Method effect in the measurement of subjective dimensions

Dipartimento di Studi Sociali — Via Cavour, 82 — 50129 Firenze

PAPER

Method Effect in the Measurement of Subjective Dimensions

Filomena Maggino

Università degli Studi di Firenze

novembre 2003

Method Effect in the Measurement of Subjective Dimensions

Filomena Maggino
Università degli Studi di Firenze
Dipartimento di Studi Sociali
filomena.maggino@unifi.it

ABSTRACT

This work concerns the exploration and the evaluation of the method effect in terms of the a) survey effect, b) scale effect (scale range, type and reference), and c) survey-scale effect.
The primary goal of the analyses presented, conducted within a study dealing with the quality of university life among students at the Faculty of Economics of the University of Florence (Italy), was to explore and compare different scales in different survey conditions in order to discriminate the individual perceptions of three different traits: 'happiness', 'general life satisfaction' and 'student life satisfaction'. Data, yielded by three different but comparable samples of students, allowed us to find that longer scale of 11-steps discriminated better than 7-steps scale; graphical scales discriminated better than rating scales independently of scale range; paper self-administrated questionnaires discriminated better than telephonic questionnaires.
Second, the study enabled the evaluation of different scale settings in the reproducing of the method effect over time (stability of method effect) and in the reproducing of individual measurement at different times (stability of individual measurement).
In order to improve comparability of different scales, a particular approach oriented to the assessment of the method effect was discussed; however the approach, that tries to find a possible correction-weight for scores yielded by different survey conditions, is applied in ability scaling context and needs to meet a strong assumption in order to be applied in attitude scaling.

1. INTRODUCTION

Individual responses to questionnaires are imputable to different components that can be essentially ascribed to individual and instrumentals components.

The former can be defined in terms of cognitive and personality characteristics, experiences, general attitudes and motivations, accessibility and openness to express them, understanding level, socio-economic and cultural levels; the latter generally concerns structure of questionnaire, subject and wording of the submitted questions, survey technique, structure of response scale (Converse and Presser, 1991; Gilbert, 1993; Biemer et al., 1991). This is particularly true in well-being studies concerning individual perception for which there is no assurance that the answers represent true individual feelings (Andrews and Withey, 1976).

These components yield answers in an undistinguishable way and with different prevalence between different individuals and between moments in the same individual. These makes difficult to understand if the obtained response represents the 'real and true' individual answer.

In any case the response cannot be explained only in psychological terms.

Usually analyses of survey data are concentrated on research objects while analyses of methodological aspects are not always confronted by a critical approach.

The present paper intends to pay attention to only one of the methodological aspects related to the internal validity of applied instruments (Carmines and Zeller, 1992; Traub, 1994), with particular reference to response scales.

The goal of the paper is comparing outcomes yielded by different types of response scale in terms of discriminant capacity. Since samples that present great internal variability do not fit to this kind of goal, the study requires a particular experimental design (Spector, 1990), rather than a hypothesis on psychological frameworks, which allows imputation of observed differences between individual responses to the applied response scales and not to individual differences.

In other words the analyses has to compare groups that use different response scales but that are homogeneous – in terms of experiences, socio-economic and cultural dimensions, of questionnaires structures and of defined questions – and needs more accurate measures in terms of discriminant capacity. This was made possible by data obtained within a study evaluating the quality of life of students at the Faculty of Economics of the University of Florence (Italy).

2. THE STUDY

The assessment of subjective measurement instruments is important in social research in order to identify and select the most sensitive indicators in specific domains. In addition, assessment allows for correct interpretations of data, especially in the meta-analysis area.

This kind of assessment needs special attention not only in the definition and selection of items, but also in the identification of more suitable response scales. These can influence not only the construction and validation of indicators but also the individual responses. This is particularly true in the measurement and assessment of the subjective characteristics like those related with perception of quality of life.

Synthetically, the aspects involved in response scale definition are:

a. scale reference (evaluation, preference, perception, image, judgment)
b. scale type (expression of scale: qualitative – quantitative, verbal – graphical)
c. scale range (number of levels for scale) in the sense of scale discriminant capacity.

Not all the combinations of these aspects can be used in all situations, since other elements also play important roles in response scale definition, such as investigated areas, semantic and cultural meanings and survey methods, making the selection of response scales complicated.

For instance, the choice among all scale forms needs to take into consideration the leading survey type and the questionnaire form, such as paper questionnaire, presence of interviewers or not, Computer Assisted Telephonic Interviews (CATI) (Saris, 1990), Web interviewing, and so on, particularly when adapting items to more than one survey method.

Since we admit the existence of a possible influence of different elements on individual responses, we can assert that individual responses are related not only to individual components but also to other components, such as cultural dimensions, scale definition, survey condition, and trait component. In other words, excluding individual components, individual responses are influenced by several different effects:

1. trait effect,
2. cultural effect,
3. method effect, defined by survey effect and scale effect.

This article concerns the investigation of the method effect in the measurement subjective characteristics related to quality of life. The presented analysis was conducted within a study evaluating the quality of university life at the Faculty of Economics of the University of Florence (Italy). The primary goals of the study were to:

- cross-validate different questionnaires (paper and CATI),
- test the reliability of different scales,

- evaluate impacts of different scores and scale meanings in selection of quality of life indicators (in university context),
- compare individual levels of satisfaction and evaluation.

Three surveys were carried out for this study in 2000, 2001 and 2002. [1]

2. THE SURVEY DESIGN

2.1 Questionnaire

In order to study the method effect we defined two different versions of the questionnaire, with the same conceptual model and the same variables (table 1), referring to two different survey conditions: a self-administrated paper-questionnaire and a CATI-questionnaire.

Table 1 Questionnaire Structure

EXTERNAL VARIABLES		○ **Gender**
		○ **Age**
		○ **University curriculum**
		○ **Employment**
		○ **Distance from University**
INDIVIDUAL TRAITS AND DISPOSITION		○ **Self-esteem**
		○ **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**

The two survey techniques required different item definitions and approaches with regard to scale reference, scale type and scale range:

a. *scale reference*: we turned some particular item references, such as judgment by graphical scales, in forms appropriate to telephonic interviews, by asking students about their agreement regarding some defined assertions;

[1] First results of this study in F.Maggino, S.Schifini D'Andrea: 2003, 'Different Scales for Different Survey Methods: Validation in Measuring Quality of University Life' in M.J.Sirgy, D.Rahtz, J.Samli (eds.) *Advances in Quality-of-Life Theory and Research* (Kluwer Academic Publisher, Dordrecht).

b. *scale type*: we changed the graphical and labeled scales of the paper-questionnaire into equivalent rating scales for telephonic interviews; for instance, in the paper-questionnaire, students evaluated their student life by the Cantril Ladder Scale (Larsen et al., 1985), in graphical form, while in the CATI-questionnaires students had to refer their agreement regarding an assertion about their student condition;
c. *scale range*: one of the hypotheses raised regarding rating scales concerns the discriminant capacities for scales with a different rating width; in order to test this hypothesis, we defined different scale ranges for our questionnaires by assigning different scale amplitude alternatively to questionnaires.

The combination of the different item characteristics allowed us to set up three different telephonic questionnaires. Consequently, we set up four questionnaires forms: one for the paper techniques (labeled *p*) and three different versions of the telephonic ones (labeled respectively *a*, *b* and *c*). Appendix A shows the four versions of the questionnaire and table 2 summarizes the overall design, showing for each variable the scales we used in the paper-questionnaire and the CATI-questionnaires. In particular, it shows the variables used for each area, the number of items defined for each variable and the scale reference, type and range used for each item.

Table 2. Different scale reference, type and range for paper- and CATI-questionnaires.

