Does it matter how life satisfaction is measured? - Evidence from a randomized experiment*

Raphael Studer [a] and Rainer Winkelmann [a,b]

[a] *Department of Economics, University of Zurich*
[b] *CESifo, Munich, and IZA, Bonn*

Octobre 2011

Abstract: Self-assessed discrete happiness data have been used for research into the determinants of subjective well-being. This study is the first to implement a continuous single item happiness question in a representative survey. Results of the randomized controlled experiment raise doubts about inferences drawn so far on correlates of happiness. We find distribution distortions for women on the numerically labeled discrete scale. Thus, we are able to explain the widely reported gender happiness inequality puzzle. Further analyses on data quality, correlates and distributions confirm the superiority of the continuous scale.

Keywords: happiness, subjective well-being, life satisfaction, likert scale, visual analogue scale, rating scales, gender inequalities, gender gap

* This paper draws on data of the LISS panel of CentERdata.
Address for correspondence: University of Zurich, Department of Economics, Zürichbergstr. 14, CH-8032 Zürich, Switzerland, ☎ +41 44 634 22 97 and +41 44 634 22 92, ✉ raphael.studer@econ.uzh.ch and rainer.winkelmann@econ.uzh.ch

1 Introduction

Interest in the determinants of subjective well-being has burgeoned in recent years. Research is often based on self-assessed happiness or life satisfaction data. Mostly a single item 11-point Likert scale (LS) well-being question is answered by survey participants. This rating scale seems to be widely accepted and little is done to find better alternatives. We propose a new continuous rating scale to measure individual happiness, the visual analogue scale (VAS). Results of the randomized controlled experiment identify stylized happiness facts as question design artifacts.

Many scholars have been interested in the design of the single item happiness question. It was Fordyce (1978) who proposed to assess happiness on a single item 11-point LS (Likert, 1932). But other rating scales have been tested (Diener, 1994). The effect of labels on LS was examined by Larsen et al. (1984). Cummins (2003) investigated LS of different discriminating powers. Andrews and Crandall (1976) assessed data quality of 7-point LS, faces and ladder scales. However, all these rating scales remained discrete.

Van Praag and Ferrer-i-Carbonell (2004) noted that individuals perceive satisfaction as a continuous phenomenon bounded by the states of complete dissatisfaction and complete satisfaction. Discretization of the underlying true happiness score into a LS score may lead to systematic transformation error. This stands in contrast to a continuous rating scale which may be perceived as a reference continuum of the latent happiness. Such a rating scale is the VAS.

The VAS (Hayes and Patterson, 1921) is simply a bounded line. Respondents assess their happiness by setting a marker on the VAS. The VAS has been extensively used in medical pain research (McCormack et al., 1988). With the development of computer based surveys the accuracy and simplicity of implementation of the VAS jumped up. A literature has recently been developed comparing LS and VAS in computer based experiments (Couper et al., 2006). But to our knowledge there have been only three happiness studies implementing the VAS (Matsubayashi et al., 1992; Bouazzaoui and Mullet, 2002; Hofmans

and Theuns, 2008). All three are small sample paper and pencil vignette studies, which do not propose any counterfactual for the VAS scores. We propose to close this gap.

This paper is the first to implement the VAS for a single item happiness question in a representative survey. A unique randomized controlled experiment enables the identification of question design effects. Thanks to a large set of socioeconomic and sociodemographic variables we can identify heterogeneous mode effects and explain patterns and puzzles which occurred so far when LS data was used. Our analyses conclude in favor of the VAS.

We start our paper by presenting the survey and question design and assessing the quality of the experiment. Section 3 reviews the existing literature on comparison of single item happiness scales and provides estimates for reliability and validity measures. We conclude on equally reliable and valid data quality for both scales. Distributional analyses are presented in Section 4. The experiment shows that the same people are on average less happy when they have to report happiness on the VAS. Moreover, wider spread happiness scores appear for the VAS, which can be explained to be due to an increased likelihood of scoring closer to the extremes. In fact, the unexplained pattern of LS high frequency categories is due to too little discriminating power. Section 5 exploits the existence of two parallel happiness questions to investigate the impact of rating scales on correlates of happiness for a common set of respondents. We find all statistically significant determinants of happiness to correlate stronger in absolute values with the VAS happiness scores than with the LS scores. But the gender gap which is present when LS data is being used vanishes with the application of VAS data. We provide insightful analyses to demonstrate the answer distortions of female respondents present in the discrete measurement. Section 6 concludes.

