Measuring stability and change in subjective quality of life. General and specific issues concerning testing reliability of measurement

Applicazioni STatistiche nella RIcerca Sociale - 3

Filomena Maggino

Measuring stability and change in subjective quality of life

General and specific issues concerning testing reliability of measurement

Università degli Studi di Firenze

Part of this work was presented at

5th Conference of the International Society for Quality-of-Life Studies

Session

Continuity and Change of Quality of Life

organized by

Working Group Social Indicators of the International Sociological Association

chaired by

Kenneth C. Land – Liz Eckermann

“Johann Wolfgang Goethe-University”

Frankfurt/Main, Germany,

July, 20-24, 2003

Abstract

In order to study dynamics in quality of life, in terms of individual stability and change, a set of methodological issues has to be encountered, considered and managed. Particularly, studies on change require accomplishment of four fundamental aims, strictly connected to each other: conceptualization, design, measurement, and data analysis. Among these, measurement issue, considered in terms of theoretical definition and coherent operational solution, has to be considered, in our opinion, crucial. Consequently, the assessment of measurement of change requires a careful and systematic consideration.

The paper attempts to review the classical literature on methodological features in measurement of subjective change, in terms of theoretical definition, and in reliability analysis approaches making reference to specific aspects and problems involved in studies concerning subjective change in quality of life.

Table of contents

Introduction

In order to study dynamics in quality of life, in terms of individual stability and change, a set of methodological issues has to be encountered, considered and managed. Particularly, studies on change require accomplishment of four fundamental aims, strictly connected to each other: conceptualization, design, measurement, and data analysis. Among these, measurement issue, considered in terms of theoretical definition and coherent operational solution, has to be considered, in our opinion, crucial. Consequently, the assessment of measurement of change requires a careful and systematic consideration.

The paper attempts to review the classical literature on methodological features in modeling, measurement, and analysis of change, making reference to specific aspect and problems involved in studies concerning subjective change in quality of life.

The first part of the paper, with no pretension to be exhaustive, provides with a reference frame as regards to general concerns and issues involved in studies on change in subjective quality of life.

The second part concerns definition and assessment of reliability of measurement over time with particular attention to classic and structural models approaches involved in assessment of reliability of change, with reference to their possible applications to subjective measurement of change in quality of life studies.

1. STUDY OF CHANGE IN SUBJECTIVE QUALITY OF LIFE: GENERAL CONCERNS

Generally, studies on individual change and stability may have different objectives, also adopted in subjective quality of life studies, such as (Goldstein, 1979; Menard, 1991):

- to describe detailed patterns of individual change,
- to predict the values of measurements at a later time from those obtained at earlier times,
- to obtain insight into underlying causal processes (Engel and Reinecke, 1996).

In order to attain the goals, study of dynamics in quality of life, as well as in other fields, requests a coherent methodological approach, including a solid theory of measurement, especially when the study of change concerns subjective dimensions. In particular, it requires the individuation of:

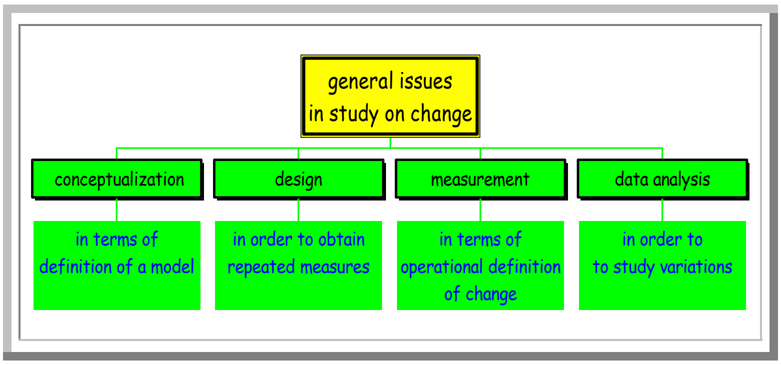
- modifiable dimensions (dimensions that may significantly change)
- adequate measures (in terms of reliability and validity)
- proper analysis approaches (in terms of methods and techniques).

In this perspective, studies on change show fundamentally four validity problems (Campbell, 1963; Visser, 1985):

a. *internal validity*, concerning the possibility to come to correct conclusion within a particular study (basic logic of the whole study),
b. *external validity*, concerning the possibility to make correct generalizations (definition of sampling design),
c. *construct validity*, concerning the representativeness of observation as regards what was supposed in theory (definition of measurement procedure),
d. *statistical conclusion validity*, concerning correct treatment of data in order to study variations in the observations (definition of data analysis procedure).

Such items correspond to the four crucial points that have to be defined and handled in studies based on change (Bryk and Raudenbush, 1992), as summarized in figure 1.

Fig. 1 Crucial issues in study on change



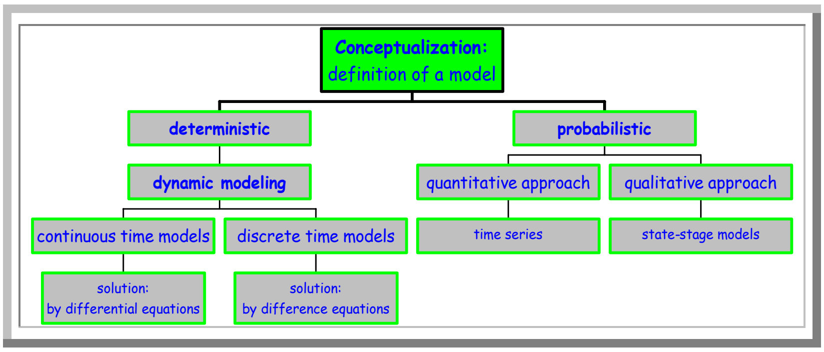
1.1 CONCEPTUALIZATION: THE DEFINITION OF A MODEL

Investigation of phenomena under the perspective of change requests the design of a model, in order to describe a change. Different kinds of models were generated in different fields, like economy, behavioral sciences, and so on.

As a rule, when the description is based on mathematical definitions (Menard, 1991; Brown, 1995), the models can be (figure 2):

- *deterministic*, when the change is conceptualized by fixed patterns or laws, defining whether, how and how much, according to some hypotheses, a variable changes as a response to change in other variables (for individuals or groups change models); the statistical approach is instrumental; deterministic approaches:
  - lead to fractions or events,
  - are more flexible in nonlinear specification allowing a more realistic understanding of the complexity of social dynamics,
  - have a series of statistical measures for evaluation of the model (measures of fit, tests of the significance of the parameters, measures of relative importance of the parameters, etc.).
- *probabilistic or statistical,* when the goal is to predict with some accuracy, at group level, the proportion/percentage of cases that will change, the proportion/percentage of cases that will change in a certain direction, the average amount by which they will change; the underlying assumption allowing the probabilistic description is that there is some influences on a process defined, at individual level, probabilistic; probabilistic approaches:
  - yield probabilities of complete events,
  - are based upon standard statistical methods, such as ordinary least squares (OLS)[1], and estimation, such as maximum likelihood approach,
  - allow including measurement error in the model.

Fig. 2 Characteristics of models of change.



Both deterministic and probabilistic models can be distinguished as regards to time, considered in terms of continuous or discrete definition:

a) deterministic continuous time models,
b) deterministic discrete models,
c) probabilistic continuous time models (which the process requires the assignment of a probability of an event in an infinitesimal period of time),
d) probabilistic discrete time models (in which the process, generally defined by generational occurrences, requires measurements separated by significant lengths of time).

The choice between continuous and discrete approaches depends upon measurement opportunities. Generally, and not only in social and quality of life studies, the measurement of phenomena are taken at discrete points of time, more or less regular, regardless of the dynamic structure (Brown, 1995).

It is also possible to identify models with randomly fluctuating parameters, applicable in situation in which parameters are functions of other variables containing stochastic components (Brown, 1995).

[1] Notice that also *ordinary least squares* method is based on a deterministic formulation for an equation defining a line (Brown, 1995).

1.1.1 Deterministic models

Deterministic models, defined in terms of quantitative change, express the values of the changing variable as functions of time. The mathematical formula/equation can be defined according to two goals (Menard, 1991):

- *description of change*: the formula includes only variable and time
- *explanation of change*: the equation involves the introduction of other variables.

The most useful approach to represent processes of change in phenomenon over time, considering time an explanatory dimension, is dynamic modeling, whose main objective is 'to understand, substantively, the mechanisms that are generating change in some observable phenomenon, and then to translate this set of ideas into mathematical language' (Huckfeldt et al., 1982). In other words, the objective is to make an adequate synthesis of observed measures in terms of a small number of parameters.

The application of dynamic modeling is based mainly upon *functional/differential models* (Visser, 1985; Menard, 1991; Huckfeldt et al., 1982) used principally as theoretical and not analytical tools; they help in consider the possible relations between unobservable variables. The difficulty in specifying a process in detail, as requested, and the treatment of time variable as a continuous variable represent one of the problems not allowing the practical use of these models; that is because applications in social sciences are essentially theoretical. Difference equation models, considering time as a discrete variable, may represent a solution to this problem.

Subsequently, according to the two different treatments of time dimension (central in dynamic modeling) we can distinguish between two different approaches.

- **Deterministic models for continuous time treatment**.
  These models define growth dynamic processes, assuming infinitely small time units and instantaneous rates of change. They are mathematically expressed in terms *differential equations* (Visser, 1985; Menard, 1991; Huckfeldt et al., 1982). An example of a deterministic model expressed in differential equations form is *the internal-influence diffusion model*. It concentrates interest on the study of innovation diffusion and expresses this as a function of time (cumulative numbers of adopters of an innovation at a given point of time):

$$\frac{dX}{dt} = ct^n$$

  where

X	cumulative number of cases that have adopted an innovation
t	time measured in appropriate unit
n,c	constant parameters (to be estimated)
dX/dt	rate of change in X

  The equivalent integrated equation is

$$X = \frac{ct^{n+1}}{n+1}$$

  When $n = 0$, X is expressed as a linear function of time and the equation becomes $X = ct$ (c may be estimated by regression techniques).
  This approach may be applied on a relatively large number of cases and a relatively small number of periods of time.

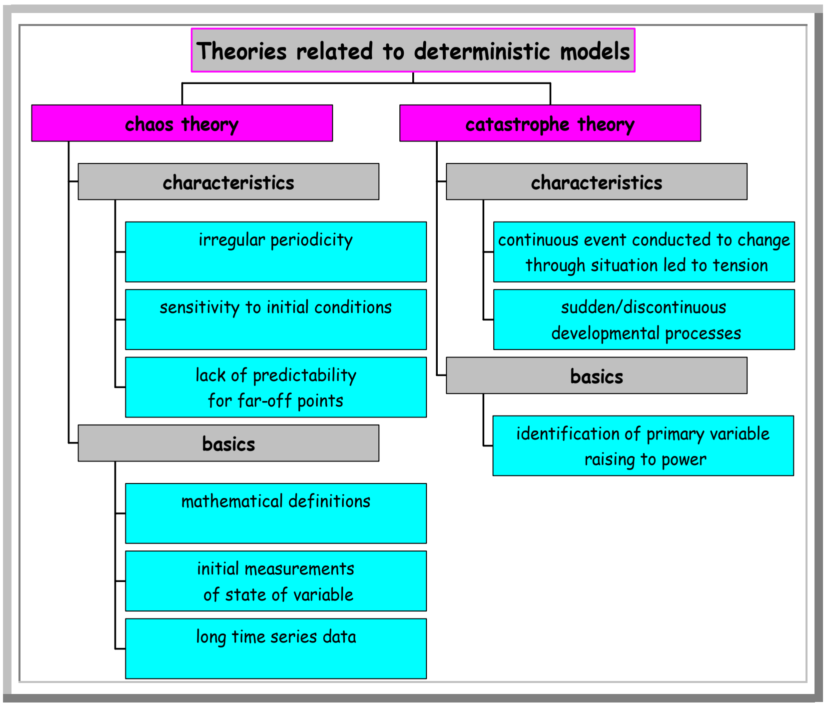
- **Deterministic models for discrete time treatment**.
  In these models the period of time between successive events is usually fixed according to calendar, simplifying the identification of time. They are mathematically expressed in terms of *difference equations*, (Huckfeldt et al., 1982) for which time is treated as a series of discrete, equally spaced units. This approach can define model as a combination of single/interdependent and linear/nonlinear equations. It seems to be more applicable in many quality of life studies since allows management of:

a. observations occurring in discrete time,
b. great diversity patterns of qualitative behavior,
c. models without requiring high level of mathematical formalizations.

In certain cases, it is possible to assume that observed discrete events are manifestations of an underlying continuous process or that the continuous process manifests itself at discrete points of time. However such approaches are applicable (Huckfeldt et al., 1982) only to change produced by a fixed dynamic structure (synchronic change). Treatment of process of moving from one structure into another (diachronic change) requires different approaches.

Among theories supporting deterministic models, two deserve a citation, for the possible applications that they may have in quality of life studies (figure 3).

Fig. 3 Characteristics and basics of theories related to deterministic models



- <u>Chaos theory</u> (Brown, 1995). Chaos is an irregular oscillatory process that can be observed in human behavior as well. Many behaviors that are repeated can be described in terms of oscillatory process (long and short cycles) both in individual (daily cycles of everyday life) and group (cyclic events like periodic vote) processes. Not always the repetition happens in regular fashion with regard to time. Consequences and effects of irregularities can be considered positive (as in some characteristics of creativity) or negative (irregular rest or food consumption cycles).

  Fundamental characteristics identifying a chaotic system are:
  a. irregular periodicity (absence of a repeated pattern),
  b. sensitivity to initial conditions (small changes in the initial conditions in a chaotic system produce dramatically different evolutionary processes),
  c. lack of predictability (sensitivity to initial conditions makes chaotic system unpredictable; however prediction is possible when the interest is concentrated on the movement between two relatively close points on a trajectory).

  Requirements of chaos modeling are
  - dimensionality (minimum of three independent variables in continuous time: differential

equations; one independent variable in discrete time: difference equations[2]);
- nonlinear models (at least one term in one equation must be nonlinear).

One of the approaches to *estimating* nonlinear interdependent systems potentially chaotic is *nonlinear least squares*.

The manageability of chaotic system, presenting these characteristics, relies
- on numeric calculations,
- initial measurements of the state of variable,
- long, and sufficiently close, time series data.[3]

These make difficult to deal with chaotic processes; in other words, chaotic systems are highly susceptible to combination of computational and measurement errors of original data.

The apparently restrictive requirements do not imply that chaos is rare in real world, on the contrary especially in social system.

➢ *Catastrophe theory* (Brown, 1995)

At the basis of catastrophe theory is the concept of bifurcation, which is an event that occurs in the evolution of a dynamic system in which the characteristic behavior is transformed (Brown, 1995). Models based on catastrophe theory allow describing discontinuous phenomenon controlled by continuous variables through graphs in *k*-dimensional spaces.[4] They are limited in treatment of processes characterized by sudden change or discontinuous development; they describe a process that can be conducted to a generalized change through a situation led to tension. This requires identification of a primary variable raising to some power. One of the approaches to *estimating* catastrophe processes is *nonlinear least squares* method as well. Different measures of fit are calculated for one-case many time points data and for many-case few time points, also two points (Brown, 1995).

In order to develop a significant catastrophe model social theory perspective is greatly usable; algebraic forms are only helping instruments in building models. (Brown, 1995)

Also catastrophe is not rare in real world even if data applications are rarely found. One of the most well-known applications of this theory in studying social processes concerned prison riots (Visser, 1985).

Applications of both theories present problems in dealing with errors of measurement of original data.

1.1.2 Probabilistic models

Probabilistic models relate directly measures made at one occasion to those made at successive occasion. In many case this approach assume causal relations between measurements. The adoption of probabilistic models requires implicit recognition of direction of time (measurements at one occasion are dependent on measurements at earlier occasions). Pre-post models can be considered to belong to these models (Visser, 1985).

Figure 4 summarized the main characteristics of the two probabilistic model approaches: quantitative and qualitative.