Areas	Variables	Paper-Questionnaire				Cati-Questionnaire											
						a				b				c			
		N. of items	Reference	Type	Range	N. of items	Reference	Type	Range	N. of items	Reference	Type	Range	N. of items	Reference	Type	Range
UNIVERSITY EVALUATION	Faculty Evaluations	23	Image	Graphical (no numerical reference)	1-7	9 (Positive adjectives)	Agreement	Numerical*	1-7	9 (Negative adjectives)	Agreement	Numerical*	1-7	9 (5 positive and 4 negative adjectives)	Agreement	Numerical*	1-7
SATISFACTION AND WELL-BEING PERCEPTION	General Life Satisfaction	1	Evaluation	Numerical	0-10	1	Evaluation	Numerical*	0-10	1	Evaluation	Numerical*	1-7	1	Evaluation	Numerical*	0-10
	Subjective Well-Being in Particular Ambits	10	Evaluation	Numerical	0-10	10	Evaluation	Numerical*	0-10	10	Evaluation	Numerical*	1-7	10	Evaluation	Numerical*	0-10
	Student Life Satisfaction	1	Judgment	Graphical (Self Anchoring Ladder Scale)	1-9	1	Agreement	Numerical*	0-10	1	Agreement	Numerical*	1-7	1	Agreement	Numerical*	0-10
	Happiness at the Present	1	Judgment	Graphical (Face Scale)	1-7	1	Evaluation	Numerical*	1-7	1	Evaluation	Numerical*	0-10	1	Evaluation	Numerical*	0-10
	Happiness One Year Ago					1	Evaluation	Numerical*	1-7	1	Evaluation	Numerical*	0-10	1	Evaluation	Numerical*	0-10
VALUES	Importance of Particular Ambits in one's Life					16	Judgment	Numerical*	1-7	1	Evaluation	Numerical*	0-10	1	Evaluation	Numerical*	0-10
INDIVIDUAL TRAITS AND DISPOSITIONS	Self-esteem	10	Agreement	Verbal	1-4	10	Agreement	Numerical*	1-5	10	Agreement	Numerical*	1-7	10	Agreement	Numerical*	0-10
	Motivation	10	Agreement	Verbal	1-5	10	Agreement	Verbal	1-4	10	Agreement	Verbal	1-4**	10	Agreement	Verbal	1-7

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

2.2 Samples

Three different random samples were 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).

2.3 Data analysis

The principal goal of data analysis presented here is to explore and compare different scales in different survey conditions in order to evaluate the existence of the method effect, in particular in discriminating individual perceptions for three different traits: 'happiness', 'general life satisfaction' and 'student life satisfaction'.

Item approaches for each variable and for each group are the followings:

- *Happiness*: students expressed their level by one of the following approaches:
  - 7 points agreement-rating scale, in the *a* CATI-questionnaire,
  - 11 points agreement-rating scale, in the *b* and *c* CATI-questionnaires,
  - *Face Scale* (7 expressions), in the paper-questionnaire.[2]

  In the CATI-questionnaires students expressed their happiness level regarding both the present and the past year.
- *General life satisfaction*: students referred their agreement as regards an assertion concerning the trait by one of the following rating scales with different ranges:
  - 11 points agreement-rating scale (from 0, *at all*, to 10, *completely satisfied*) in CATI-questionnaires, *a* and *c*, and the paper-questionnaire,

[2] *Face scale* is one of the most applied and considered single item measures of happiness (Fordyce, 1988; Larsen et al., 1985), having revealed a high level of validity (Andrews and Withey, 1976). Notice that its expressions show an inverted direction as regards the rating scales of CATI-questionnaires. In order to compare distributions, we reversed the *face-scale* codes.

- 7 points agreement-rating scale (from 1, *at all*, to 7, *completely satisfied*) *b* CATI-questionnaire,

➢ *Student life satisfaction*: students expressed their judgments about this trait by one of the following approaches:

- Self Anchoring Ladder Scale by Cantril (9 steps), in the paper-questionnaire[3],
- 11 points agreement-rating scale, in the *a* and *c* CATI-questionnaires,
- 7 points agreement-rating scale, in the *b* CATI-questionnaire.[4]

The characteristics of the sample design allowed different levels of comparison; the identified groups for these possible levels of analysis are presented in table 3.

Table 3. Samples compositions for the three surveys

Statistical model	Questionnaire type	Survey Year	Group-dimension	Group-label
Independent samples	Paper	2000	300	p-group
	CATI *a*	2001	498	a-group
	CATI *b*	2001	517	b-group
	CATI *c*	2002	675	c-group
Dependent samples	CATI *a*	2001	208	a-I-group
		2002		a-II-group
	CATI *b*	2001	220	b-I-group
		2002		b-II-group

In particular, in order to test the method effect we identified five different kinds of comparison:

1. In order to analyze the *survey effect*, we compared groups in different survey conditions using the same scale (same type, same reference and same range) for the same trait ('general life satisfaction'); this comparison is defined for independent samples: *p* group vs. *a* and *c* groups;
2. In order to analyze the *range-scale effect*, we compared groups in the same survey condition but using different scale ranges for the same traits ('happiness', 'happiness one year ago', 'general' and 'student life satisfaction') in the same year (*a* group vs. *b* group) and in a different year (*a* and *b* groups vs. *c* group);

[3] *Ladder Scale* revealed a high level of validity in almost all domains concerning subjective well-being and better performances of seven-point rating scales (see Andrews and Withey, 1976). Further, Andrews and Withey recommended Ladder scale especially because its easiness of application.

[4] We decided to compare 7-steps scale and 11-steps scale for some reasons:
- scales with less than 7 steps are well-known considered poorly discriminant (Biemer et al., 1991; Lyubomirsky and Lepper , 1999);
- the defined levels of the longer scale (from 0 to 10) are directly related to decimal system, familiar to all individuals;
- the intention to leave a midpoint step (even if the presence or absence of a midpoint score was and is object of debate within the survey group).

3. In order to analyze the combined *survey-scale effect*, we compared groups (*a-, b, c* and *p*-groups) in different survey conditions using different scales (different reference, different type and different range) for the same traits ('happiness', 'general' and 'student life satisfaction');
4. In order to analyze the reproducing of *method effect* for the four considered variables over time, we compared scale performances for the two different dependent samples (*a*-I-group vs. *b*-I-group and *a*-II-group vs. *b*-II-group); this analysis mainly treated scale range performances;
5. In order to analyze the reproducing of individual measurement by using the same scale type at different times, we compared individual responses in 2001 and 2002, taking into account the possible presence of individual change; statistically, this kind of comparison is defined in terms of the 'test-retest' approach and is measured by the correlation coefficient interpreted as a stability coefficient; the comparisons were possible for all the variables considered.

Statistical evidence for the first three analyses can be derived from the observation and evaluation of the different discriminant capacities of the scales. The analysis of the discriminant capacity (Osterlind, 1983) was made mainly in terms of both the graphical representations and statistical parameters observed for the quantitative data (from first moment to skewness and kurtosis indexes) of standardized scores. In this context, we used these statistical tools as indexes of the discriminant capacity. In particular, we considered the observation of a high kurtosis value an indicator of an inadequate scale extension. In order to evaluate the discriminant level for each scale we applied the following discrimination index for single item:

$$\delta_i = \frac{N^2 - \sum_{j=1}^{k} f_j^2}{N^2}$$

where

δ_i discriminant coefficient for item *i*

N dimension of the sample

f_j frequency of *j*-th score

k number of scores of item *i*

We derived this coefficient from the discrimination index for multi-item scale defined by Guilford (1954):

$$\delta = \frac{(n+1)\cdot\left(N^2 - \sum_{i=1}^{n}\sum_{j=1}^{k} f_{ij}^2\right)}{nN^2}$$

where *n* is the number of items of the test and f_{ij} the frequency of *j*-th score for *i*-th item.