2 Does the survey design matter?

The randomized controlled experiment used in this paper was implemented in the Longitudinal Internet Studies for the Social sciences (LISS). The LISS panel was established by CentERdata based at Tilburg University in the Netherlands. 10'150 random addresses were drawn from a 10% sample of the Dutch population register. The oldest inhabitant was approached by a letter including 10 Euros. In case of non-response the person was called or visited. 5176 households announced to participate in the survey. Households without broadband internet connection or computer were provided with it. After the first survey year in 2007 73% were found to participate on average (Scherpenzeel, 2009). Knoef and de Vos (2009) concluded on underrepresentation of elderly people and of some ethnicities. In 2009 a refreshment sample stratified by age, ethnicity and household types was successful in establishing representativeness of the LISS panel (de Vos, 2010).

A monthly e-mail invites participants to respond to the LISS panel. Monthly waves consist of three questionnaires. The Background Variable questionnaire needs only be updated if any changes in core socioeconomic or sociodemographic variables, as income, education, age, civil status or household composition, occurred. A second questionnaire contains questions on one of the twelve Core Studies repeated every year, for instance on health or religion. A third questionnaire contains so called Assembled Studies, like the experiment used for this paper. Survey respondents can choose which questionnaire they want to answer first.

The experiment was implemented during the survey months March and April 2011. The link to the assembled study directed participants to a single item happiness question. Answers had to be given either on a LS ranging from 0-9 (10 point) or a VAS (0-9 measured in 100 points). Answers scale were randomly assigned. In the subsequent month the scales were changed or again randomly assigned if people had not answered during the March wave. In the ideal case every survey participant reported its happiness using the VAS and the LS. This crossover design has two advantages. First, the dependent sample increases

power of test statistics. Second, any time effect affecting only subgroups of our sample is captured in both scales.

The crossover experiment is summarized in Figure 1. The experiment consists of 5042 participants in wave 1 and 4795 in wave 2. The paired sample, individuals who responded in March and April 2011, includes 4274 observations. In May 2011, the month after the experiment took place, the LISS Core Study was the personality questionnaire. This personality study gathers not only information on overall happiness on a LS (0-10) but also on personality traits, like emotional stability or self esteem. In this May wave 5230 responses were gathered out of which 3770 had already assessed their happiness in March and April. We will uniquely use data of the March and April waves to quantify differences in means, variances (section 4) and correlates (section 5). Data of the May wave will be used for the assessment of data quality (section 3) only.

Screenshots of the two questions implemented in the experiment are presented in Figure 2. The design of both scales was the same: No questions were asked before the happiness question; the length of both scales was approximately equivalent; the VAS had no default marker to avoid artificial high frequency regions (Treiblmaier and Filzmoser, 2009); both scales were aligned horizontally, however results should not differ to vertical scales (Funke et al., 2010; Paul-Dauphin et al., 1999) and the same anchor words were used for the LS and the VAS, in order to avoid wording effects (Weng, 2004). We expect no hidden factors to drive any differences in responses between the two rating scales.

In order to examine the question design, we asked participants to answer 5 evaluation questions after they assessed their happiness. Difficulty in answering, clearness of the question, degree of thought provocation, interest and joyfulness were rated on a LS ranging from 1 (certainly not) to 5 (certainly yes). Figure 3 shows the distributions by question types to all five evaluation questions. Distributions are very similar and we conclude on absence of design artifacts causing response problems.

Two concerns about the experiment may still be raised. First, screen resolution may

differ among survey participants. A lower resolution will lead to a wider VAS or LS. Empirical findings however suggest no effect of varying length of the VAS (Keindler et al., 2003). Second, people can decide on the order of the three questionnaires each month on their own. Order of questionnaires have been shown to have important effects on answers (Schumann and Presser, 1981), however we will show that it does not affect answers in our survey.