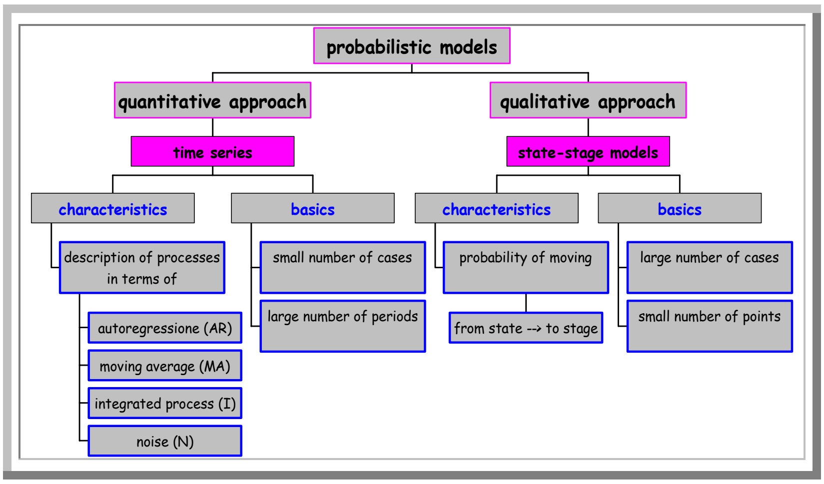
---

[2] The term 'map' in place of 'function' is usually applied with reference to difference equations that associate paired data for discrete time. General form for these functions is logistic.

[3] The concept of close points of time in a data series is relative to the time variability of phenomena under study, especially in social research.

[4] The study of qualitative differences between classes of graphs connected by continuous transformations is part of topology.

Fig. 4 Characteristics and basics of probabilistic models



Quantitative approaches

Typical quantitative descriptive approach for probabilistic models is *time-series approach*, allowing description of small numbers of cases (typical one) on a large number of periods of time. The description (Visser, 1985) of time-ordered data is carried out in order to distinguish four processes:

a. noise/random process (probabilistic component, present in all stochastic models),
b. autoregressive (*AR*) process (present values of a variable depending on past values at some specific lag/s or interval/s),
c. moving average (*MA*) process (past value of the noise influencing present values)
d. integrated (*I*) process (measurable trend over time in the values of the modeled variable, but in which there is no trend in the series detectable by subtracting values of the variable from values of the variable at later time).

Models, quantitatively expressed, always involve noise component and may incorporate one, two or all three other components; the objective is to define the change of the variable over time in terms of a stationary time-series.

Qualitative approaches

Typical probabilistic change models defined in qualitative terms (categorized classifications) are *stage-state models* (or *dynamic typologies*) in which movements of subjects from one category to another over time are modeled (Menard, 1991). In particular, *stage-state models* are concerned with the probability of moving from one value (state) to another value of a variable by a given period (stage). For multiple category variables, separate probabilities of *transition* (movement from one value to another in a given interval between periods) are calculated for each pair of *origin-destination* states (state, respectively, at the beginning and at the end of the interval). When the origin and the destination states are the same, the transition probabilities indicate the stability in the state over time.

Stage-state models can be described by using (Menard, 1991):

a. simple transition matrices, with no assumption about underlying properties of the transition matrices (based on measurement taken for few or two, periods),
b. Markov models, including Markov chains (based on measurement taken for few or two, periods),
c. log-linear models (based on measurement taken for few or two, periods),
d. univariate life table models (based on measurements taken for several periods),
e. univariate survival models (based on measurements taken for several periods); they assume that survival rate follows fixed distribution these models are useful to model processes like recidivism, labor force participation, marital history events and other transition events

among discrete states.

The *a* and *b* models are based on simple row percentages from cross-tabulations or contingency tables that compare the values of a variable for the same set of cases.

For Markov and life table models it is possible to define a particular state that, once entered, cannot be removed named *absorbing*.[5] In these cases, it is possible to calculate (Menard, 1991):

- what proportion of cases will be in the absorbing state, and each other state at a given period;
- how long it will take all cases, or a certain proportion of cases, to enter the absorbing state.

Some *stage-state models* were developed in socio-economic field with reference to particular concepts.

- *Models connected to the concept of 'wastage' (or 'turnover' or 'attrition')*
  This model is concerned with the loss (and its dynamics) of individuals from closed system, like a firm, and is dealing with a stochastic process, developing in time. The development of the model allows the comparison for the same system over two different periods; many functions may be described, like parameter of the propensity-to-leave function (Bartholomew, 1996).
- *Models connected to the concept of 'mobility'*
  This concept is typical of social researches concerning social or occupational mobility (Bartholomew, 1996). Particularly in social dimension, intergenerational, in terms of transitions of family lines from one generation to the next, and intragenerational, in terms of changes of class within the lifespan of the individual, mobility can be defined. The mobility process can be observed in other fields (movement of firms, geographic mobility of individuals, families, and so on). Each mobility model is defined by two dimensions:
  a. classes, between which movement takes place; they can be formed in a variety of ways (income, occupation status, occupational skill, satisfaction level, etc.);
  b. time, that generally is treated as discrete, since usually data are available at fixed intervals of time (months, year, contracts of employment, holiday periods, and so on).

  Consequently, definition and analysis of mobility models refer to discrete classes changing state at fixed time intervals.

  Approaches to modeling are related to two different aspects of mobility processes:
  - *pure mobility*: the number of individuals (such as fathers) in each class are treated as fixed while the number of lines moving to other classes (such as sons) are modeled;
  - *structural mobility*: the number of places of individuals in the second occasion (sons) are treated as fixed while the reverse flows of vacancies back to individuals of the first occasion (fathers) are modeled.

Both depend on the discrete time Markov chain (Markov, 1979; Bartholomew, 1996). Alternative approaches to discrete change modeling are based on loglinear models and latent class models (Lazarsfeld and Henry, 1968 Bartholomew and Knott, 1999). Wiggins and Coleman models are considered extensions of Markovian models incorporating error of measurements (Markus, 1979).

1.1.3 Modeling applications in quality of life

Generally, modeling approaches reported here found applications in different kinds of study especially concerning sociopolitical attitudes and behavior, of public opinion and voting and may find applications in subjective quality of life studies as well.

However some remarks are required.

With regard to deterministic approach, application of chaotic theory to social setting was made possible in the past only through simulations because of lack of adequate measured data. Nowadays, the availability of a system of social indicators (endowed with individual and aggregate data) makes possible to find evidence of chaotic models in social scenery especially for those social phenomena

[5] Typical absorbing state is death.

having long and periodic properties, as happens in many individual behavior models (Brown, 1995), related to subjective quality of life ambit.
On the other side, possible applications to quality of life study, as in general to social sciences studies, of catastrophe theory is seriously limited because of complicated structures of the model and technical problems in defining social processes; in other words, catastrophe theory remains an interesting but not unique model proposal (Visser, 1985).
With regards to probabilistic time related approaches, we have to notice that usually models are expressed in terms of growth (growth/development/achievement models), finding applications in achievement research and in many approaches to educational and psychology research (Rogosa et al. 1982; Rogosa and Willett, 1985; Embretson, 1994).
Main advantage in using growth models is connected with the possibility of predicting variable values (Goldstein, 1979), by individuation of patterns, linear and not linear, of events or relationship.[6] However the main limit of the approach is conceptualizing change only in terms of gains, of a 'step by step' function in analogy to a 'building block model', where no negative blocks are defined; such model may fail in describing processes underlying subjective quality of life. In other words, the adoption of growth models is not always justified, especially in quality of life studies, where the model of change cannot be defined only in terms of time related growth; moreover, differences in both directions are expected depending on different predictive factors (for examples we may expect that positive and negative life events may yield individual change in different directions on the same dimension). Moreover, growth-models parameters are directly comparable between different cases (subjects, nations, organizations, and so on) only after testing comparability of case starting conditions.
Modeling subjective quality of life change by difference (change/mobility/residual) models seems to be more interesting since this approach may help to explore presence of causal patterns.

1.2 DESIGN

The study of quality of life and, as a rule, of social dynamics requires definition of designs in order to obtain repeated measures over time on same individuals with respect to specific dimensions. Three features characterize each design:
- data are collected for each variable at two or more distinct points of time,
- involved individuals are the same, or at least comparable, from one occasion to the next.

According to these, different designs (Goldstein, 1979; Menard, 1991; Firebaugh, 1997) can be identified (figures 5 and 6).

[6] In statistical terms, this requires (Goldstein, 1979; Firebaugh, 1997; Bryk and Raudenbush, 1992):
- identifying interpolating a line connecting points for the observed group (*individual change model*),
- estimating model parameters (slope and intercept) in order to identify a general modification of characteristic and to analyze differences among individuals (*general change model*).

Fig. 5 Characteristics of typical designs

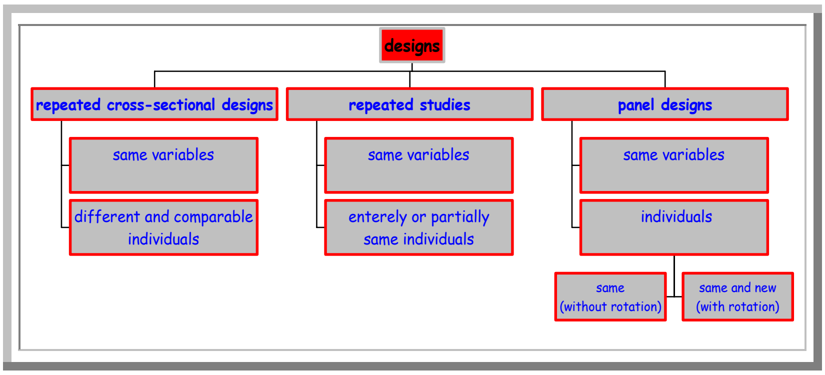
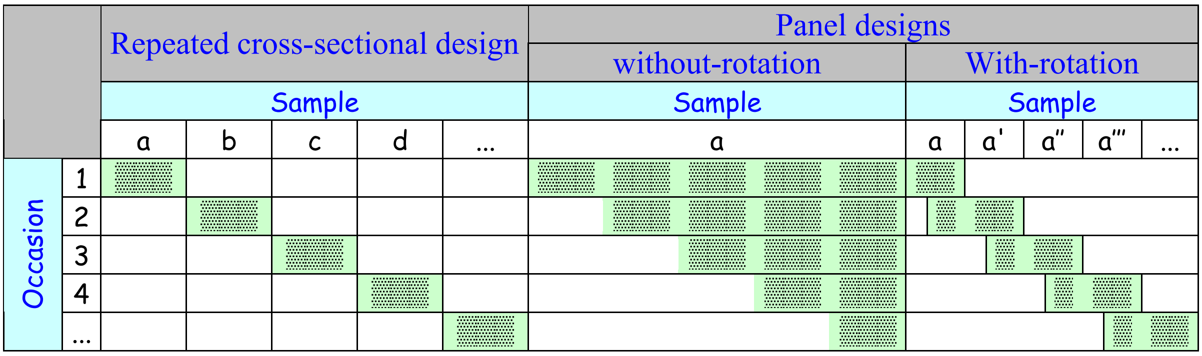


Fig. 6 Schematic representation of three typical repeated sample designs



Repeated cross-sectional designs, in which data for each occasion are observed on different but comparable individuals on same variables.

In some circumstances in change studies, *one-occasion cross-sectional survey design* can be applied (Hagenaars, 1990); these can be done by:

- using information, provided by respondents, about their past behaviors, attitudes, beliefs, etc.;
- comparing different age groups in order to observe generational change (meeting two conditions: the differences observed at one occasion can be interpreted as a difference between generations; the generational characteristics are stable over time);
- interpreting cross-sectional relations between certain variables as explanation of change over time (education level and kind of work); this can be done when one condition is meeting: the system of variable under study is in a dynamic equilibrium.

In spite of interesting application in change studies, in general this design does not provide reliable and valid evidence (Hagenaars, 1990).

Repeated designs, also called periodical or recurrent on:

- groups, entirely or partially formed by same individuals,
- groups formed by different individuals;

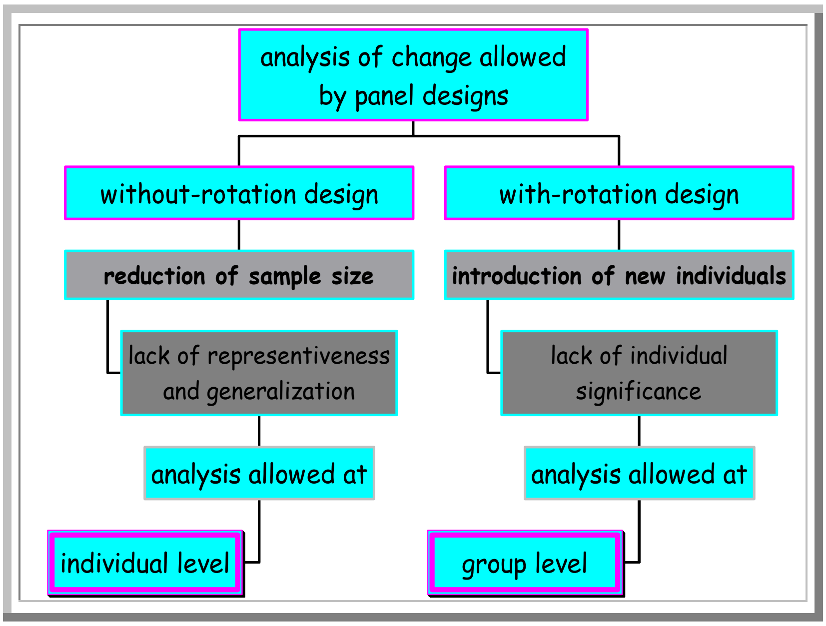
Groups defined for repeated studies are not able to keep representativeness as regard to reference population.

Panel designs, defined in two different ways:

- <u>without rotation</u>: repeated surveys on same individuals; the goal is to study change at individual level; outcomes may not be generalized because of reduction of sample size;
- <u>with rotation</u>: repeated surveys on same individuals and new individuals in order to maintain statistical generalization of sampling outcomes.

The two approaches allow respectively two different levels of change analysis (figure 7) both useful in quality of life studies.

Fig. 7 Different analysis approaches allowed by panel designs



Panel studies are defined by the designed number of points of time, generally named *waves*. The significant number of waves depends on the defined model. Usually, studies dealing with time related models require longitudinal design with more than two waves (Goldstein, 1979). On the other hand difference models can be significantly based also upon two waves.[7]
Repeated measurements designs may also be distinguished in (Lindsey, 1999):

- *prospective design*, in which data are collected at two or more distinct occasions on same individuals and variables; traditionally this kind of approach is typical of survey designs (panel and cohort designs) and experimental designs (clinical trial);
- *retrospective design*, in which values of previous explanatory variables, concerning past occasions, are collected and investigated for each respondents. Difficult to apply in social ambit as in quality of life studies, such design is typical of experimental study (case-control study in medical sciences).

In quality of life studies, repeated survey design seems to be the most applied especially in comparing data at aggregated levels.[8] On the other hand, large panel designs are recommended in order to obtain more explicative information about individual processes about subjective change in quality of life dimensions.

1.3 MEASUREMENT

In practice, measurement of change requests at least

a. two measures obtained in two different moments on same subjects or
b. one measure obtained in one different moments for each of two comparable individuals.

[7] As reported in this second part of the paper, combination of waves and number of involved variables defined different types of model; the simplest one is the two-waves two-variables (2W2V) model.

[8] Social Indicators Research, one of the journals of major reference for quality of life topics, produces many interesting works on data obtained by repeated designs in order to describe national/regional trends.

In case *a* the change is measured at individual level, in case b at group level. In both case measurement of change requires, as a rule, a coherent operational definition of change based upon definition of:

1. methods for measuring amount of change (operational definition of change),
2. requirements for comparability conditions,
3. approaches for testing reliability and validity of measures over time.

Approaches to measurement of change at individual level deserve examination[9]

1.3.1 Measurement of change at individual level

Change may be defined in terms of

1. comparison between two values observed for one variable, (comparison between original scores)
2. difference between two values observed for one variable (difference score).

Choice between original-score comparison and difference score depends principally upon (Menard, 1991) theoretical considerations concerning the use of change measure in the particular research ambit and upon measurement level.

Original-score comparison
This approach can be always applied; however is a forced choice when the study involves categorical variables; in this case we define the observed diversity between two categories (in presence of categorical values)

$$D = X \Leftrightarrow Y$$

where
X individual classification at the first occasion
Y individual classification at the second occasion.

Difference/change score
In social research, and in quality of life studies as well, many concept and hypotheses are formulated in terms of *change*, conceived as a variable itself (Menard, 1991). Requirements in order to obtain difference/change score, applied principally in panel (experimental or observational) design (Menard, 1991), concern (figure 8) measurement level, at least ordinal or metric, and observation of score distribution. Traditional calculation approaches for change score are (Menard, 1991):

1. *change score* (*D*)[10], *raw change* or *raw gain*: difference between two obtained scores:
$$D = Y - X$$
2. *residual score* (*RS*), following linear regression method: *Y* is regressed (using linear regression) on *X* in order to obtain a predicted or expected value for *Y*:
$$RS = Y - E(Y) = Y - a - bX$$
In this perspective, *D* represents a special type of residual score with *b*=1 and *a*=0.
This approach helps to identify those cases that change more or less than expected, given some initial level or value on the variable whose change is being described.[11]

[9] Measurement of change at group level may be defined principally in terms of comparison or difference between certain group-descriptions indices (mean, median, etc.). Both comparison and difference approaches may be carried out by different statistical instruments, according to measurement levels.