The coefficient varies from zero, when all individuals make the same score, to 1.0 when the distribution is rectangular.

Moreover, since we had to test the distribution form[5], and not the central tendency, we used the Kolmogorov-Smirnov non-parametric test (KS) in order to compare and verify the existence of different frequency distributions; the KS tests whether two independent samples come from the same population by comparing the two cumulative frequency distributions. The test assumes that the maximum difference between the two cumulative distributions is not significant at the defined alpha value (1%).

2.4 Testing comparability among groups

In order to test the real possibility of comparing the defined groups in the evaluation of the method effect in subjective measurement, we tested the statistical significance of the difference between samples as regards external variables applying the proper statistical test for independent samples (parametric or non-parametric test depends on measurement level and distribution shapes). None of the variables considered registered a significant difference at the defined α value (0.01). Moreover, students that form dependent samples registered no change in external variables from 2001 to 2002.

3. RESULTS

3.1 Survey effect

The comparison of *a*- and *c*-groups as to *p*-group allowed us to test the presence of the survey effect in the measurement of general life satisfaction by the same item approach (10 points agreement-rating scale).

The immediate observation of the three distributions and the analysis of descriptive statistical indexes allows us to notice the same kind of distribution, in terms of skewness, revealing the same satisfaction state for the three groups (table 4).

[5] Testing validity of measures by the technique of analysis of distribution forms was applied also other studies (see the classical and comprehensive work by Andrews and Withey, 1976, evaluating the measures of well-being).

Table 4. Survey effect in the measurement of life satisfaction: frequency distributions, statistical indexes and significant differences between observed distributions.

CATI-Questionnaire				Paper-Questionnaire	
a		c			
RATING SCALE	PCT	RATING SCALE	PCT	RATING SCALE	PCT
0		*0*		*0*	
1		*1*		*1*	
2	0.2	*2*	0.3	*2*	0.7
3	1.2	*3*	0.9	*3*	2.0
4	1.2	*4*	0.7	*4*	2.4
5	3.0	*5*	4.0	*5*	7.1
6	8.7	*6*	10.0	*6*	13.9
7	25.7	*7*	27.9	*7*	30.4
8	35.4	*8*	34.9	*8*	22.0
9	15.8	*9*	10.1	*9*	14.5
10	8.9	*10*	11.1	*10*	7.1

	CATI-Q.		Paper-Q.
	a	c	
Skewness	-0.7	-0.5	-0.5
Kurtosis	1.5	1.2	0.6
δ	0.77	0.77	0.80

Kolmogorov-Smirnov test: two-sided probabilities		
	a-group	c-group
a-group	.	
c-group	0.796	.
p-group	0.000	0.003

A more analytical examination reveals a very small difference between the two CATI groups but a significant difference between the distributions of these groups and the *p*-group, in term of kurtosis index and discriminant coefficient values. In particular, *p*-group distribution appears less concentrated than the other two. This outcome is confirmed by the statistically significant difference registered by the KS.

This first outcome seems to indicate the presence of the survey effect; in particular, the paper questionnaire, on the same scale approach, reveals a better individual discrimination.

3.2 Scale effect

As pointed out above, in this experience the analysis of scale effect is possible essentially in terms of scale-range effect, comparing groups in the same survey condition (telephonic questionnaire) and in the measurement of the same traits (‘happiness’, happiness one year ago’, general life satisfaction’ and ‘student life satisfaction’).

Happiness. The observation of almost the same form for the three distributions (table 5) allows us to conclude that almost all the students expressed a high happiness level. After careful examination, the two groups using the same rating-scale range (*b*- and *c*-groups) show almost the same dicriminant capacity level, in spite of different kurtosis values, which appears higher than the value registered by the third group (*a*-group using shorter rating scale); this outcome is confirmed by the statistically significant difference registered by the KS between both ‘longer rating scale’ groups and *a*-group.

Table 5. Scale range effect in the measurement of happiness: frequency distributions, statistical indexes and significant differences between observed distributions

CATI-Questionnaire					
a		b		c	
RATING SCALE	PCT	RATING SCALE	PCT	RATING SCALE	PCT
		0	0.6	*0*	0.1
1	0.6	*1*	0.4	*1*	
2	0.6	*2*	0.8	*2*	0.3
3	3.0	*3*	0.8	*3*	0.6
4	9.9	*4*	1.5	*4*	1.5
5	29.1	*5*	6.0	*5*	7.0
6	39.9	*6*	10.1	*6*	8.0
7	16.8	*7*	27.1	*7*	26.2
		8	36.9	*8*	35.4
		9	8.3	*9*	11.9
		10	7.5	*10*	9.1

	CATI-Q.		
	a	b	c
Skewness	-0.9	-1.3	-0.7
Kurtosis	1.6	3.8	1.6
δ	0.72	0.76	0.77

Kolmogorov-Smirnov test: two-sided probabilities			
	a-group	b-group	c-group
a-group	.		
b-group	0.000	.	
c-group	0.000	0.436	.

Happiness one year ago. Observed outcomes (table 6) confirm the preceding analysis.

Table 6. Scale range effect in the measurement of happiness (one year ago): frequency distributions, statistical indexes and significant differences between observed distributions

CATI-Questionnaire					
a		b		c	
RATING SCALE	PCT	RATING SCALE	PCT	RATING SCALE	PCT
		0	0.6	*0*	0.6
1	1.8	*1*	0.8	*1*	1.0
2	6.3	*2*	1.9	*2*	0.9
3	6.9	*3*	3.1	*3*	1.8
4	13.6	*4*	4.3	*4*	4.0
5	25.9	*5*	9.1	*5*	8.6
6	30.1	*6*	17.7	*6*	14.0
7	15.3	*7*	23.2	*7*	24.7
		8	24.1	*8*	27.7
		9	8.4	*9*	11.2
		10	6.8	*10*	5.5

	CATI-Q.		
	a	b	c
Skewness	-0.8	-0.8	-1.0
Kurtosis	0.0	0.9	1.7
δ	0.79	0.83	0.82

Kolmogorov-Smirnov test: two-sided probabilities			
	a-group	b-group	c-group
a-group	.		
b-group	0.000	.	
c-group	0.000	0.159	.

General life satisfaction. Observed outcomes (table 7) confirm the preceding analyses although the assignation of rating-scale ranges to each group presents some differences (exchange of scale range between *a*- and *b*-group).

Table 7. Scale range effect in the measurement of life satisfaction: frequency distributions, statistical indexes and significant differences between observed distributions

CATI-Questionnaire					
a		b		c	
RATING SCALE	PCT	RATING SCALE	PCT	RATING SCALE	PCT
0				*0*	
1		*1*	0.8	*1*	
2	0.2	*2*	1.5	*2*	0.3
3	1.2	*3*	2.3	*3*	0.9
4	1.2	*4*	6.6	*4*	0.7
5	3.0	*5*	23.2	*5*	4.0
6	8.7	*6*	38.7	*6*	10.0
7	25.7	*7*	26.9	*7*	27.9
8	35.4			*8*	34.9
9	15.8			*9*	10.1
10	8.9			*10*	11.1

	CATI-Q.		
	a	b	c
Skewness	-0.7	-1.3	-0.5
Kurtosis	1.5	2.3	1.2
δ	0.77	0.72	0.77

Kolmogorov-Smirnov test: two-sided probabilities			
	a-group	b-group	c-group
a-group	.		
b-group	0.000	.	
c-group	0.796	0.000	.

Student life satisfaction. Again, the observed outcomes (table 8) confirm the preceding analyses; the assignation of rating-scale ranges is the same as 'general life satisfaction'.