Table 1 and 2 examine the quality of the present experiment. First, table 1 evaluates whether the subsamples are truly random by comparing the means of ex-ante characteristics by question types. We cannot reject equality of means for any of the variables, but the variables out of labor force, working and foreigner. However, the point estimates are very similar in magnitude for the two groups and differ only by 2-3 percentage points for these three variables. If randomization seems not complete in these three variables statistically, it is practically. The picture is similar for the April wave. Second, table 2 reports estimates of the parameters identifying a time or questionnaire order effect. The model presented in equation (1) is estimated using the paired sample.

$$\begin{aligned} s_{it} &= \beta_0 + \beta_1 \cdot april_{it} + \beta_2 \cdot vas_{it} + \beta_3 \cdot april_{it} \cdot vas_{it} \\ &+ \beta_4 \cdot experiment2_{it}^{nd} + \beta_5 \cdot vas_{it} \cdot experiment2_{it}^{nd} + \epsilon_{it} \end{aligned} \tag{1}$$

As dependent variable s_{it} happiness scores are used. The parameter β_1 estimates a time effect, which may be captured by every scale differently through the interaction effect β_3. The variable experiment2_{it}^{nd} takes the value 1 if during the 2 hours preceding the response of the happiness question, the background variable questionnaire was opened. Also this questionnaire order effect may vary by scales through the interaction coefficient β_5. Table 2 shows that neither a time nor a questionnaire order effect exist for one or the other scale. We therefore conclude that the experiment was successfully implemented.

3 Does the VAS matter for data quality?

Many different methods have been used to assess data quality. We start with the most popular, examination of validity and reliability. The true score model will be implemented afterwards and recently developed methods will be employed at the end of this section.

Validity measures the degree of transformation from the true latent happiness h_i into a theoretical score on the rating scale t_i. A systematic error due to a nonconformity of a rating scale harms validity. Intuitively, the LS, requesting the categorization of a continuous feeling, may have lower validity than the VAS. Reliability is the extent to which the transformed score t_i is captured in the observed score s_i. Low reliability is due to a random measurement error. The high sensitivity of the VAS may lead to lower reliability. Different methodologies have been established to investigate validity and reliability of rating scales.

A huge body of literature on reliability and validity of the LS and the VAS can be found in medical pain research. Validity has been tested by the administration of different intensities of heat (Price et al., 1994) or sound (Lara-Munoz et al., 2004) in a randomized order. Survey participants had to rate the felt pain on a LS or VAS. Flint et al. (2000) assessed reliability by letting people judge hunger feelings during two subsequent days while the authors controlled ingested energy. Like these three studies all studies, which we have been reviewing, have not been advising against the VAS.

Happiness cannot be measured objectively. The presence of validity in single item happiness questions has therefore been evaluated through construct or convergent validity (see Diener (1994) for a review). On one hand, convergent validity is the correlation between happiness scores assessed on different rating scales. Our results suggest no difference between the three implemented scales regarding convergent validity. First, the VAS correlates with the Likert 11 and 10 point scales by 0.69, whereas the two discrete measurements correlate by 0.72. Magnitude of these point estimates is in line with earlier findings (e.g., Larsen et al., 1984). This positive correlation may indicate that all three measures assess

the same latent construct, but no more. In fact, convergent validity cannot conclude which scale is best. On the other hand, construct validity is the correlation between happiness scores and traits of individuals. Estimates of correlations between rating scales (columns) and the BIG V inventory (Goldberger, 1992) and the self-esteem scale (Rosenberg, 1965) are reported in Table 3. These six trait variables were gathered in the personality study of the May wave. The time gap between the assessment of happiness (March, April, May) and personality traits (May) causes no problems as personality traits are judged stable (e.g., Srivastava et al., 2003). The levels of correlation we report are similar to what has been found in earlier research (e.g., Larsen et al., 1984 or Abdel-Khalek, 2006). Moreover, we do not find any pattern in lower or higher convergent validity for one or the other scale. However, also convergent validity measurement is not a satisfactory analysis. It relies implicitly on the assumption of valid measurements of the external criterions, which are in this case multiple item LS questions. None of the so far proposed methods is satisfactory. Nevertheless, the VAS has to be considered as valid, if the 10 or 11 point LS are. Moreover, the theoretical argument should be emphasized again. Higher (theoretical) validity should be attributed to the VAS, because the line length acts as a reference continuum to represent perceived happiness.