[10] In case of multi-items measures then $D = Y - X = \sum_{i=1}^{k}(y_i - x_i)$ where *k* represents the number of measures.

[11] Residual score allows eliminating negative correlation between pretest and difference score (Engel and Reinecke, 1996).

3. *percent change* (*PC*), for ratio scales:

$$PC = 100\frac{Y - X}{\max(s)}$$

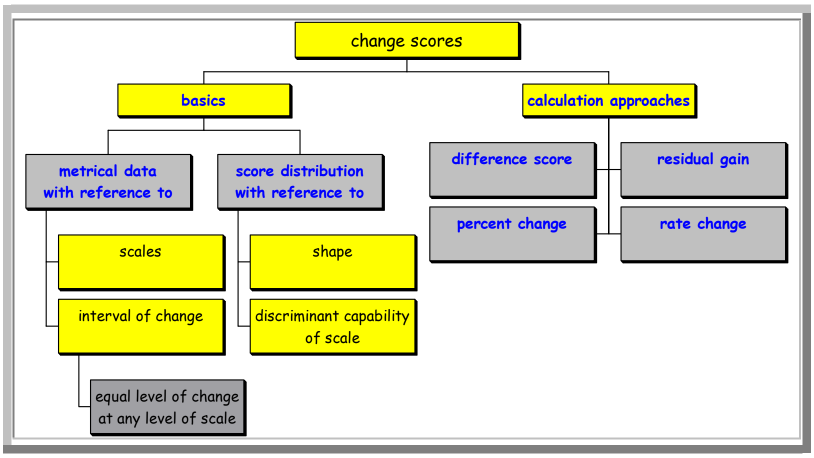
where $\max(s)$ is the maximum value on used scale for both X and Y

4. *rate change* (*RC*), for ratio scales:

$$RC = \frac{Y}{X}$$

where
X individual value at the first occasion
Y individual value at the second occasion.

Fig. 8 Basics and calculation approaches to change scores.



Change variable may be used over more than one occasion and can be used for descriptive purposes, moreover may be used also at group level, if individuals are the same.
Even if adoption of change score has an intuitive appeal, the decision to use one of the approaches is not simple since it yields problems of comparability. Moreover, there is some disagreement in social sciences on the use of change as a variable in order to analyze, predict or explain change (e.g., in causal model). Particularly, in Cronbach and Furby opinion (Cronbach and Furby, 1970, 'How we should measure "change" – should we?' in *Psychological Bulletin*, 74, pp. 68-80), the use of change score is not recommendable since change score are systematically related to any error of measurement and are typically less reliable then original scores; the unreliability of change score may lead to fallacious conclusions or false inference. They admit only the use of residual change in order to compare individuals regarding amount of difference between observed and expected change. Difference scores present difficulties with respect not only to reliability but also to correlation between change and initial score.[12]

[12] The problem can be statistically analyzed by studying bivariate distribution of X and Y. Particularly, we may determine the covariance between X and the difference score D:

$$\operatorname{cov}_{XD} = \operatorname{cov}_{XY} - \sigma_X^2 = \sigma_X^2\left(-1 + r_{XY}\frac{\sigma_Y}{\sigma_X}\right)$$

that usually is negative when σ_Y^2 is not large in comparison to σ_X^2.
So, we have to consider that:

- with a positive correlation between X and Y there will be a negative correlation between X and D;
- individuals with high scores on X tend to have lower score on Y ("regression-to-mean" effect), consequently, a negative difference score;
- the higher r_{xy}, the lower σ_D^2, in other words with high r_{xy}, D has a low reliability; on the other hand if r_{xy} is low, D is reliable; this situation produce some doubts about validity of original observations because of low correlation ("reliability-validity dilemma", Bereiter, 1963).

However, actual possibility to calculate change score seems to be related to reliability of original score. Continuous works done in order to test validity of many measures applied in subjective measurement, also with regard to different cultural contexts, may encourages use of change score.[13]

1.3.2 Comparability

In order to measure change, the measurement should meet conditions of comparability, defined in terms of characteristic, individual, procedure, instrument and measurement approach.
Characteristic: comparability of characteristic over time has to consider that characteristic can be time correlated, age-correlated or events-correlated; in each case, the observed change may lead to different interpretations.
Consequently, in some studies, in order to maintain comparability of characteristic at group and/or individual levels, different instruments for different conditions (in terms of point of time, of individual age, of survey condition, and so) could have to be applied (Goldstein, 1979).
Subject: each individual should have to be comparable with him/herself; frequently even if comparable instruments are available, subjects may result not comparable, for instance in their level of understanding with regard to questions.
Procedure: the measurement procedure, in terms of data collection techniques (e.g.: paper-, CATI-, web-questionnaire) should have to be comparable from one occasion to the other; generally the procedure regards survey method; for example, the interviewers may be different from one occasion to the other, each interviewer may use different approach from one occasion to the other or each subject may be interviewed by different interviewers at each occasion; this problems may also produce errors in classification and in coding; careful training and detailed instructions can help to reduce differences yielded by interviewers errors.
Instrument: in terms of comparability, instruments can be (Webster and Bereiter, 1963):

- o identical: forms that have identically worded items presented in the same order and the same format to the same person;
- o matched: forms that have item statistically matched by statistical criterion; in this perspective, identical instruments represent a special case of matched ones;
- o not-matched or randomly matched instruments.

The presence of other factors (events occurring between the two occasions, memory, learning, survey different conditions, change in characteristic definition, and so on) could prevent from meeting comparability of instrument and make impracticable the use and application of the same instrument.
Measurement approach, defined in terms of level of measurement, types of measures and type of response data.
Level of measurement: the comparability regards level of measurement defined in terms of

- scaling techniques: each technique is defined by scale reference (evaluation, preference, perception, image, judgment), scale type (expression of scale: qualitative/quantitative,

---

It seems that the solution to this problem is the adoption of original scores instead of difference scores. Particularly conditional distribution of *Y*, given *X*, can be used. This can be done considering the purposes of the use of difference scores; they can be:

a. to provide a dependent variable in an experiment,
b. to provide a criterion variable in a correlation study,
c. to provide an indicator of deviant development,
d. to provide an indicator of a construct that is thought to have significance in a certain theoretical framework.

In the last situation a multivariate approach may be preferred, while in the other ones a conditional or partial correlation approach may be sufficient.

[13] For this purpose, World Database of Happiness, Catalog of Happiness in Nations, by Ruut Veenhoven, (www.eur.nl/fsw/research/happiness) represents a landmark.

verbal/graphical), scale range (number of levels for scale – in the sense of scale discriminant capacity);
- scale unit adopted in both occasions; achievement of comparability requests
  - adopting metrical or, at least, ordinal, information;
  - observing shape of score distribution;
  - testing metric interval of change (equal level of change at any level of scale).

  Nevertheless, the adoption of categorical scale in social researches cannot be disregarded. In these cases testing the comparability of scales requires a soft approach, in contraposition to the statistical one, considered hard.

  Comparability conditions expressed in terms of scale unit is particularly important in calculating and interpreting difference scores. In fact, usually studies on change adopt scales that were validated in order to discriminate individuals at one occasion but not necessarily in order to discriminate individual differences. In other words, comparability of original scores does not produce necessarily comparable differences since the same amount of difference may have different meaning at different points of the scale.

Types of measures: with regard to measurement of change it is also important to define two different kinds of measures:
- point measures, obtained for a single point in time; the problem is define the span of time to which assign the point measure;
- interval measures, involving a count of events, or frequency, measured for an interval of time; these measures are defined in terms of the amount of time over which the measurement is taken.

Types of response data[14], that, in subjective measurement, can be defined, with reference to Coomb's theory of data (Jacoby, 1991)[15], in terms of
- single stimulus data (subjects answering to a question on a rating scale),
- stimulus comparison data (subject 'possesses' *x* units of some characteristic),

---

[14] Another approach to classification of type of data produces following categories of data (Lindsey, 1999):
- general continuous data: repeated data that take the form of quantitative measurements supposed to have any value on the real line; very few measurements are able to meet the assumption, but at a high theoretical level; this is particularly true in subjective measurement, also in quality of life ambit;
- categorical and count data: this data are produced when the response is an indicator of which of a number of events has occurred on the same individual; when such response are distinguished for the same individual by no explanatory variable, events can be aggregated as counts (one category of event is being observed, like a subjective behavior related to quality of life perception); one common use of counts is to measure rates; if individual events occur in time and this provide additional information, then the type of data is
- duration and survival data: a duration is the waiting time to some event; an event history is a series of successive events, with the accompanying duration between them; the study of such processes in subjective quality of life requires: a) a continuous variable measuring the time has passed, b) a discrete variable indicating whether an event has occur, or not, at each point of time, c) one or more variables indicating relevant information about event (what kind of event it is)

Traditionally these kinds of data apply to
  - survival and reliability studies, when the observation on an individual begins when the characteristics of interest is first diagnosed;
  - incidence studies, when the duration is measured from the origin (rate of occurrence, obtained by longitudinal follow up);
  - prevalence studies, when the interest is on the frequency in a population at a given point of time (in a fixed population, prevalence=incidence X duration).

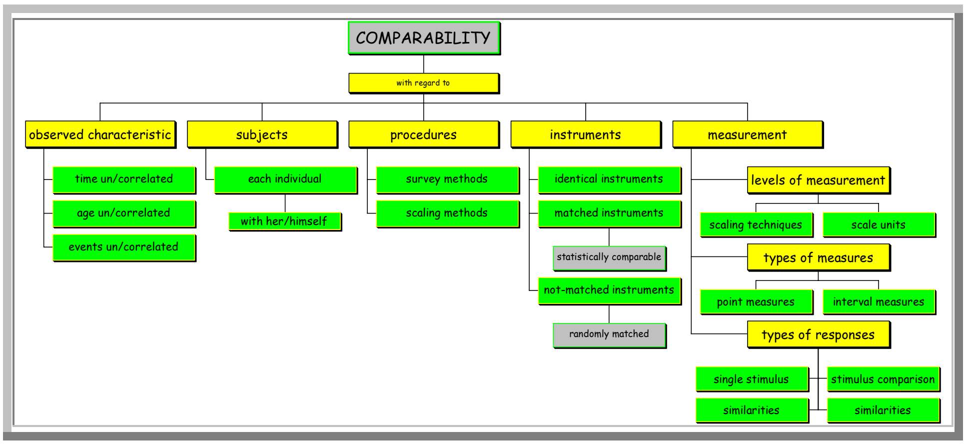
Survival and incidence are modeled by intensity whereas prevalence by probability.
These kinds of data are useful, according to the measured characteristic, in quality of life studies, at aggregated and disaggregated level, however are difficult to be applied in subjective dimension measurement.

[15] Clyde Coombs developed several theories based on geometric interpretation of data (Jacoby, 1991). Synthetically, two entities in a single datum can vary in two ways: a) with regards the set to which the entities belong that can be different (a stimulus and a response) or same (two individual who take the same test); b) with regard to the relation in which the entities are involved that can be a dominance relation (an individual answers a question by reporting a level exceeding a defined measure) or a proximity relation (two individual share an event).

- similarities data (the degree to which two subjects exhibit the same behavior, level of life satisfaction, and so on),
- preferential choice data (a subjects likes or prefers a particular stimulus concerning quality of life).

Testing comparability of types of response data is particularly important in study on change of subjective quality of life perception because, on same topic, we may reach different conclusion on change condition according to data approach applied. This is particularly true in trend studies when there could be a risk to compare individual aggregate data on change over time, with regard to a specific characteristic, obtained by different data approaches (even if by, for example, same questionnaire).
All these elements (schematically represented in figure 9) play a crucial role in testing reliability and validity of measurements over time.

Fig. 9 Elements of comparability to be considered in a study on change.



Meeting comparability condition is particularly important in study on change of subjective quality of life perception since we may reach different conclusions on change condition according to different data approaches both in trend (involving group comparisons) and process (involving individual comparisons) approaches. Even if level of comparability seems to be less severe in trend approach, comparing data obtained by different, for example, scaling techniques (even if by same questionnaire) has to consider the different scale performance in discriminating individuals (Maggino, 2003).

1.4 DATA ANALYSIS

The purpose of data analysis in change studies is to represent observations in order to show the regularities of empirical data and to come to valid statement about it. Choice between the approaches to change data analysis depends upon a) defined model, b) adopted design.
Another criterion useful in distinguishing data analysis approaches is related to types of measurement, particularly with regard to continuous, categorical, count and duration data (Lindsey, 1999).

1.4.1 Data analysis approaches related to model definition

Data analysis approaches can be distinguished with reference to model, in particular:
- time-related models, typical probabilistic approaches are time series analysis (quantitative approach) and turnover table (qualitative approach);
- models for relating measurements at different occasions (analysis of change in structure), for which multivariate approaches exist.

➢ **Time-series approach**

Time series analysis is typically applied in analysis of dynamic system when many repeated measures are observed; the principal goals of this analysis are: forecast future trends and drawing a possible internal structure of the series (Visser, 1985).
Time series analysis can also be applied in analysis of individual change even if produces three characteristics that make the approach difficult to apply especially when each series regard one individual (Holtzman, 1963):

a. time variable can bring other uncontrolled variables;
b. time intervals are arbitrary and, as a result, difficult to allow control of continuous variables;
c. repeated observations are not independent producing a serial correlation and making unusable the great part of statistical models.[16]

The analysis of individual change through time-series approach shows a further problem regarding the generalization by statistical inference.
The individualization of a model for a homogeneous group allows for comparison with other analogous model for other comparable groups.
One of the mainly problems these models have to deal is the presence of missing data.

➢ **Mobility/rotation/turnover tables** (Chazel et al., 1970; Hout, 1983; Hagenaars, 1990)

When change is observed through categorized data instead of metric scores, the analysis needs to adopt different approaches. Among these we may note the *mobility table* approach whose goal is to measure the rotation index instead of line parameter estimates.

➢ **Multivariate approach**

The multivariate approach is applied when the main point of interest is the analysis of change in coherence (in correlational terms) of a set of variables, that is analysis of change in structure. Multivariate procedures may be distinguished in (Visser, 1985):

- *simple descriptive procedure*; technically we may distinguish two techniques in order to describe data:
  - numerical, when the description is based on numerical values of some measures of change;
  - graphical, when the description is based on plotting values of a variable for different moments on a graph (horizontal and vertical axes represent respectively time and variable of interest);
- *exploratory procedures*, whose goal is to represents the change in structure in a simple way with low assumptions about data; among exploratory methods, Principal Component Analysis and Multidimensional Scaling have found particular versions in order to deal with three dimensional matrices;
- *confirmatory procedures*, able to test more specific models, in particular those which explain correlation by causal relationships (analysis of covariance structures).

[16] If the systematic variance of a given observation depends upon preceding observations the time-series defines a Markoff chain (the order of which depends upon the number of number of dependence preceding observations).

In multivariate data analysis many procedures are available to deal with complex structure such as those yielded by repeated measures. This is particularly true when the number of occasions is limited. Many examples of three-mode analysis are available for explorative and confirmative factor analysis[17] (Tucker, 1963; Harris, 1963b, Kaiser, 1963, Cattell, 1963; Bartholomew and Knott, 1999) and for multidimensional scaling (Cox and Cox, 1994).
An important role in selection of the techniques to be used in multivariate analysis is played by data characteristics in terms of size of data matrix. Datasets present particular complexities derived by the typical three-way data matrix (Visser, 1985) in which each dimension of the matrix represents respectively occasions, individuals, and variables.
Since generally multivariate data analysis approaches deal with more than two dimensions matrix, flattening out one dimension in required. This can be done:
- *selecting* one dimension (selection of one individual, one variable, one occasion) on which conducting the analysis,
- *reducing* one dimension in a single value; this approach requires a reducing model for individuals, variables or occasions;
- *flattening out*: in order to avoid the arbitrary decision involved in preceding approaches, yielding arbitrary approximations, the data matrix is sliced in three different direction; each arrangement is a table in which one dimension represents one of the original matrix dimension and the other represents the Cartesian product of the other two dimensions; the resulting three tables are transposed in such a way that the rows may be interpreted as observations and the columns as variables.