The observed outcomes seem to point out the presence of a scale-range effect. At this point we need to test the presence of a combined survey and scale effect.

Table 8. Scale range effect in the measurement of student life satisfaction: frequency distributions, statistical indexes and significant differences between observed distributions

CATI-Questionnaire					
a		b		c	
RATING SCALE	PCT	RATING SCALE	PCT	RATING SCALE	PCT
0	3.6			*0*	1.8
1	0.4	*1*	7.0	*1*	2.2
2	2.3	*2*	10.8	*2*	4.5
3	5.2	*3*	18.8	*3*	3.9
4	8.2	*4*	21.7	*4*	8.2
5	20.1	*5*	26.2	*5*	20.8
6	17.4	*6*	9.0	*6*	17.7
7	20.9	*7*	6.5	*7*	20.1
8	14.2			*8*	14.6
9	4.0			*9*	2.1
10	3.8			*10*	4.2

	CATI-Q.		
	a	b	c
Skewness	-0.6	-0.1	-0.5
Kurtosis	0.7	-0.6	0.3
δ	0.85	0.82	0.85

Kolmogorov-Smirnov test: two-sided probabilities			
	a-group	b-group	c-group
a-group	.		
b-group	0.000	.	
c-group	0.999	0.000	.

3.3 Survey and scale effect

The comparison between all the groups allows us to test the presence of a combined effect 'survey-scale' using a different scale approach in the measurement of general life satisfaction, happiness and student life satisfaction.

General life satisfaction. The comparison between the four groups involves two survey techniques (self-administrated paper and telephonic questionnaires) and two different rating-

scale ranges (7 points and 11 points). Analyzing descriptive statistical indexes (table 9) we can observe that groups using the longer rating scale (*a*, *c* and *p*-group) show the same kind of distribution shape (low concentration, long tails and low kurtosis value) while the group using the shorter rating scale (*b*) registers a compression of extremely high scores (high skewness value in comparison of the other groups).

Table 9. Survey-scale effect in the measurement of life satisfaction: frequency distributions, statistical indexes and significant differences between observed distributions

CATI-Questionnaire						Paper-Questionnaire	
a		b		c			
RATING SCALE	PCT	RATING SCALE	PCT	RATING SCALE	PCT	RATING SCALE	PCT
0				*0*		*0*	
1		*1*	0.8	*1*		*1*	
2	0.2	*2*	1.5	*2*	0.3	*2*	0.7
3	1.2	*3*	2.3	*3*	0.9	*3*	2.0
4	1.2	*4*	6.6	*4*	0.7	*4*	2.4
5	3.0	*5*	23.2	*5*	4.0	*5*	7.1
6	8.7	*6*	38.7	*6*	10.0	*6*	13.9
7	25.7	*7*	26.9	*7*	27.9	*7*	30.4
8	35.4			*8*	34.9	*8*	22.0
9	15.8			*9*	10.1	*9*	14.5
10	8.9			*10*	11.1	*10*	7.1

	CATI-Q.			Paper-Q.
	a	b	c	
Skewness	-0.7	-1.3	-0.5	-0.5
Kurtosis	1.5	2.3	1.2	0.6
δ	0.77	0.72	0.77	0.80

Kolmogorov-Smirnov test: two-sided probabilities			
	a-group	b-group	c-group
a-group	.		
b-group	0.000	.	
c-group	0.796	0.000	.
p-group	0.000	0.000	0.003

The observation of *p*-group outcomes reveals the possible presence of the combined effect (survey-scale-range effect) since it shows a better discriminant capacity in comparison with all other groups, independently of the rating-scale ranges used.

Moreover, students with the longer rating scale did not use low score points. The lack of extremely low values in longer rating scales can be interpreted as a clear positive group trend (nobody expressed a very low life satisfaction); the shorter rating scale does not allow us to reach the same conclusion, although students, using the shorter scale, show the same trend. In other words, longer rating scales are more efficient in individual measurement than the shorter rating scale, which, in this case, seems unable to discriminate among extreme levels of satisfaction.

Happiness. The comparison between the four groups involves two survey techniques (self-administrated paper and telephonic questionnaires), two different rating-scale ranges (7 points and 11 points). Notice that the two survey techniques caused the use, for this trait, of two different scale types (numerical and graphical) and different scale references (evaluation and judgment).

The four observed distributions show different scale intensities (different distributions of frequencies among scale values). The differences (table 10) are statistically significant (according to the KS results) except between the groups with the same scale range (*b* and *c*).

Table 10. Survey-scale effect in the measurement of happiness: frequency distributions, statistical indexes and significant differences between observed distributions

CATI-Questionnaire						Paper-Questionnaire	
a		b		c			
RATING SCALE	PCT	RATING SCALE	PCT	RATING SCALE	PCT	FACE SCALE	PCT
		0	0.6	*0*	0.1		
1	0.6	*1*	0.4	*1*		*Face 7*	1.7
2	0.6	*2*	0.8	*2*	0.3	*Face 6*	2.7
3	3.0	*3*	0.8	*3*	0.6	*Face 5*	8.8
4	9.9	*4*	1.5	*4*	1.5	*Face 4*	18.6
5	29.1	*5*	6.0	*5*	7.0	*Face 3*	34.6
6	39.9	*6*	10.1	*6*	8.0	*Face 2*	25.4
7	16.8	*7*	27.1	*7*	26.2	*Face 1*	8.1
		8	36.9	*8*	35.4		
		9	8.3	*9*	11.9		
		10	7.5	*10*	9.1		

	CATI-Q.			Paper-Q.
	a	b	c	
Skewness	-0.9	-1.3	-0.7	-0.7
Kurtosis	1.6	3.8	1.6	0.5
δ	0.72	0.76	0.77	0.77

Kolmogorov-Smirnov test: two-sided probabilities			
	a-group	b-group	c-group
a-group	.		
b-group	0.000	.	
c-group	0.000	0.436	.
p-group	0.000	0.000	0.000

Moreover, when comparing CATI-questionnaire groups, we see a greater dispersion among extreme scores for the longer scales; this may mean that students using the shorter scale had to compress their expressions; this is particularly evident among low scores, because of the strong concentration along the high happiness levels registered for all students. As we can see, *b* group distribution appears more concentrated (high kurtosis value) and with a long tail in correspondence with low happiness levels.

The comparison between paper-questionnaire group and *a*-group distributions, using the same range scale, allows us to compare different scale types (respectively graphical and rating) and different scale references (respectively judgment evaluation). While they registered almost the same skewness values, but the kurtosis values were different, revealing a less concentrated distribution for *face-scale*. The low kurtosis value of *face scale* is indicative of the better discriminant capacity of this scale. Since we cannot assume different psychological conditions, *face-scale* outcomes seem to reveal a better individual 'identification' of happiness perceptions; in this perspective, score 3 for rating scale may be perceived as the more serious position than the corresponding position 5 on the face scale and face 1 may be perceived as 'too pleased' in comparison with score 7.

The comparison of outcomes registered by the four-item versions allows us to infer that the discriminant capacity is related first to scale-type and then to scale-range.

Student life satisfaction. The comparison between the four groups involves two survey techniques (self-administrated paper and telephonic questionnaires), three different rating-scale ranges (7 points, 9 points and 11 points). Notice that, for this trait too, the two survey techniques implied the use of two different scale types (numerical and graphical) and two different scale references (agreement and judgment).

The analysis allows us to highlight (table 11), even if this is less clear, the better capacity of longer scales in discriminating extreme agreement/disagreement levels. As we can see, the extreme scores for the groups using the longer scale (*a* and *c* groups) show a greater dispersion; this can mean that students using the shorter scale had to compress, once again, their expressions, especially in low scores (higher frequency values for this group compared with low frequency values for the other two groups).

