manifest variable score, b = standardized regression coefficient, F = score on the common factor, u = score on the specific (unique) factor and d = regression coefficient of the specific factor on X . A *path model* is thus defined as a model that specifies how common and specific factors produce the observed score on a manifest variable (fig. 3).

[Insert fig. 3 about here]

It is important to note that all information available is the correlation of the manifest variables and their variances and that unique factors do not affect the covariance between observed variables, which is assumed to be the outcome of the sole common factor influence.

In fig. 3 b_1 and b_2 are standardized regression coefficients, that is, correlation coefficients between the observed variables and the factors. In other words, they represent the importance of a variable in determining the factor. In Factor Analysis they are called *factor loadings* and the statistical procedure is aimed to determine their values.

A Factor Analysis always begins with the calculation of the observed variable *correlation or covariance matrix*. Then the *factor extraction* process takes place and, after *rotation*, factors are *interpreted*. Let's see these steps in a little detail.

Extraction can be performed in a number of ways, since several algorithms have been developed. One of the oldest ones is the *Centroid* method, while *Maximum Likelihood* and *Least Squares* are nowadays among the most widely used. *Alpha Factoring* and *Image Factoring* are less popular, while *Principal Component Analysis* (PCA) deserves a qualification. While all the other methods aim to extract from the correlation matrix all the *common* variance, that is, that portion of variance which is accounted for by factors, PCA extracts *all* the variance, regardless of whether it is common or not. Take a look to fig. 4. The relationship between the two variables can be described by two dimensions: some kind of width (full line) and some kind of height (dashed line).

[Insert fig. 4 about here]

When applied to n variables, PCA always describes the covariance pattern in terms of n dimensions. At first sight, there is no gain in parsimony, but not all dimensions extract a substantial amount of

variance. The first component removes the highest amount of variance it is capable of, the second one removes the highest amount of variance still to be accounted for, and so on. In fig. 4 the greater importance of the 'width' dimension is apparent. Of course, not all components will remove a substantial amount of variance, for which only $p < n$ components will be held in the factor structure. PCA can be a useful tool for detecting the dimensionality underlying a pattern of covariances, but it must be noted that it is something different from a proper Factor Analysis. Moreover, it does not make much sense applying a rotation procedure to a PCA solution (see below).

Table 1 depicts a Factor Loading Matrix taken from a study on the *Multidimensional Locus of Control Scale*, form C, by Wallston.

[Table 1 about here]

Factor loadings in bold are those considered *substantive*, that means, higher than .30. This is an arbitrary cut-off value (some scientists use .40) for considering a variable a good indicator of the factor. As it can be seen from table 1, item 1 has a substantive correlation with factor *Internal* and negligible relationships with the other three factors. Likewise, item 2 is a good indicator of the factor *Chance* and so on.

If we sum the squared factor loadings for each column, we obtain what is called an *eigen-value*, that is, a value that, when divided by the number of variables, tells the proportion of variance accounted for by a factor. On the other hand, if we sum the squared factor loadings for each row, we obtain what is called *communality*, that is, the proportion of variance of a variable accounted for by common factors.

Eigenvalues and communalities are used to determine how many factors must be held in the model, since the procedure extracts all those possible. Following the *communality criterion*, there are as many factors that must be extracted in order to reach a certain value of communality for all variables (e.g., .60). Far more used are those criteria following *eigen-value criteria*, like the Kaiser-Guttman's method (hold all factors with eigen-values greater than 1) or fixing the proportion of variance accounted for by all the factors extracted (e.g., at least 40%).

Once we have determined the smallest number of latent dimensions, i.e., factors, that adequately accounts for the observed correlations, the next step is *interpreting* the psychological meaning of factors. This is not trivial, since all factorial solutions provide latent dimensions that are orthogonal, i.e. independent, each other and in decreasing order of importance, i.e., proportion of variance accounted for. Such properties are not intrinsic to data structure, but they are arbitrary constrains applied to data in order to find unique solutions for the statistical procedure. This implies that in an un-rotated factor solution factor complexity of indicators will be (almost) always greater than 1, thus violating the simplicity criterion (see above).

Rotation is a statistical procedure that removes such a problem without modifying the number of factors or communalities or percentage of variance accounted for. To put it in simpler words, it just changes the point of view on factors in order to see them more clearly. Take a look to fig. 5.

[Insert fig. 5 about here]

Fig. 5a represents graphically an un-rotated factor solution with two factors (I and II, since factors, unless they are provided with a label, are indicated with Roman numerals). The position of each point in the Cartesian plane indicates the importance of each item in measuring each factor. Since coordinates are factor loadings, the higher is the factor loading, the more important is the item in measuring the factor. The simplicity criterion requires that an item should be an indicator of only one factor, that is, each item should have a substantive factor loading on only one factor. In the graph, we should observe a cloud of points around the II axis (thus showing high factor loadings on

the II factor and negligible factor loadings on the I factor) and another cloud around the I axis (i.e., nearly zero factor loadings on the II factor and high factor loadings on the I factor). Actually, in fig. 5 we observe two clouds of points, though not placed around factor axes. But what happens if we rotate the axes by a given angle? Fig. 5b displays a new Cartesian plane defined by axes I' and II' that meets the simplicity criterion. Note that the number of factors or communalities or proportion of variance accounted for remain unchanged. The table beside the graph shows the *rotated factor solution* with factor loadings on the new factors, I' and II' (provided numbers in both tables in fig. 5 are fictitious).

The above performed rotation preserved the orthogonal factors, that is, the two factors remained independent of each other even after the rotation procedure. If we interpret factors as psychological dimensions, we should conclude that dimension I and dimension II are two psychological constructs not related. Whatever the psychological field of research may be, it is often hard to sustain such hypothesis from a theoretical point of view. Consider fig. 6.