➢ **Multivariate approach for experimental data: analysis of variance**
Analysis of variance allows dealing with the analysis of experimental data representing characteristics in different, systematically varied conditions at different points of time (Visser, 1985). The approach requires meeting assumption regarding
- experimental conditions: each individual is assumed to be randomly assigned to treatments and to stay in the same treatment group during the experiment;
- statistical characteristics of data: normal distribution of observed variables and equal variances and covariances between observations over time for all individuals.

When the interest is on differences between treatment effects (repeated observation experimental design) the approach is multivariate analysis of variance (MANOVA); when the interest is also on development of the effects over time, the approach is a combination of analysis of variance and regression analysis (Bock, 1963).

1.4.2 Data analysis approaches related to design definition

Data analysis approaches can be mainly distinguished according to the adopted design; particularly, each design allows analysis at different level:
- *macro change level*: analysis of change at group level, allowed by repeated studies;
- *micro change level*: analysis of change at individual level, allowed only by panel studies.

As described below, analysis at each level allows attainment of different goals (Engel and Reinecke, 1996).

[17] Factor analysis applied to longitudinal data may have different kind of approaches, as:
- factor analysis applied to a reduced matrix,
- analysis of covariance structures (confirmatory approach),
- three-mode factor analysis,
- dynamical factor analysis (designed for the analysis of multiple time series).

- **Analysis of macro change: trend analysis**

The objectives of analysis of macro change are mainly detecting and comparing trends, defined as change expressed as function of time; the analysis is finalized to the decomposition of real trends and irregular trends; particularly the time variable can be observed in term of, or better, the real trend can be decomposed in terms of (Glenn, 1977; Menard, 1991; Firebaugh, 1997):

- *period* of observation (for instance "year"), interpretable as "changes over time" effect, that refers to change produced by influenced related to historical age under study;
- *age* of observed individuals, interpretable as "life cycle and developmental changes" effect, that refers to change produced by influenced related to age (considered as individual life-cycle status);
- *cohort* of which each observed individual is a member (defined in different terms like geographical, event, time of particular event, generation, year of birth, and so on), interpretable in different kind of effects, that refers to cohort differences that result from common experiences or reactions of a cohort (according to the cohort definition).

Each of these may represent an explanatory effect of change to be considered (separately or in combination) in the model.
Synthetically:

$$\textit{year of birth} = \textit{year of measurement} - \textit{years}$$
$$\textit{cohort} = \textit{period} - \textit{age}$$

Four methods can be individuated in order to study group trend (Firebaugh, 1997):

1. <u>trend analysis</u>: analysis of a presence of an average change in a group over time and comparison between different trends (coincident, parallel, converging, diverging, crossed trends);
2. <u>proximate decomposition</u> of trends (proximate source of change), distinguishing between net change among individuals and gross change due to group turnover; the analysis is based on linear regression or algebra approach;
3. <u>change decomposition</u> of aggregate change in one variable in terms of change in the levels and effects of other variables; the analysis is based on the decomposition equation regression;
4. <u>changing-parameter method</u>: analysis of change in the effects of variables at individual level, in order to determine the time-dependence of individual-level relationships.

Since repeated studies are based on substantially independent samples, trend analysis provides sufficient description of change but fails in providing empirical explanations of the change process.

- **Analysis of micro change: process analysis**

The analysis is not a simply descriptive but explanatory one (process analysis or internal analysis). The analysis is accomplished at individual-level by investigation of covariation over time; this approach is allowed essentially by dependent samples (panel studies).
Taking into account fundamental requirements for establishing a causal relationship (covariation, temporal precedence and nonspuriousness) we can observe following types of change (Menard, 1991):

- *initiation*, referring to the first time that a case enters a particular state,
- *escalation/reduction*, referring to the entry of a, respectively, higher or lower state (on an ordinal scale)
- *suspension*, referring to a permanent or temporary exit from all states that indicate involvement a particular state; this type of change is not always significantly present.

Because of the possible long term of occurring, collecting data in order to study and analyze the processes described by this model may be difficult. The problem, called 'left-hand censoring', indicates the failure to detect a change because it happened before the period of data collection.
Consequently in order to unravel causal relationships, it is important to qualify the model in terms of adequateness of lag time (meant as interval between data collection periods of time) in order to allow

- detection of change in a variable clearly separated from change in another;[18]
- occurrence of cause to produce an effect.

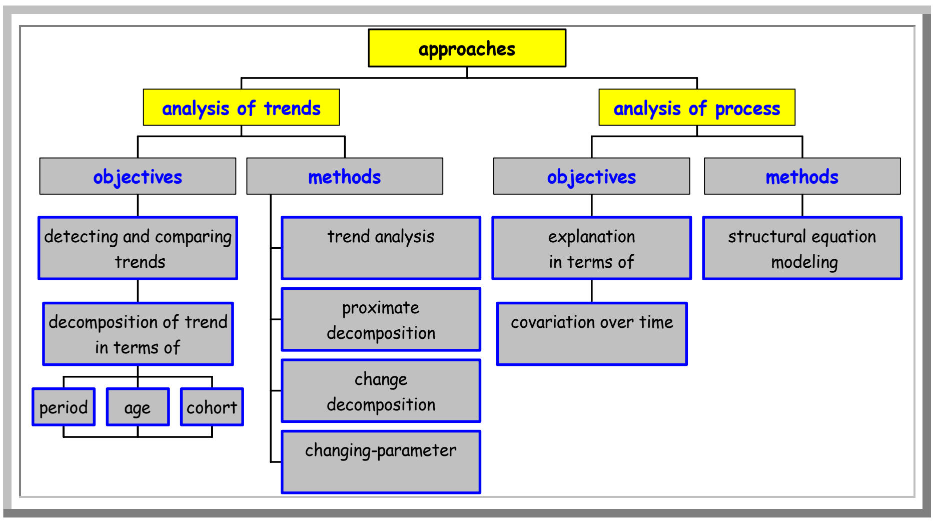
Appendix A shows the rational of structural model approach to panel data analysis (Shingles, 1985), joining measurement and structural model definitions (see second part of the paper).
In order to unravel supposed causal relationships by collected data, traditional data analysis methods can be classified as represented in table I (Menard, 1991).

Tab. I Data analysis approaches for causal relationship hypotheses

Dependent variable	Independent variable	Methods of analysis
Quantitative/continuous	Quantitative/continuous	Differential equations
		Regression
		Multivariate ARIMA time-series analysis
		Latent variable structural equation models
	Mixed continuous and categorical	ANOVA with ANCOVA
		Regression with dummy variables
	Qualitative/categorical	ANOVA
		Nonparametric ANOVA
		Dummy variable regression
Qualitative/categorical	Quantitative/continuous	Discriminant analysis
		Logit or probit analysis
		Logistic regression
		Survival/event history analysis
	Mixed continuous and categorical	Log-linear analysis
		Logistic regression
		Survival/event history analysis
	Qualitative/categorical	Log-linear analysis
		Multistate life table models
		Survival/event history analysis

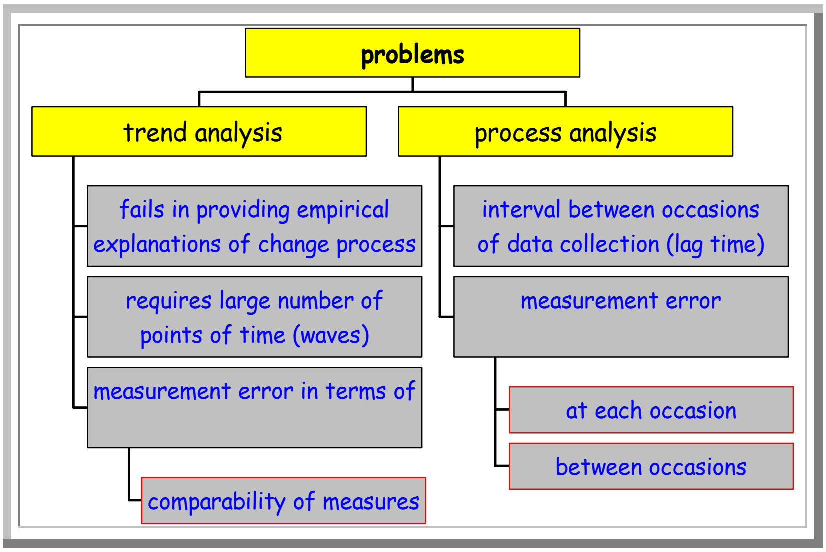
Moreover, figures 10 and 11 summary, respectively, characteristics and problems of the two levels of analysis.

Fig. 10 Objectives and methods to analysis of trend and analysis of process.



[18] If the change in both variables occurs in the same period, there are different possible explanations (Menard, 1991):
a. the two variable measure the same thing,
b. the two variables are spuriously related, having a common cause producing changes in both,
c. the length of measurement period does not allow to separate the two changes.

Fig. 11 Problems in analysis of trend and analysis of process.



Peculiarities of data analysis in study changing of subjective quality of life concern elements not so different from that we will see about reliability analysis approaches.

2. RELIABILITY OF SUBJECTIVE MEASUREMENT OVER TIME

The success of any study on change in subjective quality of life, apart from the defined model and the adopted design, is strictly connected to the possibility of having reliable measures of change of a "modifiable" dimension (Goldstein, 1979). In other words, separating 'true change' from 'error component' is required in order to meet the principal purposes of analyses. In fact, the undetected presence of error in the measurement of a single variable at two different occasions, may lead to two extreme situations:

- the two observed values are identical, even if they are actually different,
- the two observed values are different, even if they are actually identical.

The detection of error of measurement of change in quality of life studies, and the subsequent correct interpretation of a difference between two subsequent measures, is relevant not only in descriptive studies but particularly in studies whose goals are defined in predictive and planning terms.

The theoretical and statistical debate regarding the complex problem of reliability in measuring change has a long history and found numerous solutions. However, even if important works on measurement of change can be found in literature, we are not able to assert to have solved problems of measurement (Harris, 1963a) since the majority of works has focused principally on methods of analyzing change (Embretson, 1994; Schutz, 1994). However, some clear fundamentals can be singled out.

Testing reliability of subjective change requires the definition of an experimental design and a data analysis approach.

2.1 TESTING RELIABILITY OF MEASUREMENT OVER TIME: EXPERIMENTAL APPROACHES

Definition of an experimental design in testing reliability of change is not simple. It has to allow explaining the observed change and distinguishing different sources of change (schematically represented as in table II):

a) <u>change in individual</u>: the subject has changed (*real change effect*),
b) <u>change in characteristic</u>: the second measurement does not observe the same characteristic (*trait effect*), that maybe needs
    - another definition,
    - another measurement procedure for the second moment;
c) <u>change in survey condition and measurement procedure</u>, that is not reliable in the measurement of the differential (*method effect*).

The experimental design for testing reliability of measurement over time should make possible distinction between different effects or, better, detection of method effect.

Different approaches to experimental design are possible according to the level of measurement of change that can at:

a. individual level, allowed by panel design,
b. group level, allowed by repeated design.

Assumptions requested by the former are (Harris, 1963a):

- same (or comparable) instrument in the two measurement procedures for the same characteristic,
- same measurement unit (or comparable measurement units) in the two measurement procedures for the same dimension,
- other variables that may explain or test the real change.

Tab. II Identification of the source of change in a reliability study.

Source of change	Interpretation	Defined effect	Solution
Change in characteristic	In the second occasion the trait is not the same	*Trait effect*	Need of another theoretical definition
Change in survey condition	Lack of comparability	*Method effect*	Assessment of survey condition
Change in measurement procedure	Lack of reliability on measurement of differential		Assessment of measurement procedure
Change in individual		*Real change effect*	No solution requested

2.2 TESTING RELIABILITY OF MEASUREMENT OVER TIME: ANALYSIS APPROACHES

Since observing high reliability values for each measure at each occasion does not allow assuming, as a consequence, a reliable measurement of change in terms of comparison of two scores or in terms of difference score, particular analysis approaches are required also according to the operational definitions of change adopted (original score or change/difference score).

2.2.1 Reliability of original scores

Two traditional approaches are identifiable in testing reliability of measurement of change: classical model and structural modeling.

2.2.1.1 Classical approach

Classical model of measurement of change is based upon the definition of three variables:
- an initial measure (*pretest*),
- a final measure (*post-test*) by a identical or matched instrument,[1]
- an independent variable, explaining the change.

Synthetically, the considered variables (Bereiter, 1963) are:

X_{O_j} Observed total score for *pretest* for subject j[2]

Y_{O_j} Observed score for *post-test* for subject j[3]

[1] Using identical forms enhances the measurement of change, even though in this case true and error components are not independent.

[2] In the case of multi-item approach the observed score for pretest is obtained by $\sum_{i=1}^{k} x_{ji}$

X_{T_j} True score for *pretest* for subject *j*

Y_{T_j} True score for *post-test* for subject *j*

e_X Error for *pretest*

e_Y Error for *post-test*

According to the assumptions of classical theory of measurement we know that (Goldstein, 1979):

- each observed score is different from true score because of error:

$$X_{O_j} = X_{T_j} + e_X \qquad (1a)$$

$$Y_{O_j} = Y_{T_j} + e_Y \qquad (1b)$$

- *e* is not a methodological characteristic;
- *e* is uncorrelated to and independent from individual score or other individual variable
- observed score tends to true score in the limit by increasing number of measures; otherwise the true value can be estimated.

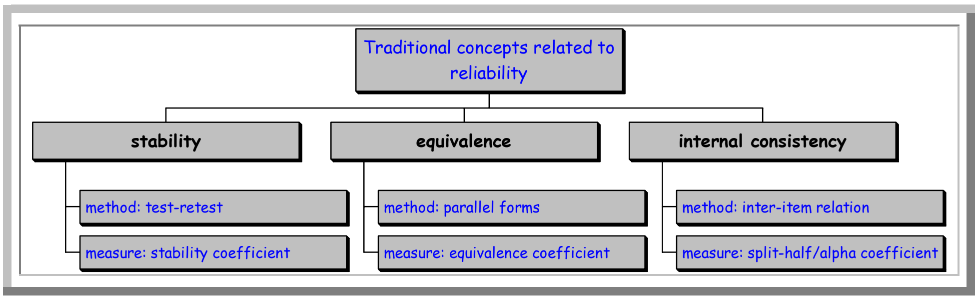
The reliability estimates are respectively defined by the following ratios:

$$rho_X^2 = \frac{\sigma_{T_X}^2}{\sigma_{T_X}^2 + \sigma_{e_X}^2} = \frac{\sigma_{T_X}^2}{\sigma_{O_X}^2} = \frac{\sigma_{O_X}^2 - \sigma_{e_X}^2}{\sigma_{O_X}^2} \qquad (2a)$$

$$rho_Y^2 = \frac{\sigma_{T_Y}^2}{\sigma_{T_Y}^2 + \sigma_{e_Y}^2} = \frac{\sigma_{T_Y}^2}{\sigma_{O_Y}^2} = \frac{\sigma_{O_Y}^2 - \sigma_{e_Y}^2}{\sigma_{O_Y}^2} \qquad (2b)$$

Traditionally, estimate of defined reliability can be conducted by repeated measures in terms of stability, equivalence, consistency, as represented in figure 12.

Fig. 12 Classical concepts in testing reliability



Estimating reliability in terms of stability. One of the measurement attributes that can be used in estimation of reliability is the concept of *stability* defined in terms of

- stability of measurement over time (short and long time),
- stability of measurement over different survey methods.

Test-retest approach is considered the traditional rationale in order to estimate reliability in terms of stability of measurement. *Stability coefficient*, based upon a correlation measure, is an estimate of reliability of measure. Since the basic assumption of *test-retest* approach is the stability of measured traits, estimate of reliability in terms of stability is not applicable in the assessment of measurement instruments in change studies, where the interest is generally concentrated on modifiable traits.

Estimating reliability in terms of equivalence. Another measurement attribute that can be used in estimation of reliability is the notion of *equivalence* between the instrument and another one that can be considered parallel to the first one and whose reliability was already tested. Estimate of reliability is considered the *equivalence coefficient* defined in terms of correlation between the two parallel instruments. The approach assumes for both parallel instruments:

---

[3] In the case of multi-item approach the observed score for post-test is obtained by $\sum_{i=1}^{k} y_{ji}$

- equal observed score and equal true expected score for each subject,
- equal variances or, equivalently, equal standard errors,
- equal covariances between observed and true scores.

Restrictive assumptions, together with the difficult practical application, make impracticable this approach in the estimate reliability of measurement in change studies.

Estimating reliability in terms of internal consistency. The estimate of reliability of measurement in change studies in terms of internal consistency approach (by alpha coefficient), as well as the parallel component approach (by split-half technique), may represent an overcoming of problems pointed out from previous approaches but only for multi-item instruments.