Table 11. Survey-scale effect in the measurement of student life satisfaction: frequency distributions, statistical indexes and significant differences between observed distributions

CATI-Questionnaire						Paper-Questionnaire	
a		b		c			
RATING SCALE	PCT	RATING SCALE	PCT	RATING SCALE	PCT	LADDER SCALE	PCT
0	3.6			*0*	1.8		
1	0.4	*1*	7.0	*1*	2.2	*Step 1*	0.7
2	2.3	*2*	10.8	*2*	4.5	*Step 2*	1.3
3	5.2	*3*	18.8	*3*	3.9	*Step 3*	4.0
4	8.2	*4*	21.7	*4*	8.2	*Step 4*	9.3
5	20.1	*5*	26.2	*5*	20.8	*Step 5*	24.3
6	17.4	*6*	9.0	*6*	17.7	*Step 6*	20.7
7	20.9	*7*	6.5	*7*	20.1	*Step 7*	26.3
8	14.2			*8*	14.6	*Step 8*	9.3
9	4.0			*9*	2.1	*Step 9*	4.0
10	3.8			*10*	4.2		

	CATI-Q.			Paper-Q.
	a	b	c	
Skewness	-0.6	-0.1	-0.5	-0.3
Kurtosis	0.7	-0.6	0.3	0.1
δ	0.85	0.82	0.85	0.98

Kolmogorov-Smirnov test: two-sided probabilities			
	a-group	b-group	c-group
a-group	.		
b-group	0.000	.	
c-group	0.999	0.000	.
p-group	0.000	0.000	0.000

A particular anomalous outcome can be observed for *a* group: score 0 shows a proportionally high frequency in comparison with the low scores of other groups.

Once again the results of discriminant coefficient values and KS are concordant, confirming the possible presence of a scale-type effect that is not directly related to survey effect.

The different amounts in discriminant values registered by groups using the same scale approach (type, reference, range scale) suggest the possible existence of a trait effect. This was, however, difficult to evaluate with our data.

3.4 In/stability of method effect

In order to test the reproducing of scale performances and effects over time, we applied three different analysis approaches:

a. comparison of surveys data for each of the two groups considered, in terms of discriminant capacity (reproducing of scale-range effect),
b. comparison of surveys data for each student in terms of correlation, interpreted as the stability coefficient in test-retest model (reproducing of individual measurement); this

analysis was completed by applying structural equation model for panel studies in order to distinguish between stability and reliability parameters.

3.4.1 Reproducing of scale effect

The comparison between the two surveys data for each group allows us to observe clearly that the two groups registered almost the same scale performances for the considered variables and for the two surveys (the observed differences are not statistically significant).
The two groups reproduce same relation between scale performances and discriminant capacities that we saw in the previous analyses (tables from 12 to 15).

Table 12. Happiness: frequency distributions and descriptive statistical indexes (dependent samples)

a			b		
RATING SCALE	2001	2002	RATING SCALE	2001	2002
			0	0.9	0.5
1		1.0	*1*	0.5	0.5
2			*2*	1.4	
3	3.4	3.9	*3*	1.4	2.3
4	8.7	11.6	*4*	1.4	2.7
5	35.6	45.4	*5*	4.5	4.1
6	37.5	39.1	*6*	10.9	10.0
7	14.9	15.5	*7*	21.4	27.3
			8	44.1	35.5
			9	6.4	11.4
			10	7.3	5.9

	a		b	
	2001	2002	2001	2002
Skewness	-0.4	-1.0	-1.6	-1.3
Kurtosis	0.0	2.6	4.3	3.2
δ	.70	.70	.74	.77

Table 13. Happiness (one year ago): frequency distributions and descriptive statistical indexes (dependent samples)

a			b		
RATING SCALE	2001	2002	RATING SCALE	2001	2002
			0	0.5	0.5
1	1.9	1.0	*1*	0.9	0.9
2	6.8	3.4	*2*	1.8	1.4
3	5.3	6.8	*3*	2.8	2.7
4	12.1	9.7	*4*	5.0	4.5
5	30.4	40.6	*5*	9.6	11.4
6	28.0	24.2	*6*	18.3	19.1
7	15.5	14.5	*7*	20.6	25.5
			8	26.1	21.4
			9	7.8	9.1
			10	6.4	3.6

	a		b	
	2001	2002	2001	2002
Skewness	-0.8	-0.8	-0.7	-0.8
Kurtosis	0.3	0.7	0.8	1.1
δ	0.78	0.76	0.83	0.83

Table 14. Life satisfaction: frequency distributions and descriptive statistical indexes (dependent samples)

a			b		
RATING SCALE	2001	2002	RATING SCALE	2001	2002
0					
1			*1*	0.9	2.3
2		0.5	*2*	3.2	
3	1.0		*3*	1.4	1.4
4	1.0	0.5	*4*	7.3	6.4
5	1.9	2.4	*5*	22.7	23.2
6	8.7	8.2	*6*	39.5	46.4
7	27.4	28.8	*7*	25.0	20.5
8	38.9	37.0			
9	11.5	11.5			
10	9.6	11.1			

	a		b	
	2001	2002	2001	2002
Skewness	-0.5	-0.4	-1.4	-1.7
Kurtosis	1.5	1.7	2.4	4.9
δ	0.74	0.75	0.72	0.69

Table 15. Student life satisfaction: frequency distributions and descriptive statistical indexes (dependent samples)

a			b		
RATING SCALE	2001	2002	RATING SCALE	2001	2002
0	2.5	1.9			
1		1.4	*1*	6.8	10.5
2	1.0	4.8	*2*	11.4	6.4
3	4.5	6.2	*3*	13.7	17.3
4	9.5	2.4	*4*	24.2	19.5
5	20.4	26.4	*5*	26.0	29.1
6	21.9	18.3	*6*	11.4	10.9
7	19.4	19.7	*7*	6.4	6.4
8	14.4	11.1			
9	4.0	2.9			
10	2.5	4.8			

	a		b	
	2001	2002	2001	2002
Skewness	-0.6	-0.4	-0.2	-0.3
Kurtosis	1.1	0.4	-0.6	-0.6
δ	0.84	0.84	0.82	0.82

3.4.2 Reproducing of individual measurement

The variables observed register higher correlation values between the two surveys for the group using the longer rating scales (table 16) except for ‘general life satisfaction’. This is also valid in the case of correlation between happiness perception in 2001 and past year happiness perception in 2002.

Table 16. Stability coefficient values between the two surveys for each variable.

Variables	**Scale characteristics**		***a* group**		***b* group**
Happiness (2001 vs. 2002)	reference	evaluation	0.360	evaluation	0.569
	type	*numerical*		*numerical*	
	range	*1-7*		*0-10*	
Happiness one year ago (2001 vs. 2002)	reference	*evaluation*	0.386	*evaluation*	0.404
	type	*numerical**		*numerical**	
	range	*1-7*		*0-10*	
General life satisfaction (2001 vs. 2002)	reference	*evaluation*	0.542	*evaluation*	0.563
	type	*numerical*		*numerical*	
	range	*0-10*		*1-7*	
Student life satisfaction (2001 vs. 2002)	reference	*agreement*	0.405	*agreement*	0.351
	type	*numerical*		*numerical*	
	range	*0-10*		*1-7*	
Happiness (2001) vs. happiness one year ago (2002)	reference	evaluation	0.326	evaluation	0.555
	type	*numerical*		*numerical*	
	range	*1-7*		*0-10*	
	reference	evaluation		evaluation	

Of course, this analysis of scale performances is affected by a bias: a possible individual change occurring between the two surveys. In order to control this bias, we measured the individual perception of change in life satisfaction in 2002 with respect to 2001.[6] This variable allows us to distinguish three subgroups:

1. students who perceived a worsening (16 in *a* group and 19 in *b* group),
2. students who did not perceive any change (86 in *a* group and 73 in *b* group),
3. students who perceived an improvement (106 in *a* group and 128 in *b* group),

in life satisfaction.