[Insert fig. 6 about here]

The u-rotated solution provided two factors, but the position of the two clouds in the Cartesian plane is inconsistent with a rotation of the axes that preserves the orthogonal (fig. 6a). Beyond rotating axes, changing the angle between them is also necessary to meet the simplicity criterion. Fig. 6b shows a rotated factor solution which is the outcome of an *oblique* rotation. Oblique rotations imply stating a relationship between factors, which, most of the time, is more consistent with the later theoretical interpretation.

What has been explained so far is the usual procedure when we do not know how many factors may account for the observed correlations between manifest variables, which are the relationships between factors and which are the relationships between factors and variables. To say in one word, it is an *Exploratory Factor Analysis*. If, on the basis of theory or literature or past experience, we knew something about the number of factors or their relationships, we could proceed in a different way, that is, testing if the factor structure we hypothesize fits the observed data. This is called *Confirmatory Factor Analysis*.

To perform a Confirmatory Factor Analysis we need, at least, five steps:

1. Specify the model
2. Identify the model
3. Parameter estimation
4. Goodness of fit evaluation
5. Improve the model

Specifying the model implies defining constraints on pairs of correlated factors and on relationships between factors and observed variables. This is a purely theoretical task, since each arrow put in the model must be theoretically meaningful (see fig. 7).

[Insert fig. 7 about here]

Next step is to verify if the model we specified is *identified*, that is, if, on the basis of observed covariances, there exists a unique solution for the parameter estimates. Parameters are factor loadings and factor correlations that correspond to each arrow in the model. The statistical procedure is aimed to determine the value of each parameter, but, if the model does not meet certain constraints, a unique solution, that is, one and only one value for each parameter, cannot be found. This can be tested algebraically (very complex) or with some indexes usually provided by software. If the model is identified, *parameter estimates* are provided and *goodness of fit of the model* can be evaluated. Goodness of fit is tested by a number of indexes that tell how much the correlation

between observed variables predicted by the model and those actually observed overlap. Normally, predicted values are not exactly equal to observed ones, since too many parameters would be needed to obtain this result and the aim of the model is *to simplify*. In other words, the specified model must reach the highest fit to the observed data using the least number of parameters. If the discrepancy between the predicted and the observed data is negligible, according to goodness of fit indexes, we can conclude that the model is a good representation of the phenomenon under investigation, otherwise we will have to specify another model re-thinking to theoretical assumptions.

An example: in a well-know study published in 1977, Weathon and co-workers hypothesized that Alienation could be a factor determining high scores on Anomia and Helplessness tests (observed variables). Variables were measured twice, one in 1967 and one in 1971 and a relationship between scores on Alienation in 1967 and in 1971 had been hypothesized (fig. 8a).

[Insert fig. 8 about here]

Besides, they used the number of education years and Duncan's Occupational Status Index to measure Socio-Economic Status (fig. 8b). We have therefore two measurement models and we can define linear equations that specify the relationships in each model (fig. 8c and d). By applying the procedure described above, we can test if both models are adequate theoretical representations of the relationships between observed variables by using Confirmatory Factor Analysis.

But we can go further. Confirmatory Factor Analysis Models aim to explain relationships between a set of observed variables by means of a smaller set of unobserved (latent) variables, but they do not tell anything about the possible *causal* relationships between latent variables. They just test if they are *related*. *Structural Equation Modelling* is a covariance structure modelling technique that allows to test causal relationships between latent variables. The steps needed to perform it are the same as those for Confirmatory Factor Analysis. The model is made up of three parts: two measurement models and a structural model. The measurement models are called *exogenous* and *endogenous*. Exogenous variables are those which are determined or set outside the model, and, usually, they are not fully psychological. These are contrasted by the endogenous variables which are determined inside the model and are properly psychological. The value of the endogenous variables will change when the exogenous variables change, that is, when we consider exogenous variables as independent variables and endogenous variables as dependent variables. If, in the example, we would consider Socio-Economic Status as the independent variable and the values of Alienation in 1967 and 1971 as the dependent variables, we would define the *structural model*, that is, the pattern of causal relationships between latent variables (fig. 9).

[Insert fig. 9 about here]

The procedure to test the goodness of fit of a Structural Equation Model is the same as for Confirmatory Factor Analysis. If we obtain a good fit, we can consider the model as a good and likely representation of what is going on. Moreover, it supports our hypotheses about causal relationships between variables.

To conclude, it must be noted that what we actually observe is always *co-variation* between variables, *not causality*. Causality can be hypothesized on the basis of our knowledge about the phenomenon under investigation and can be tested statistically. However, a significant statistical result tells only that our theory is plausible, and not that it is *true*. The model used by Wheaton and co-workers proved to be a good model, but things are not *necessarily* continuing that way. Never ever forget it.

Suggested readings

- Carmines, E.G., & Zeller, R.A. (1979). *Reliability and Validity Assessment*. (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-017). Newbury Park, CA: Sage.
- Comrey, A.L., & Lee, H.B. (1992). *A First Course in Factor Analysis*. Laurence Erlbaum Associates.
- Kim, J.-O., & Mueller, C.W. (1978). *Factor analysis: statistical methods and practical issues* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-014). Newbury Park, CA: Sage.
- Kim, J.-O., & Mueller, C.W. (1978). *Introduction to factor analysis* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-013). Newbury Park, CA: Sage.
- Long., J.S. (1983). *Confirmatory factor analysis* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-033). Newbury Park, CA: Sage.
- Long., J.S. (1983). *Covariance structure models* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-034). Newbury Park, CA: Sage.

Acknowledgements

I would like to thank prof. Riccardo Luccio and dr. Grazia Micale for reviewing the drafts.