Estimating reliability in terms of error variance. Generally another estimator of stability between two subsequent measures is assumed the error variance. Assuming that the two measurements are independent and identically distributed, the usual estimator of error variance is the *Gross Difference Rate* (*GDR*) divided by two (Biemer and Lynne Stokes, 1991):

$$\hat{\sigma}_e^2 = \frac{GDR}{2} = \frac{\sum_{i=1}^{n}\left(X_{O_i} - Y_{O_i}\right)^2}{2n}$$

In case of categorical variable the measure of stability is based upon the cross-classification of two measurements of each individual at two occasions (table III); particularly, the estimator of error variance (Forsman and Schreiner, 1991) is[4]:

$$\hat{\sigma}_e^2 = \frac{GDR}{2} = \frac{\sum_{i=1}^{n}\left(X_{O_i} - Y_{O_i}\right)^2}{2n} = \frac{b+c}{2n}$$

Tab. III Cross-classification of two measurements at two occasions for a categorical variable.

		1st occasion		
		1	0	total
2nd occasion	1	a	b	**a+b**
	0	c	d	**c+d**
	total	**a+c**	**b+d**	**a+b+c+d=n**

The interpretation of these measures in terms of measurement errors depends on the experimental design. In other words, the approach allows detection of error of measurement only for stable characteristics, in short or long terms but does not allow testing the reliability of a modifiable characteristic.

Since the classical reliability approach is based upon the concept of repeated measures and the considered variables are assumed to change significantly in change studies, the approach founded upon correlation method, interpreted in terms of stability, equivalence or consistency, is difficult to use and to apply in testing reliability of measurement of change.

Moreover, the assessment of original score contains a factor difficult to control by classical procedure: related errors of measurement between occasions. However, as we have seen, the reliability of measurement of a construct, assumed to change over time, can be determined only by interindividual variances at a fixed time (Schnabel, 1996).

[4] Replacing the previous assumptions with the assumption that the second occasion provides true values, the estimator of response bias of the first measurement is given as the Net Difference Rate (*NDR*)

$$NDR = \frac{b-c}{n}$$

2.2.1.2 Structural modeling approach[5]

One of the solutions to detection of error of measurement of change refers to structural modeling, here considered in terms of latent variable model (Engel and Reinecke, 1996).
Particularly, latent variable model represents the alternative approach for analyzing complex longitudinal designs, because of the, generally considered, great strength of causal inferences that are grounded on longitudinal analysis (Engel and Reinecke, 1996; Schnabel, 1996).
Even if analyses and interpretations of both change and causal relations present shared difficulties and even if this approach presents critical remarks (Schnabel, 1996), the latent variable model seems to be suitable to face the measurement error problem in the analysis of change.
By making few assumptions, the approach allows to obtain estimates of the reliability of the measure and the temporal stability of the true score (Sullivan and Feldman, 1981). If the defined model holds, estimates of true scores can be considered to have identical characteristics of the true latent variable.
Moreover, one of the advantages of latent variable approach is its possible application on nominal and ordinal data combining latent variable/s and log-linear models in latent class analysis (Lazarsfeld and Henry et al., 1968; Markus, 1979; Hagenaars, 1990).

Error of measurement and causal modeling

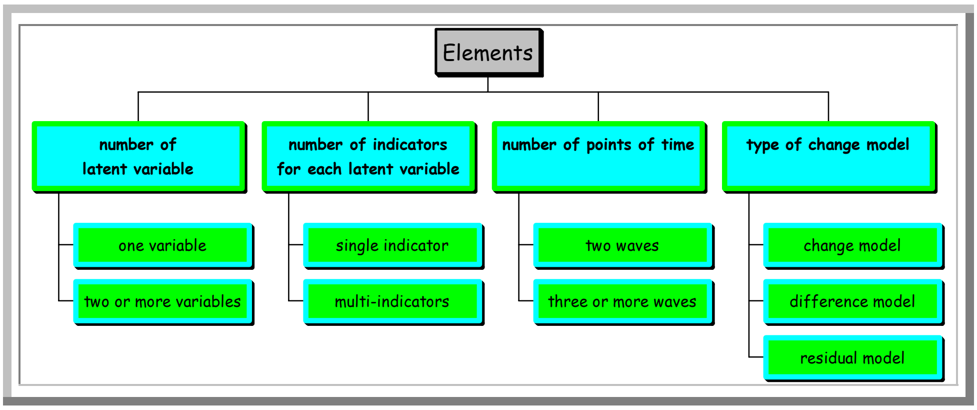
The estimation of true casual effect of the latent variable will be biased if observed indicator is affected by random measurement error. Particularly, unreliable measures will lead to underestimation (attenuation) of a variable effect (Finkel, 1995). In a bivariate regression model, the estimate coefficient will be equal to

*(true coefficient) * (indicator reliability)*

In multivariate models, however, the direction of the bias may be in either direction, and the presence of measurement error in any one of the independent variable can yield bias estimations even in those connecting reliable variables.
Moreover, reliability analysis of longitudinal measures (defined in panel terms) following structural modeling approach has to consider Y_t as a function of Y_{t-1} (*lagged variable*) and some independent variables (X_t, X_{t-1} , and so on).
Since measurement error in lagged endogenous variable may represent a serious problem for causal inference, the definition of the model has to take into account measurement error in estimating structural effects. Seeing it in another perspective, this means that the definition of a causal model taking into account measurement error may help in disentangling the reliability testing in the measurement of change.
Each model is defined by different aspects, as shown in figure 13.

[5] LISREL notation (Saris and Stronkhorst, 1990) adopted here does not take into consideration the distinction between exogenous (ξ) and endogenous (η) structural (latent) variables, as usually is done in structural equation modeling. All latent variables are treated as endogenous variables (and denoted with η symbol).

Fig. 13 Elements defining a structural model of change



Reliability and stability estimates

Classical equation of measurement (1a and 1b) can re-expressed in terms of one latent variable; particularly, the error model is (Finkel, 1995):

$$y_{kt} = \lambda_{kt}\eta_t + \varepsilon_{kt} \qquad (3)$$

where

k indicator
t point of time
y_{kt} observed score for indicator k at time t of true score (endogenous structural variable)
η_t true score (endogenous structural variable – latent variable)
λ_{kt} unstardardized coefficient linking observed score and true score
ε_{kt} random error

The variance of the observed indicator is

$$\sigma^2_{y_{kt}} = \sigma^2_{\eta_t} + \sigma^2_{\varepsilon_{kt}}$$

and the reliability of, respectively, single-indicator and multiple-indicator models

$$rho^2_y = \frac{\sigma^2_{\eta_t}}{\sigma^2_{\eta_t} + \sigma^2_{\varepsilon_{kt}}} \qquad (4a)$$

$$rho^2_y = \frac{\lambda^2_{kt}\sigma^2_{\eta_t}}{\lambda^2_{kt}\sigma^2_{\eta_t} + \sigma^2_{\varepsilon_{kt}}} \qquad (4b)$$

Different analytical strategies in order to estimate reliability can be defined according to the considered model.[6] The strategies considered here (Finkel, 1995) are defined as regards to the number of indicators:

- single-indicator approach, in which one measure for each latent variable is employed;
- multiple-indicators approach, in which several measures of the same latent variable are applied in order to estimate structural effects and measurement parameters (Finkel, 1995).

Both approach are combined with the number of points of time (generally two ore three waves).

[6] About the two techniques:
- instrumental variable technique does not provide any direct information about reliability, error variance or other measurement properties of indicators;
- Two Stage Least Square procedure is based upon assumptions about the independence of the instrumental variable and the disturbance term (assumption not sustainable in longitudinal data).

Single-indicator approach

The simplest model is the **two-wave single-indicator** approach; using observed variables with error of measurement, the estimation model will contain the true score η as well as measurement error; however the defined model is not identified and the restrictions that can be defined are, however, problematic (Finkel, 1995). The **three-wave single-indicator** approach, represented as in figure 14 (Finkel, 1995), is easily extendible to longer single-indicator models.

Fig. 14 Three-Wave Single-Indicator model

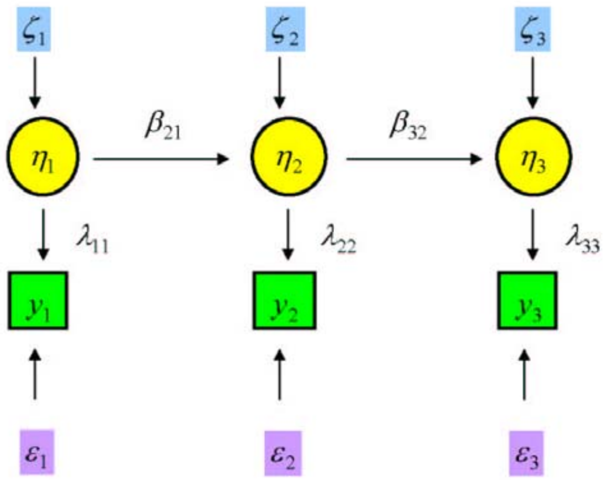


Table IV describes its equations and assumptions. Since there are

- 11 unknown or free parameters (3 ε_t, 3 λ_{tt}, 3 variances of ζ_t, 2 β coefficients)
- 6 'knowns' (variances and covariances of the observed indicators),

not enough information are available to obtain a unique estimates of parameters values; in other words, the model is considered unidentified.

Tab. IV Equations and assumptions describing three-wave single-indicator model

Measurement equations linking indicators to their latent variables	$y_1 = \lambda_{11}\eta_1 + \varepsilon_1$	$y_2 = \lambda_{22}\eta_2 + \varepsilon_2$	$y_3 = \lambda_{33}\eta_3 + \varepsilon_3$
Structural equations describing causal linkages	$\eta_1 = \zeta_1$	$\eta_2 = \beta_{21}\eta_1 + \zeta_2$	$\eta_3 = \beta_{32}\eta_2 + \zeta_3$
Assumptions	$\text{cov}(\eta_t \varepsilon_t) = 0 \qquad \text{cov}(\varepsilon_t \zeta_t) = 0$		
	Measurement errors are randomly distributed with mean zero and constant variance		
	Measurement errors are uncorrelated with one another over time		
	ζ_t are randomly distributed with mean zero and constant variance		
	ζ_t are uncorrelated with one another over time		

In order to achieve identification, two different approaches are available (Finkel, 1995): Heise and Wiley and Wiley.
The Heise approach (Heise, 1985) presents principally two sets of assumptions (Sullivan and Feldman, 1981):
- ➢ Equal reliabilities across measurements. The assumption of constant reliability of a measure along time seems to be quite reasonable; moreover, it allows model identification.
- ➢ Absence of correlations of the error terms across the waves. Particularly, it is assumed that all ζ are uncorrelated among themselves and that all ε are uncorrelated among themselves.

From the analytical point of view, the approach can be summarized as follow (Heise, 1985):
- ✓ standardization of latent variables (variances of η_t are 1);
- ✓ standardization of observed variables;
- ✓ reliability of y is considered equal for the three points of time.

At this point the free parameters are
- stability coefficients, β_{21} and β_{32} (structural portion of the model)
- λ_{tt} , for variances of $\varepsilon_t = 1 - \lambda_{tt}^2$ (measurement portion of the model).

The free parameters can be determined as presented in table V.

Tab. V Heise approach: determination of reliability and stability parameters.

reliability	stability	
	Wave	
	2--1	3--2
$\lambda_{tt}^2 = \frac{r_{y_1y_2} r_{y_2y_3}}{r_{y_1y_3}}$	$\beta_{21} = \frac{r_{y_1y_3}}{r_{y_2y_3}}$	$\beta_{32} = \frac{r_{y_1y_3}}{r_{y_1y_2}}$

Wiley and Wiley considered the first Heise's assumption (constant reliability across time) incorrect, since it yields biased estimates (Sullivan and Feldman, 1981); in fact, as known, the reliability coefficient of an indicator *X* can be interpreted as (Nunnally, 1978) the proportion of variance in the indicator accounted for by the true variable or the ratio of the true score variance to the total variance (this being equal to the sum of true score variance and error variance),

$$rho_y^2 = \frac{\sigma_{T_y}^2}{\sigma_{T_y}^2 + \sigma_{e_y}^2} = \frac{\mathrm{var}(\eta)}{\mathrm{var}(\eta) + \mathrm{var}(\varepsilon)} \qquad (5)$$

In this terms, if there is a change in the true score variance but not in the error variance, it is possible for the reliability of an indicator to change (as it can occur between different populations or between two different moments for the same population using the same measurement instrument).
Moreover, the second assumption appears weak (absence of correlations of the error terms across the waves) because of the presence, especially in subjective measurement, of factors like response set or social desirability that cannot be considered independent over time in individuals. However the inclusion of correlated error terms produces an unidentified model.
Also assuming all ζ terms as completely uncorrelated over time seems to be quite unlikely especially if it is possible to suppose that these error terms are produced by another latent variable (Sullivan and Feldman, 1981).
Possible solutions are (Sullivan and Feldman, 1981) more waves (Heise proposal) or more indicators (Blalock proposal).
In order to face the problem of the first Heise assumption, Wiley and Wiley approach proposes to use unstandardized data[7] (variances and covariances) and to assume constant error variance of the

[7] On possible problems produced by adoption of standardized or unstandardized data see Sullivan and Feldman, (1981).

indicator over time (Markus, 1979; Sullivan and Feldman, 1981; Wiley and Wiley, 1985; Werts et al, 1985).
Parameter estimates are reached through algebraic manipulation of the variances and covariances of observed variables; in particular there will be six free parameters (variances of ζ_1, ζ_2, ζ_3, β_{21}, β_{32} and ε) to be estimated by six observed variances and covariances. From the analytical point of view, the approach can be summarized, in the appropriate recursive order, as follow (Wiley and Wiley, 1985):

$$\beta_{32} = \frac{\mathrm{cov}(y_1 y_3)}{\mathrm{cov}(y_1 y_2)}$$

$$\mathrm{var}(\varepsilon) = \mathrm{var}(y_2) - \frac{\mathrm{cov}(y_2 y_3)\mathrm{cov}(y_1 y_2)}{\mathrm{cov}(y_1 y_3)} = \mathrm{var}(y_2) - \frac{\mathrm{cov}(y_2 y_3)}{\beta_{32}}$$

$$\mathrm{var}(\eta_1) = \mathrm{var}(y_1) - \mathrm{var}(\varepsilon)$$

$$\beta_{21} = \frac{\mathrm{cov}(y_1 y_2)}{\mathrm{var}(y_1) - \mathrm{var}(\varepsilon)} = \frac{\mathrm{cov}(y_1 y_2)}{\mathrm{var}(\eta_1)}$$

$$\mathrm{var}(\eta_2) = \mathrm{var}(y_2) - [\beta_{21}\,\mathrm{cov}(y_1 y_2) + \mathrm{var}(\varepsilon)]$$

$$\mathrm{var}(\eta_3) = \mathrm{var}(y_3) - [\beta_{32}\,\mathrm{cov}(y_2 y_3) + \mathrm{var}(\varepsilon)]$$

Reliability of the indicator at each point and true stability of latent variable over time can be calculated as presented in table VI (Sullivan and Feldman, 1981, Wiley and Wiley, 1985).