The exiguous dimension, for both *a* and *b* groups, suggests the exclusion of the first subgroup from interpretation.

As expected, the observation of the other two subgroups allows us to notice higher correlation values in the 'unchanged' subgroup in both *a* and *b* groups (table 17 and figures 1 and 2).

Table 17. General life satisfaction: stability coefficient values between the two surveys for the three groups defined by different past perceptions.

Variables	**Scale characteristics**	***a* group**				***b* group**			
		Scale characteristics	1. negative change	2. no change	3. positive change	Scale characteristics	1. negative change	2. no change	3. positive change
General life satisfaction (2001 vs. 2002)	**reference**	*evaluation*	0.007 (0.042)	0.761 (0.607)	0.528 (0.501)	*evaluation*	0.875 (0.759)	0.661 (0.537)	0.435 (0.303)
	type	*numerical*				*numerical*			
	range	*0-10*				*1-7*			

4. students which perceived a worsening (16 in *a* group and 19 in *b* group),
5. students which didn't perceive any change (86 in *a* group and 73 in *b* group),
6. students which perceived a betterment (106 in *a* group and 128 in *b* group)

in general life satisfaction.
In brackets Kendall's tau coefficient values.

[6] Applied questionnaires have no analogous items for happiness and student life satisfaction variables.

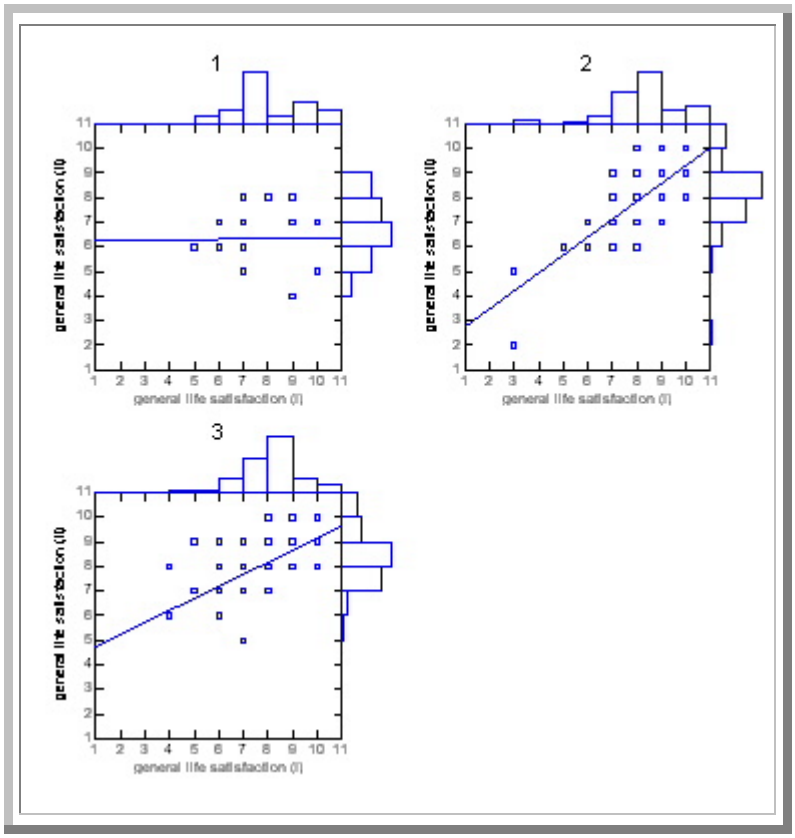


Fig.1 General life satisfaction: comparison between the two surveys scores for the three groups defined by different past perception for a-group.

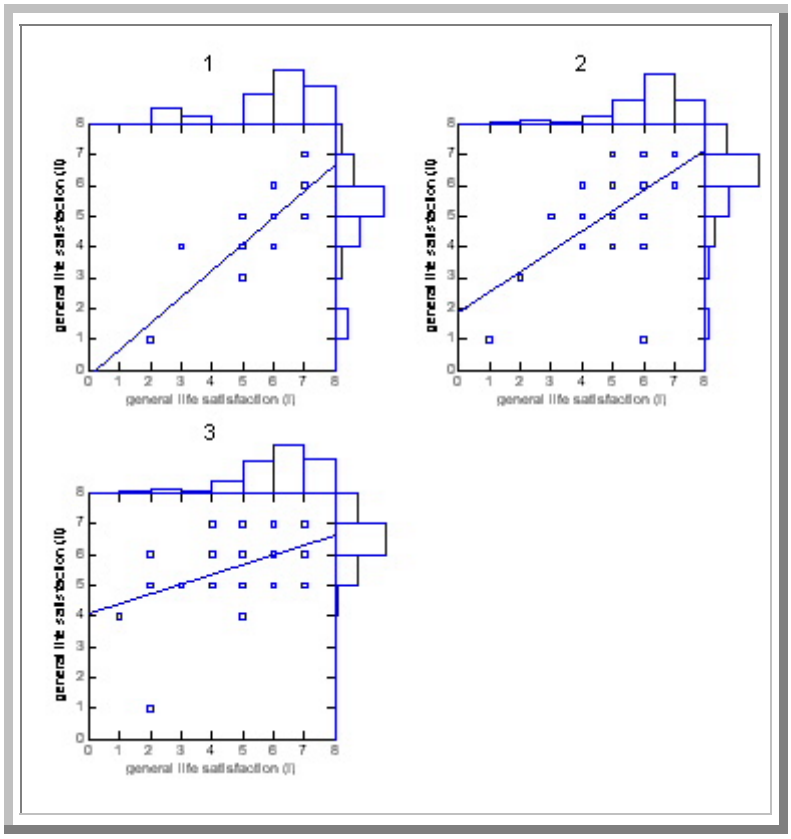


Fig. 2 General life satisfaction: comparison between the two surveys scores for the three groups defined by different past perception for b-group.

Moreover, the higher correlation value observed for the unchanged subgroup using the longer scale (*a*) can be interpreted as evidence of the better discriminant capacity of the longer rating scale. However, this interpretation can present a distortion, represented by other psychological dimensions, playing a possible role with regard to the individual perception of present life satisfaction and of past life satisfaction, which may well modify the possible explanation of these outcomes; as a result, these dimensions need to be analyzed with a different approach.

Finally, application of structural equation modeling in order to compare the two surveys data for each group allows us to distinguish between stability and reliability parameters; outcomes[7] (table 18) show clearly that higher stability values are observed in combination with lower reliability value. However this analysis does not help us to connect somehow reliability parameter values and discrimination index values.

Table 18. Comparison between discrimination and stability coefficients and stability and reliability parameters (structural equation modelling) for the two dependent groups.

Variables	Group	Survey Method	Rating-Scale range	Discrimination Index		Stability	Structural equation modeling	
							Reliability	Stability
Happiness	a	CATI Questionnaire	1-7	0.70	0.70	0.36	0.67	0.80
	b		0-10	0.74	0.77	0.57	0.89	0.71
General Life Satisfaction	a		0-11	0.74	0.75	0.54	0.90	0.67
	b		1-7	0.72	0.69	0.56	0.91	0.68
Student Life Satisfaction	a		0-11	0.84	0.84	0.41	0.72	0.82
	b		1-7	0.82	0.82	0.35	0.79	0.57

3.5 Statistical assessment of method effect

The outcomes presented show the existence of a method effect even if this effect does not yield a great influence on the global evaluation of the phenomenon (the overall tendency is clearly observable and comparable between different survey method conditions). However, the outcomes can provide suggestions for an aware choice between the different kinds of scales to use in the measurement of sentiment and attitude traits.