Reliability of single item happiness questions has been assessed through test-retest reliability. The test-retest method suggests to use the same sample and the same measurement on two occasions. Larsen et al. (1984) and Krueger and Schkade (2008) have concluded on test-retest reliability coefficients ranging from 0.4 to 0.6 for single item discrete measurements. The data structure of this study does not allow to present test-retest reliability coefficients.

Reliability can also be exploited using our experiment. We have shown that randomization was successful and that no time effect exists. Sample distributions in happiness scores should be equal between the March wave and April for each rating scale if the scales are reliable. In order to compare distributions among the two scales, the VAS score was split

up in 10 equal intervals. Figure 4 shows the histograms for both waves for both scales. We observe a huge agreement in the distribution in scores. It could be judged marginally stronger for the VAS. The VAS is consider to be a reliable rating scale for happiness.

A further quality measurement of rating data is provided by classical test theory. The true score model (e.g., Saris and Gallhofer, 2007) is a simpler version of multitrait-multimethod approaches as employed by Andrews and Crandall (1976) for single item happiness questions. Consider the observed score for individual i being a noisy measure of the transformed score: $s_{ij} = t_{ij} + \zeta_{ij}$. If the transformation for every measurement method $j = \{vas, ls10, ls11\}$ is a function of the latent happiness $t_{ij} = v_j \cdot h_i + \eta_{ij}$, then substitution yields $s_{ij} = v_j \cdot h_i + \epsilon_{ij}$. The three parameters of interest v_j^2 are identifiable through the three correlations between different s_{ij}. In fact $corr(s_{i,vas}; s_{i,ls10}) = \frac{v_{ls10} \cdot v_{vas} \cdot Var(h_i)}{\sqrt{Var(s_{i,ls10}) \cdot Var(s_{i,vas})}}$, which reduces to $corr(s_{i,vas}; s_{i,ls10}) = v_{ls10} \cdot v_{vas}$ if equality of variances is assumed. We find the lowest quality for the VAS (0.66) and equal quality for the two discrete measurements (0.71). However, this less sound performance of the VAS should not be weighted too heavily. In fact, identification of the parameters relies on ungentle assumptions. First, equality of variances, which will be shown in the next section to not hold true. Second, a linear transformation function, which leads to a violation of the orthogonality condition ($\mathrm{E}(\epsilon_{ij} \cdot h_i) = 0$). For instance, if $v_j^2 \to 1$ and $h^{max} > s_j^{max}$, than ϵ_{ij} has to be negative in order to make the transformation fit the support of the rating scale. In other words, the estimator of v_j^2 is biased and inconsistent.

Recent computer surveys experimentally implementing the LS and VAS used various methods to examine survey data. Research has recorded item response time and found no difference (Funke and Reips, forthcoming) or a longer response time for VAS (Cook et al., 2001; Couper et al., 2006). Completion rates of questionnaires have been found to be lower and skipped questions to be more if the VAS instead of the LS was used (Couper et al., 2006). Answers were modified nearly twice as often with the VAS (Funke and Reips, forthcoming). However, it is not clear how these new methods should be interpreted. A

longer item response time may reflect more self-reflection or answer problems.

Our data structure does not allow the analysis of all of these indicators. Randomization took only place when participants accessed the questionnaire. Therefore, item non-response cannot be assessed. On the other hand, we also find higher average item response times for the VAS (16 seconds) than for the LS (10 seconds). However, this may be due to a difference in question design: the VAS question had one sentence more to read (Figure 2). A higher fraction of survey participants was found to move back to adjust the happiness score for the VAS (2.3%) than for the LS (1.4%). However, this finding may only reflect that people are less familiar with the VAS than with the LS.