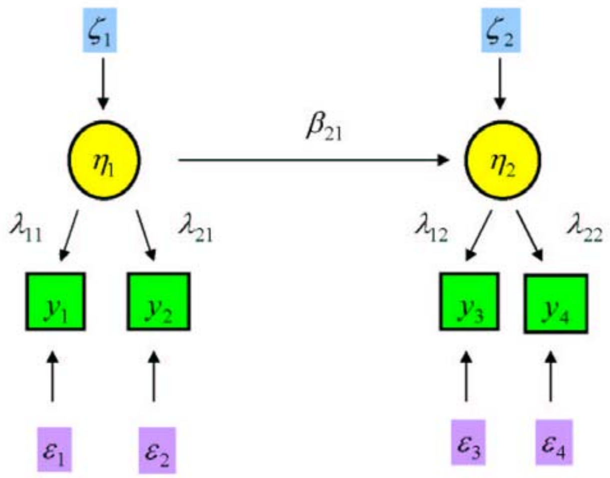
Tab. VI Wiley&Wiley approach: determination of reliability and stability parameters

wave	Reliability estimate	Stability estimate
1	$\frac{\mathrm{var}(\eta_1)}{\mathrm{var}(\eta_1) + \mathrm{var}(\varepsilon)}$	$\beta_{21}\frac{\sqrt{\mathrm{var}(\eta_1)}}{\sqrt{\beta_{21}^2\,\mathrm{var}(\eta_1) + \mathrm{var}(\eta_2)}}$
2	$\frac{\beta_{21}^2\,\mathrm{var}(\eta_1) + \mathrm{var}(\eta_2)}{\beta_{21}^2\,\mathrm{var}(\eta_1) + \mathrm{var}(\eta_2) + \mathrm{var}(\varepsilon)}$	
3	$\frac{\beta_{32}^2\left[\beta_{21}^2\,\mathrm{var}(\eta_1) + \mathrm{var}(\eta_2)\right] + \mathrm{var}(\eta_3)}{\beta_{32}^2\left[\beta_{21}^2\,\mathrm{var}(\eta_1) + \mathrm{var}(\eta_2)\right] + \mathrm{var}(\eta_3) + \mathrm{var}(\varepsilon)}$	$\beta_{32}\frac{\sqrt{\beta_{21}^2\,\mathrm{var}(\eta_1) + \mathrm{var}(\eta_2)}}{\sqrt{\beta_{32}^2\left[\beta_{21}^2\,\mathrm{var}(\eta_1) + \mathrm{var}(\eta_2) + \mathrm{var}(\eta_3)\right]}}$

Multiple-indicators approach

From a general point of view and according to *random sampling theory*, applying multiple-indicator instruments represents a possible solution to reduce error in measurement (Bejar, 1983; Thompson, 2003). The classical approach to test reliability of multi-item instruments, as we have seen, is based upon internal consistency method. By structural approach to test reliability of multiple-indicator instrument allows to assess each distinct indicator.
Availability of more than one indicator for a latent variable allows defining a **multiple-indicator** approach in order to assess reliability of change, overcoming the limitations of single-indicator approach; figure 15 (Finkel, 1995) shows the simple two-wave case (y_3 and y_4 represent respectively the repeated measures of y_1 and y_2).

Fig. 15 Two-Wave Two-Indicator model



The adoption of a multiple-indicators approach makes possible to solve the problems of model identification that we considered for single-indicator approach (Werts et al., 1974; Blalock, 1985). Particularly:

- parameters estimates can be obtained from only two waves,
- two-wave two-indicator model is less sensitive to correlations of ζ terms,
- no assumptions need to be made about equality of reliability over time.

However, computations become and parameter estimations more difficult as the complexity of the model increases.

In order to estimate stability and reliability parameters, Blalock (1985), combining Costner multiple-indicator solution with Heise multi-wave solution, proposed the solution presented in table VII. Moreover, it is possible to 'impose a variety of constraints on the parameters in the models and test these restrictions by relaxing them' (Finkel, 1995).

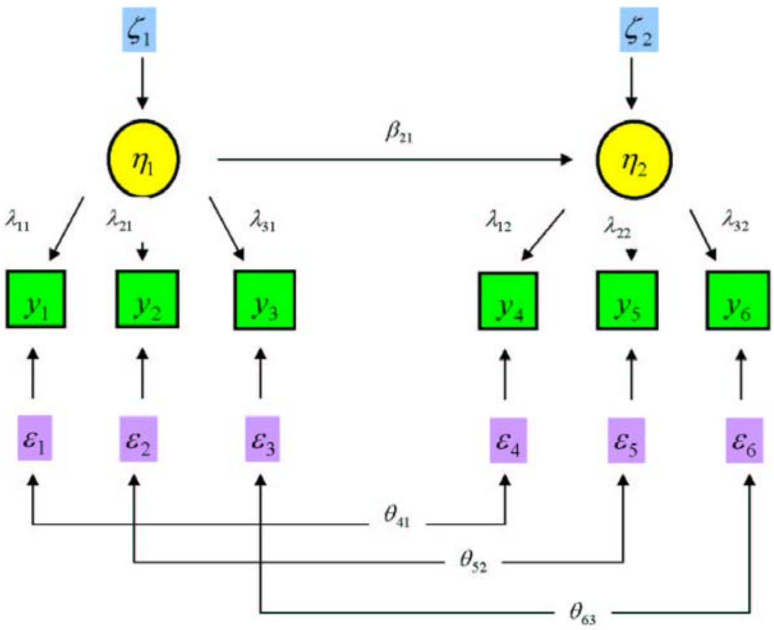
Tab. VII Estimate stability and reliability parameters in two-wave two-indicator approach: Blalock's solution.

Reliability		**Stability**
Indicator		Wave
1	2	2--1
$\lambda^2 = \frac{r_{y_1y_3} r_{y_1y_2}}{r_{y_1y_4}}$	$\lambda^2 = \frac{r_{y_1y_4} r_{y_1y_2}}{r_{y_1y_3}}$	$\beta_{21} = \frac{r_{y_1y_4}}{r_{y_1y_2}}$

The approach allows for more than two waves, more than two indicators for each latent variable, and different kind of related errors, at different levels of combinations.

Such flexibility makes the approach powerful in estimating and evaluating measurement models; in fact, more indicators and more waves allows managing estimate approaches incorporating correlated measurement errors over time (figure 16, where y_4, y_5 and y_6 represent respectively the repeated measures of y_1, y_2 and y_3 - Finkel, 1995).

Fig. 16 Two-Wave Three-Indicator model (correlated errors of measurements)



Reliability and stability parameters presented in table VIII concern one latent variable measured by two indicators for three waves (respectively y_1 and y_2 for first point of time, y_3 and y_4 for second point of time, y_5 and y_6 for third point of time) with, respectively, errors of measurement (ε_1, ε_2, ε_3, ε_4, ε_5, ε_6) correlated in different combinations (Blalock, 1985).

Tab. VIII Determination of reliability and stability parameters in Three-Waves Two-Indicators model with correlated errors of measurement (Blalock, 1985).

Reliability		**Stability**	
Indicator		Wave	
1	2	2--1	3--2
$\lambda^2 = \left(\frac{r_{y_1y_3} - r_{y_3y_5}}{r_{y_2y_4} - r_{y_4y_6}}\right)\left(\frac{r_{y_1y_4} r_{y_3y_6}}{r_{y_1y_6}}\right)$	$\lambda^2 = \left(\frac{r_{y_2y_4} - r_{y_4y_6}}{r_{y_1y_3} - r_{y_3y_5}}\right)\left(\frac{r_{y_1y_4} r_{y_3y_6}}{r_{y_1y_6}}\right)$	$\beta_{21} = \frac{r_{y_1y_6}}{r_{y_3y_6}}$	$\beta_{32} = \frac{r_{y_1y_6}}{r_{y_1y_4}}$

2.2.2 Reliability of difference score

Generally, the important requirements for a reliability index of a score (Webster and Bereiter, 1963) are:

a) it should estimate the proportion of true variance in the observed score;
b) it should allow a prediction of the true score;
c) it should serve as a measure of consistency and reproducibility of items with respect to individuals;
d) it should estimate the "intraclass" correlation among individual repeated measures.

Generally, the theory of the reliability of change/difference scores considers the first three requirements; the fourth cannot be considered since the sequence of repeated score within each individual is very important in measuring change for which we are interesting in differential change. Assumptions for testing reliability (Bereiter, 1963) of measurement of change are synthetically summarized in table IX.

Table IX Testing reliability of measurement of change: requested assumptions

1. measured characteristic a. has same definition at each occasion b. can significantly change over time in each individual
2. individual change (positive, negative or null) can be explained by other variable/s (*W*)
3. applied instrument a. is the same, or a matched one, at each occasion b. has same scale, or comparable one, at each occasion
4. reliability of instrument a. is constant over occasions b. does not depend on different kind of change(constancy of reliability of differences in different conditions, defined by variable *W*)

According to the classical theory of measurement both measures (*pretest* and *post-test*) has

- a *true* score (theoretical and abstract); difference between two true scores is the *true change*;
- an *observed* score; difference between two observed scores is the *observed change*.

By convention, positive changes are defined as gains or growth and negative change looses.
Synthetically, the considered variables (Bereiter, 1963) are:

X_{O_j} Observed total score for *pretest* for subject *j*

Y_{O_j} Observed score for *post-test* for subject *j*

X_{T_j} True score for *pretest* for subject *j*

Y_{T_j} True score for *post-test* for subject *j*

D_{O_j} Observed change score for subject *j* $\left(Y_{O_j} - X_{O_j}\right)$

D_{T_j} True change for subject *j* $\left(Y_{T_j} - X_{T_j}\right)$

e_X Error for *pretest*

e_Y Error for *post-test*

e_D Error for change $\left(e_X + e_Y\right)$

Under the assumptions of the classical theory of measurement, the difference can be defined as following:

$$D_{O_j} = \left(Y_{T_j} + e_Y\right) - \left(X_{T_j} + e_X\right) \;\rightarrow\; D_{O_j} = \left(Y_{T_j} - X_{T_j}\right) + \left(e_Y + e_X\right) \;\rightarrow\; D_{O_j} = D_{T_j} + e_D$$

D_{O_j} is assumed to be an estimate for D_{T_j}, that is

$$\hat{D}_{T_j} = D_{O_j}$$

Assuming e_X and e_Y are random and not correlated, the reliability of change $\left(rho_D^2\right)$ is traditionally expressed as

$$rho_D^2 = \frac{\sigma_{T_D}^2}{\sigma_{T_D}^2 + \sigma_{e_D}^2} = \frac{\sigma_{T_D}^2}{\sigma_{O_D}^2} \qquad (6)$$

Consequently, by knowing that

- $\sigma_{T_d}^2 = \sigma_{T_Y}^2 + \sigma_{T_X}^2 - 2\sigma_{T_Y T_X}$
- $\sigma_{O_D}^2 = \sigma_{O_Y}^2 + \sigma_{O_X}^2 - 2\sigma_{O_Y O_X}$

rho_D^2 can be defined as

$$rho_D^2 = \frac{\sigma_{T_D}^2}{\sigma_{O_D}^2} = \frac{\sigma_{T_Y}^2 + \sigma_{T_X}^2 - 2\sigma_{T_Y T_X}}{\sigma_{O_Y}^2 + \sigma_{O_X}^2 - 2\sigma_{O_Y O_X}} \qquad (7)$$

Observation of such expression helps to point out the factors that influence rho_D^2:

- relative amount of the two true score variances;
- amount of covariance;
- change in post-test variance.

On the other hand, as we know, this approach in order to calculate reliability is not actually applicable since only observed values are available.

Two comparable coefficients for reliability of change can be defined from the (7) as function of observed values and reliability coefficients (rho_X^2 e rho_Y^2)

$$rho_D^2 = \frac{rho_X^2 + rho_Y^2 - 2r_{XY}}{2\left(1 - r_{XY}\right)} \qquad (8)^8$$

$$rho_D^2 = \frac{\sigma_{O_Y}^2 rho_Y^2 + \sigma_{O_Y}^2 rho_X^2 - 2\sigma_{O_Y O_X}}{\sigma_{O_Y}^2 + \sigma_{O_X}^2 - 2\sigma_{O_Y O_X}} \qquad (9)^9$$

Below, another expression for the reliability of difference scores (Zimmerman and Williams, 1982.):

$$rho_D^2 = \frac{rho_X^2 \lambda + rho_Y^2 \lambda^{-1} - 2r_{XY}}{\lambda + \lambda^{-1} - r_{XY}} \qquad (10)$$

where

λ $\quad \sigma_{O_X} / \sigma_{O_Y}$

From (8) we deduce that if intercorrelation is as great as the mean of the reliabilities, rho_D^2 is zero (no reliability in the difference scores). If the intercorrelation is zero, the reliability of the difference scores equals the mean of rho_X^2 and rho_Y^2. This emphasizes the need for univocal uncorrelated scores in differential prediction (Guilford, 1954).

---

[8] *Reliability of Difference Score* (Guilford, 1954, pp 393-394).

[9] Lord, 1963. Notice that when $\sigma_X^2 = \sigma_Y^2$ then

$$rho_D^2 = \frac{rho_Y^2 - 2r_{XY} + rho_X^2}{2\left(1 - 2r_{XY}\right)} = \frac{\frac{rho_Y^2 + rho_X^2}{2} - r_{XY}}{1 - r_{XY}} \qquad (9a)$$

even though the assumption is rarely met by empirical data, especially with regard to growth functions.

Bereiter (1963) observes that a reliability estimate for difference scores tends to be lower than that for the component scores and that measuring change by change-score presents three problematic issues:

1) *reliability paradox*: high reliability of change scores is inversely related to magnitude of pretest-posttest correlation;
2) *scaling problem*: comparability of changes from different initial levels is questionable;
3) *negative bias in change-initial status relation*: individuals with lower initial status scores artificially change more.

Since the different operational definition of change (change score, residual score, percent score, rate score) at analytical level are not considered, the classical approach to reliability of change may be applied before calculating the different scores.

Structural modeling approach seems to allow avoiding problems of classical approach by defining change/difference variable at latent variable level; however, it requires adoption of more specific assumptions concerning the conceptual validity of change/difference/residual latent variable and requests a more complex model involving other latent variables, as shown in appendix A.

2.3 SPECIFIC ISSUES IN TESTING RELIABILITY OF CHANGE IN SUBJECTIVE QUALITY OF LIFE: AN APPLICATION

From an applicative point of view, especially in studies concerning subjective perception, many problematic aspects do not find solution in the theoretical literature on reliability of change in subjective measurement. Thus, even if structural modeling seems to be substantially a more satisfying approach in estimating stability and reliability parameters, some questions concerning assumptions of the model, and particularly connected to subjective measurement, deserve to be mentioned.

The approaches commonly applied to estimating stability and reliability parameters by structural modeling, maximum likelihood and normal theory generalized least squares, assume that the measured variables are continuous, have a multivariate normal distribution and are linearly related.[10] It has to be added that reliability models of subjective measurement are often implemented in educational and achievement fields where linear-growth models are frequently adopted (Rogosa et al., 1982; Bejar, 1983; Rogosa and Willett, 1985; Willet, 1988; Laveault, 1994). This can cause some legitimate doubts on adequacy of applications of reliability models in fields, like subjective quality of life studies, where is not always possible to assume linearity (and normality) of phenomena; moreover, even if variables can be assumed underlying continuous, the actual measurement is discrete and defined by different response scale (e.g. labeled scale, rating scale).[11] This means consequently that applications of reliability models often involve violations of the assumptions. In order to find a remedy for normal assumption violation, corrective approaches were developed at analytical levels (West, 1995); on the other hand, also different approaches should have to be defined providing explanatory variables when defining hypothesis of measurement.

In testing reliability of complex instruments, multiple-indicator approach is surely better than one-indicator approach (Asher, 1983; Bartholomew and Knott, 1999); however, in practice, the one-indicator definition is often applied to multi-item instruments. In other words, the total score,

[10] Also classical model is based upon assumption of linearity (McIver and Carmines, 1979).

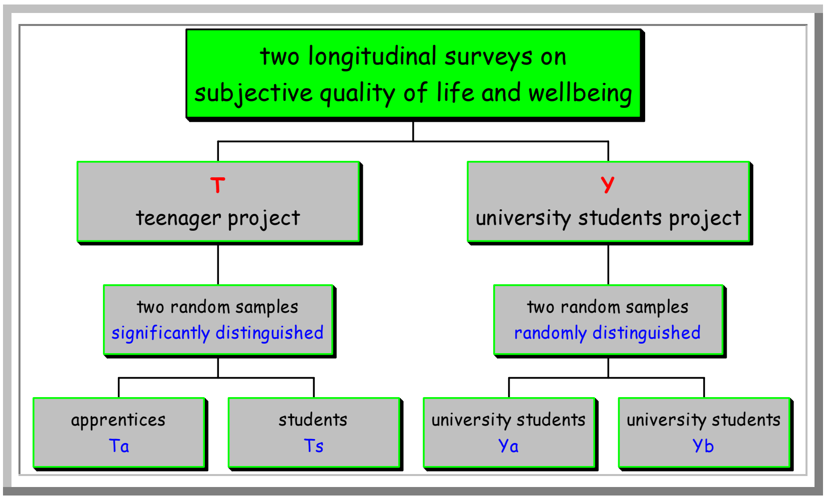
[11] The influence of different capability in discriminating individuals of each different response scale on reliability analysis is not sufficiently known.

obtained by linearly summing up scores of items, is assumed to be a single indicator; this approach allows global reliability evaluation of the instrument in the measurement of change (both in classical and structural models), but fails in detecting the presence of biased items. On the other hand, adoption of multiple-indicator approach to multi-item instrument does not allow a global evaluation of reliability of measurement.
The presentation of an empirical application on longitudinal data, concerning two studies on subjective quality of life and well-being, having involved four different samples of Italian young people, may help to make clear such issues. In particular, presentation of some outcomes regarding the two projects shows some rising issues concerning detection of measurement errors in panel surveys and multi-wave data collections, and, involving also violation of assumptions in reliability analysis.