Making a synthesis of outcomes, we verified, observing the difference between our scales in terms of discriminant capacity, by using the statistic δ, that longer scales registered higher

[7] In order to achieve identification of the model, Heise approach was applied (Heise, 1985; Finkel, 1995). Assumptions of this procedure are: standardization of latent variables, standardization of observed variables, reliability of observed variable is considered equal for considered points of time. The Wiley & Wiley procedure (Wiley and Wiley, 1985) yielded comparable outcomes.

discrimination levels (scale-range effect). However, the results for the combinations of different scale types and scale ranges showed themselves to be quite different: graphical scales registered the greatest discrimination values (scale-type effect), independently of scale-range. Also, for each variable, paper-questionnaires obtained greater discrimination levels vs. CATI-questionnaires (survey effect). Moreover, a possible trait effect seems to be quite clear if we observe and compare the δ values for each sample.

The observation of the method effect can lead to important considerations about the comparison of outcomes yielded by different method conditions, as these usually occur in international comparison analyses and meta-analyses. In this perspective, we need not only to verify the existence of the method effect, but also to quantify it.

In this perspective, the problem of comparability concerns both individual measures and comprehensive indexes, yielded in different method conditions. The solution of both problems is generally solved only by the simple standardization of values.[8] In order to make individual measures or synthetic indexes, affected by method effect, comparable, the standardization has to be supported by a correction weight.

In this perspective, it could be interesting to quantify method effect testing the different efficiency of scales and evaluating the item capacity. This would allow a direct comparison among outcomes yielded by different scales for the same variable. One of the possible approaches is that of evaluating the weight with regard to the effect of guessing for each item. In order to find a possible correction-weight for scores yielded by different survey methods, we referred to the approach known as *Correction of Proportions of Correct Responses for Chance*.

Guilford (1954) first defined this approach, subsequently re-proposed by Nunnally (1978), in reference to abilities scales, for which it is possible to define or clearly identify a 'correct response' (see Appendix B). On the other hand, the identification of correct response is not possible in the measurement of attitude and sentiment traits; the attempt to convert the notion of 'correctness' into the idea of 'group trend' needs to meet the assumption concerning the modal value that might be influenced by scale range.[9] In our opinion the acceptance of this assumption is almost never possible in study concerning attitude and sentiment traits.

[8] Common approaches used in meta-analyses to convert diverse measures into a common metric (standardization procedures) do not resolve the problem since do not consider the effect of using different scales.

[9] Another approach, based on Item Response Theory, considers chance parameter on individual measurement values, although they refer to the measuring of abilities (Osterlind, 1983).

4. CONCLUSIONS

Synthetically, the experience presented above allowed for the identification of the presence of method effect, which can be distinguished in:

1. The survey effect: the comparison of scores registered in the same scale settings but in different survey conditions (self-administered and telephonic) revealed a better performance, in terms of discriminant capacity, for the rating scale of paper questionnaires;
2. The scale-range effect: longer rating scales revealed a better discriminant capacity than shorter ones and allowed extreme values to be avoided (lack of extreme individual positions);[10]
3. The survey-scale effect: scale-type effect is difficult to distinguish from survey effect since not all scale types can be administrated in both survey techniques;[11] so we tested only the existence of the combined survey-scale effect; in this perspective, graphical scales outcomes revealed a better discriminant capacity than rating scales.[12]

These outcomes suggest the need for further analysis in order to evaluate the influence of the observed effects on the reliability of multi-item variables.[13] The analysis of a reference-scale effect requires a different approach since it is also related to item definition and cultural factors; in this study, the analysis of this particular effect as regard another scaling technique (semantic differential scale) showed that it is possible to identify a different outcome between positive and negative judgment references. This effect needs further study, focusing on comparisons between different cultural contexts.

The analysis of the in/stability of method effect revealed the substantial stability of the different effects previously observed (reproduction of scale effect over time) and a better stability of longer rating scales (reproduction of individual measurement). Observing the stability of scale

[10] Better outcomes of 0-10 points rating scales compared with 1-7 points rating scales seem not to confirm other experiences (Biemer et al., 1991).

[11] In order to unravel and analyze the relation between the two effects, we are planning to carry out a web-survey.

[12] In the study presented we were also able to compare labeled scales and rating scales applied in the measurement of other variables; outcomes revealed better performances of labeled scales in the paper questionnaire context than in the telephonic questionnaire context, but better performances of rating scales than those of labeled scales in the same telephonic context.

[13] In the same study, we were able to analyze the *alpha* value in different method conditions for two particular variables, 'self-esteem' (measured by the Rosenberg ten item scale) and 'motivation towards study' (measured by a ten item scale). This analysis showed (table 19) a remarkable difference between *p*-group and two CATI groups (*a*- and *b*-group), as regards 'motivation towards study' variable, using same item approach (and almost the same scale range), in favour of paper questionnaires. This difference remains, even if to a smaller degree, when CATI questionnaires adopt rating scale instead of labeled scales ('self-esteem' variable). These outcomes seem to suggest that different survey techniques require different item approaches.

effect also presented in the two previously multi-items variables (table 19), the only remarkable difference between the two surveys which emerged was for the *b*-group, which used different scale ranges in the measurement of 'motivation towards study'.

Table 19 Alpha values for two multi-items variables in different method conditions for defined independent (A) and dependent groups (B).

A

Variables	Scale characteristics	a-group		b-group		c-group		p-group	
Self-esteem	**reference**	agreement	0.73	agreement	0.75	agreement	0.77	agreement	0.85
	type	*numerical**		*numerical**		*numerical**		*verbal*	
	range	*1-5*		*1-7*		*0-10*		*1-4*	
	n. of items	*10*		*10*		*10*		*10*	
Motivation	**reference**	*agreement*	0.53	*agreement*	0.58	*agreement*	0.74	*agreement*	0.82
	type	*verbal*		*verbal*		*verbal*		*verbal*	
	range	*1-4*		*1-4*		*1-7*		*1-5*	
	n. of items	*10*		*10*		*10*		*10*	

B

Variables	Scale characteristics		a-I-group	a-II-group		b-I-group	b-II-group
Self-esteem	**reference**	agreement	0.74	0.74	agreement	0.79	0.79
	type	*numerical**			*numerical**		
	range	*1-5*			*1-7*		
	n. of items	*10*			*10*		
Motivation	**reference**	*agreement*	0.48	0.48	*agreement*	0.60	0.43
	type	*numerical**			*verbal*		
	range	*1-4*			*1-4 (2001)* *1-2 (2002)*		
	n. of items	*10*			*10*		

In order to find a correction weight for comparing outcomes yielded by different scale ranges, a possible application of the *Correction of Proportions of Correct Responses for Chance* was considered not acceptable in the context of attitude and sentiment study since it is based on the notion of 'response correctness'.

In conclusion, identifying a rational correction as regards the whole method effect would appear to be a complex task and needs further experimentation and analyses, in terms of reliability, as considered here, but also in the perspective of validity (Zumbo, 1999), whose analysis and discussion was not the object of the presented work.

REFERENCES

Andrews, F.M. and S.B. Withey (1976). *Social Indicators of Well-being: Americans' Perceptions of Life Quality*. Plenum Press, New York-London.

Andrich, D. (1988). *Rasch Models for Measurement*. CA:Sage, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-068, Newbury Park.

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

Carmines, E.C. and R.A. Zeller (1992). *Reliability and Validity Assessment*. CA:Sage, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-017, Newbury Park.