We conclude that VAS happiness data is as valid and reliable as data gathered by a 10 or 11 point LS.

4 Does the VAS matter for the distribution of happiness scores?

Several computer based studies have reported equality in distributions between VAS and LS scores (e.g., Couper et al. 2006 or Funke and Reips, forthcoming). However, these studies have not asked people to assess subjective feelings, as happiness. All of them have used objective judgments, as questions on clothing style or vignettes on behavior. But a paper and pencil study on self reported individual coping reported lower mean values for the VAS (Flynn et al., 2006). Hence, distributions may differ.

Table 4 reports the first and second moments of both happiness scales as well as for the discretized VAS. We present t-tests on the equality of means as well as Levine's test on the equality of variances between the VAS scores and LS scores for each wave and for the paired sample. All three samples show the same picture: lower mean but wider spread happiness scores in the case of the VAS. We can reject all null hypotheses of equality of means and variances. The random assignment of response scales creates a control group

(LS) of individuals that should have the same outcomes as what the treatment group (VAS) would have had if they had answered on the LS. The simple comparison in means therefore suggests that participants would have been 0.45 points happier on the LS than on the VAS. In fact, controlling for a large set of socioeconomic and sociodemographic variables, such as age, household size, employment and marital status, gender, origin and education, does not change the point estimate. This effect is also not only an artifact of the different supports of the scale (continuous vs. discrete). In reality, when the supports are equalized, i.e. the VAS discretized, the difference in means becomes 0.2 points lower, but stays significant and large. For a decomposition of the treatment effect in population subgroups we refer to the next section.

The in Table 4 reported increase in variances of 0.8 points, when moving from the LS to the VAS, is particularly interesting. In fact wider spread happiness scores contain more information. However, higher variance in the VAS scores may simply be due to the high sensitivity of the scale. For instance, it is likely that people would like to cross the equivalent of a 7 but crossed 6,8 instead. The thesis of inexactness can be easily tested. First, if it holds true, than the categorized VAS should have lower variance. But Table 4 shows the contrary: the variance increases by another 0.5 points. Second, an examination of the distributions of the March waves gives more evidence that not the higher sensitivity is the trigger of higher variances. In Figure 6 two patterns can be found. First, response densities are lower in the categories 7 and 8. Second, people are more likely to score close to the two anchors. Higher variance is therefore explained by a drift to the extremes.

The differences in distributions can be quantified. The discretized VAS scores and the LS scores were used to generate three indicator variables. The first took the value 1 if the score was equal to 7 or 8, the second if the score was equal to 1,2 or 3 and the last if the score was equal to 9. For each of these indicator variables a linear probability model was estimated using the paired sample. A large set of socioeconomic and demographic variables as well as a wave dummy and dummies indicating the questionnaire and question order

were included in the regression. Table 5 reports the estimates of the parameter of interest, i.e. the average effect on the probability of scoring in one of these intervals given that the VAS was employed. All effects were found to be high and significant. The probability that a participant scores a 7 or 8 is reduced by over 21 percentage points given that the VAS was used. Moreover, this effect can be divided in the two effects prevailing at the extremes. In reality, in the case of the VAS the probability of a 9 is over 8 percentage points higher and the probability of scoring either a 1, 2 or 3 is over 2 percentage points higher. It can be concluded that the low discriminating power of the LS makes people avoid the extreme categories. The high frequencies of answers at 7 and 8 are therefore an artifact of a too insensitive answer scale.

5 Does the VAS matter for the correlates of happiness?

Research into the determinants of subjective well-being has burgeoned in recent years, and valuable insights have been obtained (see for a review Kahnemann and Krueger, 2006). Among others economists have been interested in the effects of schooling (Orepolus, 2003), income (Easterlin, 1995), unemployment (Winkelmann and Winkelmann, 1998) or age (Stone et al., 2010). Many findings have been replicated for different countries and have been judged as robust (Frey and Stutzer, 2002). All these studies use discrete happiness data. Therefore, the question raises: may these findings not only rely on the properties of the LS?