Survey designs

The two longitudinal survey projects that took place in Florence, Italy (figure 17).

Fig. 17 Two longitudinal survey designs on subjective quality of life



Teenagers project (*T* project). The objective of the longitudinal project was to study well-being in evolutive-aged people; two different samples were drawn from two different Florentine adolescent populations, one composed by students (by provincial education office), and the other composed by apprentices (by social services). Three consecutive surveys on each sample were planned. Table X shows follow-up sample compositions during the three years.

Table X Teenager project: samples compositions

Project	Design	Survey method	Groups	*Surveys*		Panel group composition
T	PANEL (without rotation)	Paper questionnaire	Apprentices	***I (1992)***	531	307
				II (1993)	462	
				III (1994)	307	
			Students	***I (1992)***	215	135
				II (1993)	175	
				III (1994)	135	

A paper-questionnaire was applied on both groups with same conceptual model, same variables but some different organization and formulation of items, with regard to particular areas considering peculiarity of each sample (table XI). The same questionnaire was applied in the three surveys for each sample. Each item has labelled scale.

Table XI Teenager project : questionnaire structure

		Groups	
		STUDENTS	APPRENTICES
Areas	EXTERNAL VARIABLES	Gender	
		Age	
		School curriculum	Employment condition
	FAMILY	Composition and structure	
	LIFE STYLE AND EVENTS	Free time	
		Important personal and family events in last 12 months	
	STUDY	Evaluation of present	Evaluation of past experience
		Perceived present performances	Perceived past performances
	ATTITUDE TOWARDS WORK	Experiences	Experiences
		Wishes	Life changes in consequences of work
			Expectations (6 items scale)
			Involvement (6 items scale)
			Satisfaction (5 items scale)
	INDIVIDUAL TRAITS AND DISPOSITION	**Self-esteem** (Rosemberg scale, 10 items)	
		Attitude towards future (Hopelessness scale, 20 items)	
	ENVIRONMENT AND SOCIAL SUPPORT	Perceived family support (8 items scale)	
		Perceived friends support (8 items scale)	
	WELL-BEING	**General Subjective Well-being** (General Health Questionnaire Scale, 12 items)	
		Anxiety – laxity (Symptom Rating Test, 17 items)	
		Depression – contentment (Symptom Rating Test, 17 items)	
		Somatic symptoms – physical well-being (Symptom Rating Test, 17 items)	
		Hostility – good dispositions (Symptom Rating Test, 17 items)	

University-students project (*Y* project). The objective of the study[12] was to evaluate subjective quality of life of university students (Maggino and Schifini, 2003). Two different samples of students were randomly drawn from student population of the Faculty of Economics of the University of Florence (table XII).

Table XII University student project: samples compositions

Project	Design	Survey method	Groups	*Surveys*		Panel group composition
Y	PANEL (without rotation)	CATI	University students (a)	I (2001)	208	208
				II (2002)	208	
			University students (b)	I (2001)	220	220
				II (2002)	220	

The same questionnaire was applied on both groups, but with two different versions of response scales for each item (table XIII); consequently, two different versions (*a* and *b*) of the same questionnaire were defined. Table XIV shows the two different approaches for each item, with regard to scale reference (agreement, judgment, evaluation), scale type (rating or labeled scales),

[12] Presented panel data concern a greater longitudinal project. Three surveys were carried out for this study in 2000, 2001 and 2002 on three different random samples, drawn from the student population enrolled in at least the third year of degree of the Faculty:
- the first group was made up of 300 students to whom we submitted the paper questionnaire in 2000,
- the second was made up of 498 and 517 students to whom we submitted, respectively, the *a* and *b* CATI questionnaires in 2001,
- the third was made up of 675 students to whom we submitted *c* CATI questionnaire in 2002.

Moreover, we submitted the same version of the questionnaire to a subgroup of students from 2001 samples again in 2002, 208 from the sample of *a* questionnaire students (498) and 220 from the sample of *b* questionnaire students (517).

and scale range (in terms of number of steps for each scale). The project adopted CATI survey method.

Table XIII University student project : questionnaire structure

EXTERNAL VARIABLES		○ **Gender**
		○ **Age**
		○ **University curriculum**
		○ **Employment**
		○ **Distance from University**
INDIVIDUAL TRAITS AND DISPOSITION		○ **Self-esteem** (Rosemberg scale, 10 items)
		○ **Personal motivation towards study**
ENVIRONMENT		○ **Family support**
		○ **Friends support**
VALUES		○ **Importance of particular ambits in one's life**
SATISFACTION AND WELL-BEING PERCEPTION		○ **General Subjective Well-being** (General Life Satisfaction)
		○ **Subjective Well-being in particular life ambits** (Friendship, Family, Money, Free time, Health, Faculty, University career, University friendship)
		○ **Student Life Satisfaction**
		○ **Happiness** (at the present, one year ago)
UNIVERSITY LIFE	Career Performances	○ **Actual Performances** (Successful Examination Number, Taking Examination Number, Marks Average, Proportion of successful exams towards requested standard, Course attendances at the present)
		○ **Perceived Performances** (compared to other students, past expectations, future intentions)
		○ **Attitude towards Performances**
	University evaluation	○ **Faculty Evaluations**
		○ **Exam Perception**

Table XIV University student project: areas, variables, items approaches for the two questionnaires.

Areas	Variables	Questionnaire a: N. of items	a: Reference	a: Type	a: Range	Questionnaire b: N. of items	b: Reference	b: Type	b: Range
UNIVERSITY EVALUATION	Faculty Evaluations	9 (Positive adjectives)	Agreement	Numerical*	1-7	9 (Negative adjectives)	Agreement	Numerical*	1-7
SATISFACTION AND WELL-BEING PERCEPTION	General Life Satisfaction	1	Evaluation	Numerical*	0-10	1	Evaluation	Numerical*	1-7
	Subjective Well-Being in Particular Ambits	10	Evaluation	Numerical*	0-10	10	Evaluation	Numerical*	1-7
	Student Life Satisfaction	1	Agreement	Numerical*	0-10	1	Agreement	Numerical*	1-7
	Happiness at the Present	1	Evaluation	Numerical*	1-7	1	Evaluation	Numerical*	0-10
	Happiness One Year Ago	1	Evaluation	Numerical*	1-7	1	Evaluation	Numerical*	0-10
VALUES	Importance of Particular Ambits in one's Life	16	Judgment	Numerical*	1-7	1	Evaluation	Numerical*	0-10
INDIVIDUAL TRAITS AND DISPOSITIONS	Self-esteem	10	Agreement	Numerical*	1-5	10	Agreement	Numerical*	1-7
	Motivation	10	Agreement	Verbal	1-4	10	Agreement	Verbal	1-4**

** Items verbally anchored.*
*** 1-2 in 2002*

Reliability data analysis

The two projects show differences in objectives and conceptual models; however, the presence of some common elements allows interesting comparison at data analysis level with reference to validation of instruments for measurement of change. Moreover, survey data allow detection of

different performances, in terms of reliability of different scaling methods and different data collection approaches.
In particular, analysis of both panel data, allows comparison of reliability of measurement over time at following levels:

1. reliability of the same instrument submitted by same data collection approach (paper questionnaire with interviewers) to two groups with same age but different characteristics regarding external variables (*students* and *apprentices* – *T* project);
2. reliability of same instrument submitted by same data collection approach (CATI) to two groups with same characteristics regarding external variables (*a* and *b* sample – *Y* project) by different response scales;
3. reliability of the same instrument submitted by different data collection approaches to the two groups of both projects, by different response scales.

The trait allowing a complete level of comparisons (level 3) is *self-esteem*, measured in both projects by Rosenberg's 10-item scale (Rosember, 1965; McIver and Carmines, 1979), submitted by different data collection approaches (paper-questionnaires in *T* project and CATI questionnaires in *Y* project), and by different response scales (4 points labeled scale in *T* project and two different rating scales in *Y* project).
Outcomes, produced by classical and structural (Heise solution)[13] approaches and synthetically shown in table XV, allow two different lines of reading:

a) *Stability estimates*. Apart from the amounts not directly comparable between the two approaches, a clear higher stability in teenager students and an increasing stability between the second and the third survey are noticeable.
b) *Reliability estimates*. By classical approach, we notice a higher reliability value in teenager students group, a stable reliability over time for students groups, revealing a possible education effect or social effect; a higher reliability value in *b* group compared with *a* group, revealing a possible scale effect. By structural approach, we notice a higher reliability value in university students group, revealing different effects (age/education, scale effect or survey effect), and almost same reliability values between university students groups, revealing a lack of scale effect.

Table XV Self-esteem measured by Rosemberg scale: reliability and stability estimates (T and Y project).

Project	Group	Survey Method	Response Scale		analysis							
					Classical approach					Structural approach		
					Reliability			Stability		Reliability	Stability	
					I	II	III	II-I	III-II		II-I	III-II
T (teenagers)	apprentices	Paper Questionnaire	Labeled	4	0.71	0.79	0.79	0.56	0.61	0.81	0.68	0.75
	students				0.84	0.87	0.87	0.58	0.69	0.71	0.87	1.00
Y (university students)	a	CATI Questionnaire	Rating	1-5	0.74	0.74		0.58		0.88	0.65	
	b			1-7	0.79	0.77		0.56		0.87	0.65	

[13] Wiley and Wiley approach produced comparable outcomes; particularly, estimates seem to be averaged by estimates yielded by Heise solution. Moreover, we try to apply the structural modeling approach to a two-wave model in spite of its weakness, as previously reported from literature.

In order to detect presence of biased items, both multiple-indicator structural modelling and internal consistency analysis approaches[14] were performed; here, differently from previous outcomes concerning single-indicator approach, both approaches reveal almost same outcomes and produce almost same, or comparable, values of reliability for each item; in particular, the same item appears to be weak in all groups in both projects by both data analysis approaches.[15]

In project *T*, 'events' variable, concerning the presence, if any, of important personal or family events (positive or negative) between two surveys, was considered. The assumption is that such variable allows controlling the eventual nonlinearity of the relation between two occasions.

Outcomes reveal better estimates of reliability by structural modeling approach; in particular, teenager apprentices improve their reliability level (from 0.81 to 0.98) even in comparison with teenager students groups (from 0.71 to 0.74). It seems that the lower level of reliability recorded by such group, before introducing the 'events' variable, is imputable not simply to education/ social effect, as previously assumed, but rather to reliability analysis, planned without taking into account the nonlinearity of the variable over time.

Another issue rising from the presented outcomes concerns finding and evaluating source/s of error in measurement procedure. In fact, the reliability analysis approaches do not allow evaluating the weight, in terms of measurement error, of each source. Important and interesting attempts of measurement error assessment for each source were proposed (Groves, 1989; Biemer et al., 1991); each proposed model produces reliability estimates for one distinct possible source (questionnaire, respondent, method of data collection, interview process). However, the complex nature of measurement requires other particular (and maybe expensive) experimental designs allowing reliability estimates for each source in a unique model, especially in the field of panel survey and multi-wave data collections.

Moreover, the objective of reliability analysis should not be only to assess method effect but also to improve interpretation of data, especially in meta-analyses fields. From this point of view, relevance of statistical significance of method effect is less important than evaluation of its size; this kind of evaluation may allow definition of possible correction factors.

The presented paper has no pretence to give definitive answers but only to point out the need of more attention and efforts oriented towards definition of more explicatory approaches to reliability assessment in the measurement of subjective change, crucial in quality of life studies.

[14] Internal consistency analysis was performed by four indexes: correlation of the item with the total score, item reliability index (item-total correlation times standard deviation), item-total correlation if the item is excluded from the total, value of alpha coefficient if the item is excluded from the scale.

[15] The great amount of tables required in order to show outcomes concerning multiple-indicator and internal consistency analysis did not suggest their presentation in this paper.

A. RATIONAL OF STRUCTURAL APPROACH IN PANEL DATA ANALYSIS

Generally, study of change, in terms of process analysis, are concentrated not only on change of one variable at two or more occasions but can be finalized to estimate correlation between
- true change of one variable and a third variable, assumed to be stable over time,
- true change of, at least, two variables, assuming that change in one construct is accompanied by change in another. (Schnabel, 1996).[1]

This approach is also known as *Cross-Lagged Panel Analysis*, **CLPA** (Shingles, 1985).
All CLPA approaches attempt to draw inferences on causal relation from cross-lagged associations. In particular, depending on the kind of measured variables and time lag between two measurements of two variables (*X* and *Y*), the analysis of stability of variables over time can be tested through the estimation of two different kind of effect at the same time (Schnabel, 1996):
- *cross-lagged effect* of Y_1 on X_2 (measured by $r_{x_1y_2}$ and $r_{y_1x_2}$), also known as cross-lagged association,
- *autoregressive effect* of X_1 on X_2 (measured by $r_{x_1x_2}$, and $r_{y_1y_2}$), also known as lagged or *diachronic* association.

Cross-sectional, or *synchronous*, associations (measured by $r_{x_1y_1}$, and $r_{x_2y_2}$) are of secondary interest in CLPA, useful primarily as aids in interpreting cross-lagged relationships (Shingles, 1985).
Usually, autocorrelations (autoregressive effect) are much higher than crossed correlations (cross-lagged effect). Structural equation approach allows estimating the two effects and, consequently, avoiding ambiguous interpretation of the comparison between correlation values, since it takes difference in stability into account. The full cross-lagged model (**2W2LV, model,** Two-Wave Two-Latent-Variable) can be defined (in relation to a measurement framework) and represented as in figure 18 (extendible to more complex models; Schnabel, 1996).
Raykov defined an alternative approach (Schnabel, 1996) in which an unbiased estimate of the correlation between pretest and change is possible by applying the autoregressive approach to a difference structure model, in which a latent difference factor is defined. This identified **DS-model** (Difference Score model) creates a dummy variable, representing the estimated true difference score (Δ) (figure 19 – Schnabel, 1996).

[1] In this perspective, models involving more than one variable over time are focused not on reliability of change of one variable but rather on estimating the real change of one variable with regard to another one; assessing reliability of change of one variable have secondary relevance.

Fig. 18 Cross/lagged Two/Wave Two/Latent /Variable model

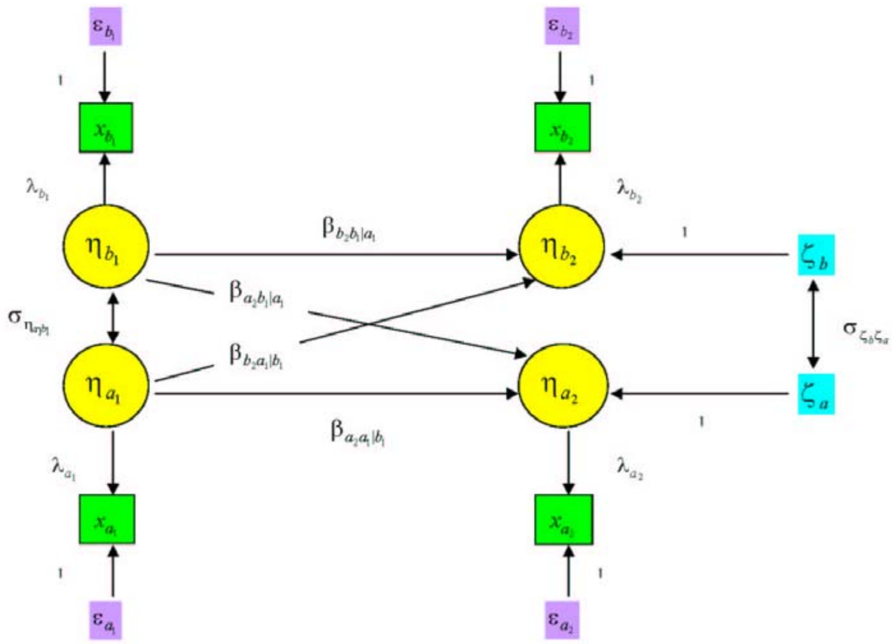
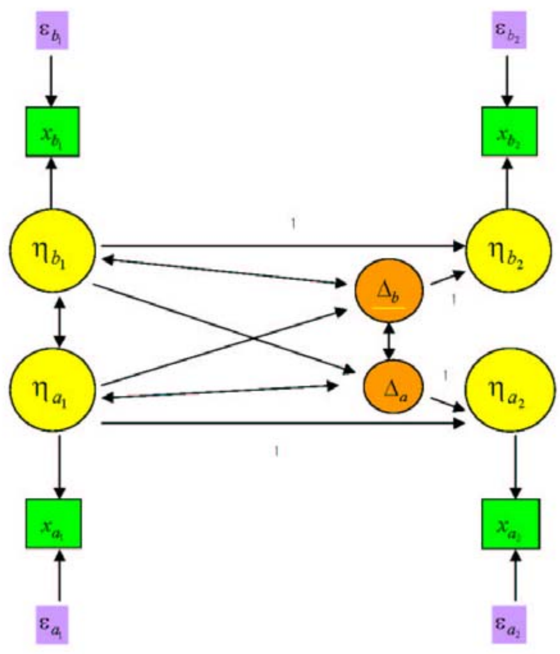
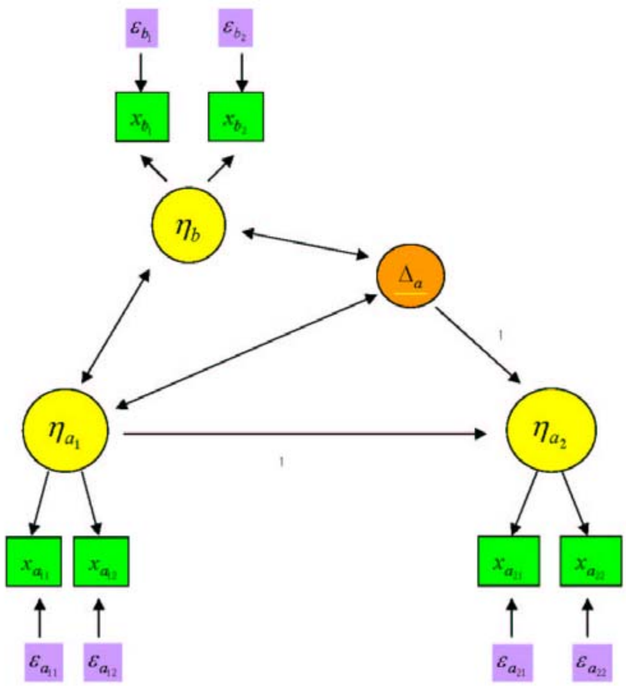


Fig. 19 Difference Score model



DS-model can be defined in terms of one changing variable and one stable variable as can be seen in figure 20 (Schnabel, 1996).