Converse, J.M. and S. Presser (1991). *Survey Questions: Handcrafting the Standardized Questionnaire*. CA:Sage, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-063, Newbury Park.

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

Fordyce, M.W. (1988). 'A review of research on the happiness measures: a sixty seconds index of happiness and mental health'. *Social Indicators Research*, n.4.

Gilbert, N. (1993). Researching *Social Life*. Sage, London.

Guilford, J.P. (1954 – 2nd edition). *Psychometric methods*. McGraw-Hill, New York–London.

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.

Larsen, E.R., J. Diener and R.A. Emmons (1985). 'An Evaluation of Subjective Well-Being Measures'. *Social Indicators Research*. Vol. 17, n.4.

Lyubomirsky, S. and H.S. Lepper. (1999). 'A measure of Subjective Happiness: Preliminary Reliability and Construct Validation'. *Social Indicators Research*. Vol. 46, n.2.

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

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

Osterlind, S.J. (1983). *Test Item Bias*. CA:Sage, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-030, Newbury Park.

Saris, W.E. (1990). *Computer-Assisted Interviewing*. CA:Sage, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-080, Newbury Park.

Spector, P.E. (1990). *Research Designs*. CA:Sage, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-023, Newbury Park.

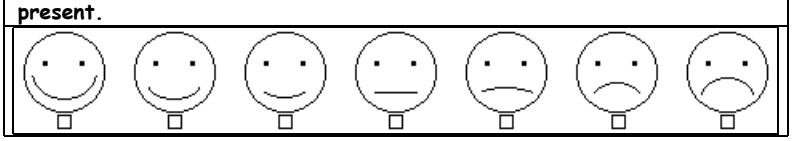
Traub, R.E. (1994). *Reliability for the Social Sciences. Theory and Applications*. SAGE, Measurement Methods for the Social Sciences series, vol. 3, London.

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.

Zumbo, B.D. (ed.) (1999). 'Validity Theory and the Methods used in Validation: Perspectives from Social and Behavioural Sciences'. *Social Indicators Research*. Vol.45, n.1-3.

Appendix A.
The different item approaches for the considered variables

1. Happiness

Paper questionnaire
Point out the face that represents your happiness condition at the present.


CATI- questionnaire – type 'a'							
Using a score between 1 (minimum level) to 7 (maximum level) indicate your happiness level	1	2	3	4	5	6	7
at the present							
one year ago							

CATI- questionnaire – type 'b'											
Using a score between 0 (minimum level) to 10 (maximum level) indicate your happiness level	0	1	2	3	4	5	6	7	8	9	10
at the present											
one year ago											

CATI- questionnaire – type 'c'											
Using a score between 0 (minimum level) to 10 (maximum level) indicate your happiness level	0	1	2	3	4	5	6	7	8	9	10
at the present											
one year ago											

2. General life satisfaction

Paper questionnaire											
Using a score from 0 (at all) to 10 (completely satisfied), can you tell how much are you satisfied for:	0	1	2	3	4	5	6	7	8	9	10
Your general life											

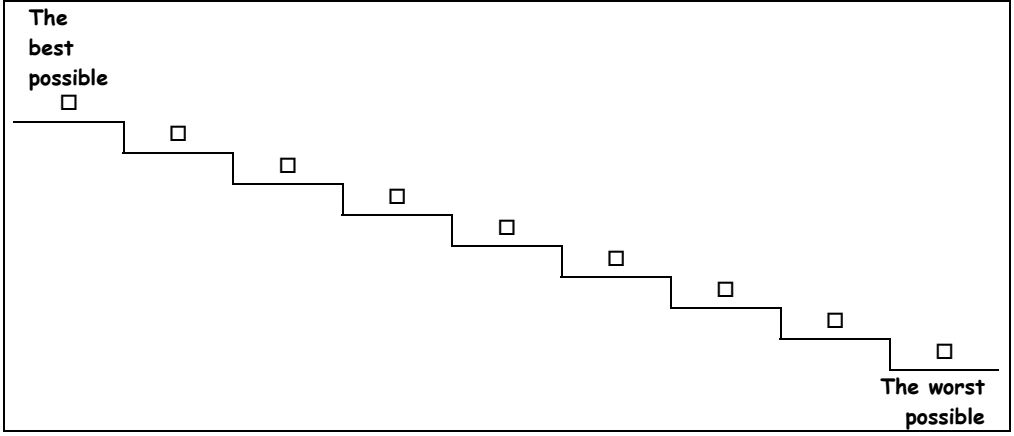
CATI- questionnaire – type 'a'											
Using a score from 0 (at all) to 10 (completely satisfied), can you tell how much are you satisfied for:	0	1	2	3	4	5	6	7	8	9	10
Your general life											

CATI- questionnaire – type 'b'							
Using a score from 1 (at all) to 7 (completely satisfied), can you tell how much are you satisfied for:	1	2	3	4	5	6	7
Your general life							

CATI- questionnaire – type 'c'											
Using a score from 0 (at all) to 10 (completely satisfied), can you tell how much are you satisfied for:	0	1	2	3	4	5	6	7	8	9	10
Your general life											

3. Student life satisfaction

Paper questionnaire
Point out the position that better represents your present student condition



CATI- questionnaire – type 'a'
***I judge my student life the best possible*. How much do you agree with this assertion?** (0 completely disagree – 10 completely agree)

CATI- questionnaire – type 'b'
***I judge my student life the best possible*. How much do you agree with this assertion?** (1 completely disagree – 7 completely agree)

CATI- questionnaire – type 'c'
***I judge my student life the best possible*. How much do you agree with this assertion?** (0 completely disagree – 10 completely agree)

4. General life satisfaction: comparison with last year (presented only in the 2002 survey)

CATI- questionnaires			
How do you feel, in comparison with last year, as regards to:	worse	same	better
Your general life			

Appendix B.
The 'Correction of Proportions of Correct Responses for Chance' approach

The *Correction of Proportions of Correct Responses for Chance* approach, whose basic idea is related to the difficulty item parameter (Osterlind, 1983), is considered important in comparing items with different numbers of alternative responses, since the chance for a 'correct response' is related to the total number of responses.

The procedure allows for the correction of the proportion or count obtained taking into consideration the possibility that an individual with lower abilities may have answered in a 'correct way'; this possibility is considered also as a function of the number of responses (scale range). The corrected proportion of 'correct responses' $\left({}_c p\right)$ can be computed directly from the response count data by applying the following formula:

$$ {}_c p = \frac{R_i - \frac{W_i}{k-1}}{R_i + W_i} $$

where

R_i number of correct responses for item *i*

W_i number of incorrect responses for item *i*

k number of alternative responses.

If all subjects answered the item, then $R_i + W_i = N$ (for *N*=sample size). Negative ${}_c p$ values indicate overcorrection for some reason (Guilford, 1954, p. 422).

After applying ${}_c p$ formula, it is possible to compare the observed proportion $\left(p_i\right)$ with the corrected proportion $\left({}_c p_i\right)$. The difference between these values $\left({}_c p_i - p_i\right)$ represents the amount of correction.

Since the identification of correct response is not possible in the measurement of attitude and sentiment traits, we can apply this approach only if we can assume that the notion of 'correctness' can be converted into the idea of 'group trend'; in other words, we need to assume that the modal value might be influenced by scale range.[14] The approach allows a possible application on Guttman scaling data (Andrich, 1988; McIver and Carmines,1979)

[14] The application of this approach to our data, just in explorative terms, showed that scales with better discriminant values registered low amounts of correction (less than 10 per cent) when shorter scales registered correction values over 10 per cent. These outcomes are constant as regards scale ranges; in fact, we observed almost the same correction values for scales with the same ranges, independently of groups, variables and time.