Fig. 20 Latent Difference Score Model for one latent variable (with multiple-indicators) related to another variable.



An alternative autoregressive model (Schnabel, 1996) presents residuals (**RS-model**) instead of dummy difference latent variables, which appear redundant; Schnabel (1996) presents an alternative identification of **RS-model** model, which differs from the full cross-lagged model respect to cross-lagged relations. Both versions are represented respectively in figure 21 and figure 22 (Schnabel, 1996).
The choice between models is not always clear, however it depends on theoretical assumptions on relation between true difference and true latent variable at first occasion $(\sigma_{\eta_1\Delta})$. If the assumption is (Schnabel, 1996):

- $\sigma_{\eta_1\Delta} = 0$ → any difference between the model variants will be due to sampling fluctuations; the **2W2LV-model** is preferable (easier to specify and more robust with respect iteration problems)
- $\sigma_{\eta_1\Delta} \neq 0$ (growth studies) → **DS-model** maybe useful if pretest-change covariance can be partialled out;
- $\sigma_{\eta_1\Delta} \neq 0$ (growth studies) → **RS-model** maybe useful if pretest-change covariance can not be partialled out.

In any case **2W2LV-model** (autoregressive model) seems to be the preferable one for its easy specification. The other one can be interesting to apply in the initial phase of analysis in order to test the existence of pretest-change covariances (Schnabel, 1996).

Fig. 21 Residual Score model (a)

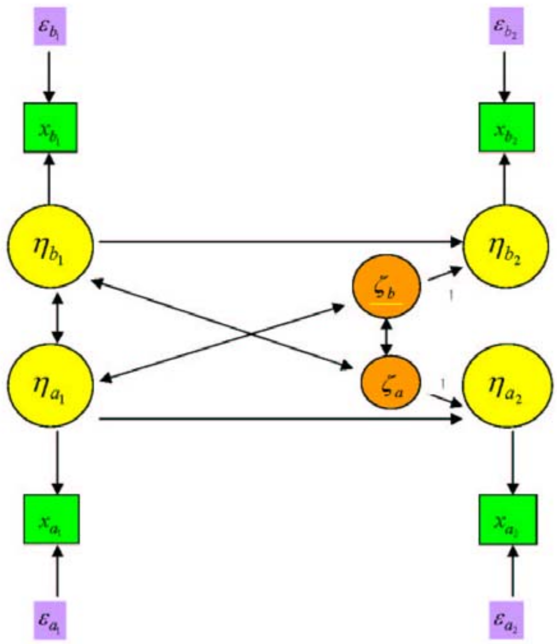
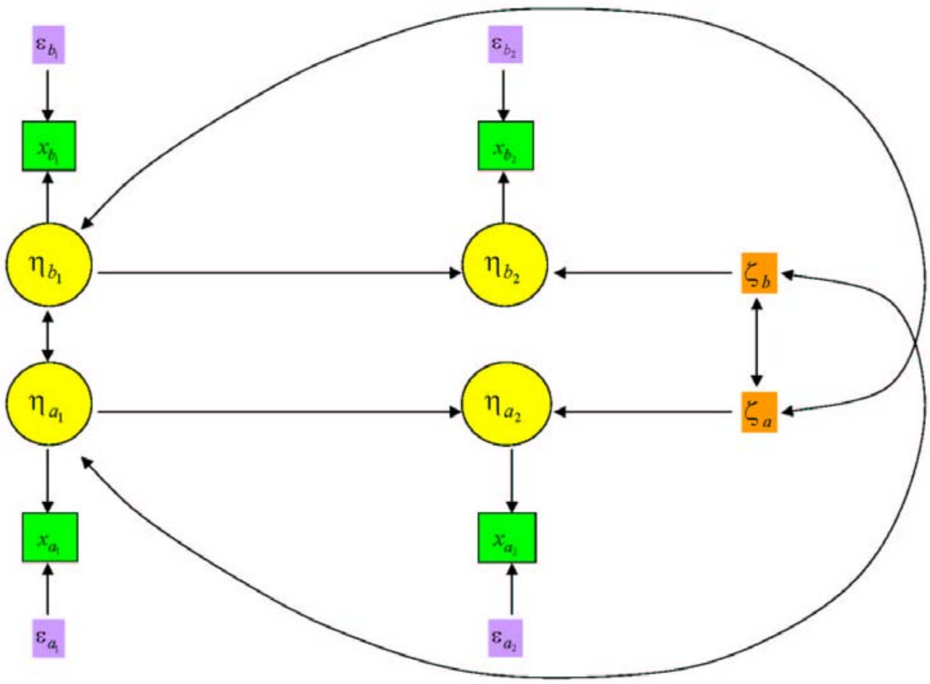


Fig. 22 Residual Score model (b)



References

Asher, H.B.: 1983, Causal modelling (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-003, Newbury Park, CA:Sage).

Bartholomew, D.J.: 1996, The Statistical Approach to Social Measurement (Academic Press, San Diego – London).

Bartholomew, D.J. and M. Knott: 1999, Latent Variable Models and Factor Analysis (Arnold, London-Sydney-Auckland).

Bejar, I.I.: 1983, Achievement testing. Recent Advances (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-036, Newbury Park, CA:Sage).

Bereiter, C.: 1963, 'Some Persisting Dilemmas in the Measurement of Change', in C.W: Harris (ed.), Problem in Measuring Change (University of Wisconsin Press, Madison).

Biemer, P.P., R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman: 1991, Measurement Errors in Surveys (John Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore).

Biemer, P.P. and S. Lynne Stokes: 1991, 'Approaches to the modelling of measurement error', in P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman (eds.), Measurement Errors in Surveys (John Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore).

Blalock, H.M. jr.: 1985, 'Estimating Measurement Error Using Multiple Indicators and Several Points in Time', in H.M. Blalock jr. (ed.) Causal Models in Panel and Experimental Design (Aldine P.C., New York).

Bock, R.D.: 1963, 'Multivariate Analysis of Variance of Repeated Measurements', in C.W. Harris (ed.), Problem in Measuring Change (University of Wisconsin Press, Madison).

Brown, C.: 1995, Chaos and Catastrophe Theories (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-107, Newbury Park, CA:Sage).

Bryk, A.S. and S.W. Raudenbush: 1992, Hierarchical Linear Models: Applications and Data Analysis Methods (Advanced Quantitative Techniques in the Social Sciences Series, Vol. 1, SAGE Pubblications, Thousand Oaks - London – New Delhi).

Campbell, D.T.: 1963, 'From Description to Experimentation: Interpreting Trends as Quasi-Experiments', in C.W. Harris (ed.) Problem in Measuring Change (University of Wisconsin Press, Madison).

Cattell, R.B.: 1963, 'The Structuring of Change by P-Technique and Incremental R-Technique', in C.W. Harris (ed.), Problem in Measuring Change (University of Wisconsin Press, Madison).

Chazel, F., R. Boudon and P. Lazarsfeld: 1970, L'analyse des processus sociaux (Collection Méthodes de la sociologie, Paris, Mouton, La Haye).

Cox, T.F. and M.A.A. Cox: 1994, Multidimensional scaling (Chapman & Hall, London).

Embretson, S.E.: 1994, 'Comparing Changes Between Groups: Some Perplexities Arising from Psychometrics', in D. Laveault, B.D. Zumbo, M.E. Gessaroli and M.W. Boss (eds.), Modern Theories of Measurement: Problems and Issues (Edumetrics Research Group, University of Ottawa, Ottawa, Canada).

Engel, U. and J. Reinecke: 1996, Analysis of Change (Walter de Gruyter, Berlin – New York).

Finkel, S.E.: 1995, Causal analysis with panel data (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-105, Newbury Park, CA:Sage).

Firebaugh, G.: 1997, Analyzing Repeated Surveys (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-115, Newbury Park, CA:Sage).

Forsman, G. and I. Schreiner: 1991, 'The Design and Analysis of reinterview: an overview', in P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman (eds.), Measurement Errors in Surveys (John Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore).

Glenn, N.D.: 1977, Cohort analysis (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-005, Newbury Park, CA:Sage).

Goldstein, H.: 1979, The Design and Analysis of Longitudinal Studies (Academic Press Inc. Ltd., London).

Groves, R.M.: 1989, Survey Errors and Survey Costs (John Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore).

Guilford, J.P.: 1954, Psychometric methods (McGraw-Hill, New York-London).

Hagenaars, J.A.: 1990, Categorical Longitudinal Data (Newbury Park, CA: Sage).

Harris, C.W. (ed.): 1963a, Problem in Measuring Change (University of Wisconsin Press, Madison).

Harris, C.W.: 1963b, 'Canonical Factor Models for the Description of Change', in C.W. Harris (ed.), Problem in Measuring Change (University of Wisconsin Press, Madison).

Heise, D.R.: 1985, 'Separating Reliability and Stability in Test-Retest Correlation', in H.M. Blalock jr. (ed.), Causal Models in Panel and Experimental Design (Aldine P.C., New York).

Holtzman, W.H.: 1963, 'Statistical Models for the Study of Change in the Single case', in C.W. Harris (ed.), Problem in Measuring Change (University of Wisconsin Press, Madison).

Horst, P.: 1963, 'Multivariate Models for Evaluating Change', in C.W. Harris (ed.), Problem in Measuring Change (University of Wisconsin Press, Madison).

Hout, M.: 1983, Mobility Tables (Sage University Paper Series on Quantitative Applications in the

Social Sciences, series no. 07-031, Beverly Hills and London, CA:Sage).

Huckfeldt, R.R., C.W. Kohfeld and T.W. Likens: 1982, Dynamic Modelling. An Introduction (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-027, Beverly Hills, CA:Sage).

Jacoby W.G.: 1991, Data Theory and Dimensional Analysis (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-078, Newbury Park, CA:Sage).

Kaiser, H.F.: 1963, 'Image Analysis', in C.W. Harris (ed.), Problem in Measuring Change (University of Wisconsin Press, Madison).

Laveault, D., B.D. Zumbo, M.E. Gessaroli and M.W. Boss (eds.): 1994, Modern Theories of Measurement: Problems and Issues (Edumetrics Research Group, University of Ottawa, Ottawa, Canada).

Lazarsfeld, P.J. and N.W. Henry: 1968, Latent Structure Analysis (Houghton M. Company, Boston)

Lindsey, J.K.: 1999 (2nd edition), Models for Repeated Measurement (Oxford University Press)

Lord, F.M.: 1963, 'Elementary Models for Measuring Change', in C.W: Harris (ed.), Problem in Measuring Change (University of Wisconsin Press, Madison).

Maggino F. and S. Schifini D'Andrea: 2003, 'Different Scales for Different Survey Methods: Validation in Measuring Quality of University Life', in M.J.Sirgy, D.Rahtz and J.Samli (eds.), Advances in Quality of Life Research (Kluwer Academic Publisher, Dordrecht).

Markus, G.B.: 1979, Analyzing Panel Data (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-018, Beverly Hills and London, CA:Sage)

McIver, J.P. and E.G. Carmines: 1979, Unidimensional Scaling (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-024, Newbury Park, CA:Sage).

Menard, S.: 1991, Longitudinal Research (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-076, Newbury Park, CA:Sage).

Nunnally, J.C.: 1978, Psychometric theory (McGraw-Hill, New York – London).

Rogosa, D.R., D. Brand and M. Zimowski: 1982, 'A growth curve approach to the measurement of change', Psychological Bullettin, n. 90, pp.726-748.

Rogosa, D.R. and B. Willett: 1985, 'Understanding correlates of change by modeling individual differences in growth', Psychometrica, n.50, pp.203-228.

Rosemberg, M.: 1965, Society and the Adolescent Self Image (Princeton University Press, Princeton).

Saris, W.E. and L.H. Stronkhorst: 1990, Causal Modelling in Nonexperimental Research. An introduction to the LISREL Approach (Sociometric Research Foundation, Amsterdam).

Schnabel, K.: 1996, ‘Latent Difference Models’, in U. Engel and J. Reinecke (eds.), Analysis of Change (Walter de Gruyter, Berlin – New York).

Shingles, R.D.: 1985, ’Causal Inference in Cross-Lagged Panel Analysis’, in H.M. Blalock jr. (ed.), Causal Models in Panel and Experimental Design (Aldine P.C., New York).

Shutz, R.W.: 1994, ‘The Measurement, Modelling and Analysis of Change’, in D. Laveault, B.D. Zumbo, M.E., Gessaroli and M.W. Boss (eds.), Modern Theories of Measurement: Problems and Issues (Edumetrics Research Group, University of Ottawa, Ottawa, Canada).

Sullivan, J.L. and S. Feldman: 1981, Multiple Indicators. An Introduction (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-015, Newbury Park, CA:Sage).

Thompson, B. (ed.): 2003, Score Reliability. Contemporary thinking on reliability issues (SAGE Publications, London).

Tucker, L.R.: 1963, ‘Implications of Factor Analysis of Three-Way Matrices for Measurement of Change’, in C.W. Harris (ed.), Problem in Measuring Change (University of Wisconsin Press, Madison).

Visser, R.A.: 1985, Analysis of longitudinal data in behavioural and social research (DSWO Press, Leiden).

Webster, H. and C. Bereiter: 1963, The Reliability of Change Measured by Mental Test Scores’, in C.W. Harris (ed.), Problem in Measuring Change (University of Wisconsin Press, Madison).

Werts, C.E., R.L. Linn and K.G. Jöreskog: 1974, ‘Quantifying Unmeasured Variables’, in H.M. Blalock jr. (ed.), Measurement in the Social Sciences. Theories and Strategies (Aldine Publishing Company, Chicago).

Werts, C.E., K.G. Jöreskog and R.L. Linn: 1985, ‘Comment on “Estimation of Measurement Error in Panel Data”’, in H.M. Blalock jr. (ed.), Causal Models in Panel and Experimental Design (Aldine P.C., New York).

West S.G., J.F. Finch and P.J. Curran: 1995, ‘Structural Equation Models with Nonnormal Variables’, in R. H. Hoyle (ed.), Structural Equation Modeling. Concept, Issues, and Applications (Sage Pubblications, Thousand Oaks).

Wiley, D.E. and J.A. Wiley: 1985, ‘The Estimation of Measurement Error in Panel Data’, in H.M. Blalock jr. (ed.), Causal Models in Panel and Experimental Design (Aldine P.C., New York).

Willet, J.B.: 1988, ‘Questions and answers in the measurement of change’, Review of Research in Education, pp. 345-422.

Zimmerman and Williams: 1982, ‘Gain scores in research can be highly reliable’, Journal of Educational Measurement, 19, p. 149-154.