Table 6 shows estimates by question types of a linear regression modeling happiness scores as the sum of socioeconomic and sociodemographic variables. Results for the LS are in line with the research literature (e.g. Kahnemann and Krueger, 2006 or Frey and Stutzer, 2002). We find that happiness is U-shaped in age, foreigners and unemployed are less and women more happy. Marriage and house ownership, the latter may be interpreted

as a proxy for savings, have a positive effect on happiness. However, these findings do not carry over to the VAS regression.

Comparison of the LS correlation coefficients with those of the VAS sample reveals some striking findings. Signs of statistically significant explanatory variables stay the same. But the equality of coefficients is rejected by a Chow test (p-value=0.012). Effects of statistically significant variables are in absolute values stronger in the VAS regression, which may also be an indicator that individuals perceive the VAS as more intuitive and scores represent their living circumstances better. The most powerful finding, however, is that all but one variables keep their significance. Suddenly the huge gender gap of 0.162 points found in the LS setup vanishes. This is even more astonishing if one keeps in mind that the samples consist of the same persons, simply asked one month apart.

The disappearance of the gender gap brings up the question if some subgroups of the population are influenced to different degrees by the VAS or the LS. In order to identify heterogeneous effects, the difference in happiness scores were computed for every individual. The LS score was substracted from the VAS score and the difference regressed on the same explanatory variables as before. Table 7 reports the estimates of this regression. The constant indicates that the reference group scored on average lower on the VAS. Three variables drive participants to score relatively higher on the VAS than the reference group: Marriage, houseownership and most important being of male sex. Hence, question design affects subgroups of the population differently. But the issue if women are scoring relatively lower or men scoring relatively higher remains open.

To see which part of the distribution drives the heterogeneous effect, we computed three indicator variables. The first took the value 1 if the difference in happiness scores was < -1, the second if it was > 1 and the last if it was in between -1 and 1. For each of these dependent variables a linear probability model was estimated. Estimates for the average effect of gender are presented in Table 8. Women are over 6 percentage points likelier to score more than 1 point lower on the VAS than male counterparts. Men have a 5

percentage points higher probability than women of having minimal changes between the VAS and the LS. This finding suggests that women are the trigger. If everybody scores on average lower on the VAS, women's happiness scores fall even more.

The previous analyses suggest that women overrate their happiness as soon as a numbered scale is used. This is in line with the results of a recent article. Conti and Pudney (2011) found that numerically labeled response categories induced a distortion in the distribution of women's job satisfaction scores. Doubts raise about the reliability of inferences that have been drawn earlier on the gender gap.

If there really exists a gender gap it is an important finding as it may reflect gender inequalities caused by society not measurable by econometricians. But would one not expect female to be less happy? It has been a often disputed topic in the literature. Some studies have found female to be better off (e.g., Gerdtham and Johannesson, 2001; Lalive and Stutzer, 2010) others concluded on equally happy gender (e.g., Fujita et al., 1994). And others said that female were happier decades ago, but that this gap vanished over the last years and should not be present anymore in recent crossectional studies (Stevenson and Wolfers, 2009). In an early attempt, Wood et al. (1989) reviewed nearly 100 studies and concluded that a gender gap was only found in representative surveys when single item happiness questions were used. But what if all these studies finding a gender gap, were simply using a single item LS happiness question?

In fact all the papers, which concluded on a gender gap have been using a LS to assess happiness. We conclude that a gender gap may exist but it is truly an artifact of numerically labeled LS. Therefore, the puzzle of finding females happier even if they are still perceived as being less privileged in western societies is not existing. Once the VAS is used the gender gap vanishes and female and male counterparts are equally off ceteris paribus.

6 Conclusion

Most of the studies interested in determinants of happiness have used discrete satisfaction scores as dependent variables. This may be because they are widely available in crossectional or panel surveys. This paper suggests to move away from the discrete Likert scale. The visual analogue scale, a continuous measurement, was implemented in the Dutch Longitudinal Internet Study for Social sciences. This paper is the first to exploit a randomized controlled experiment to compare a single item happiness question assessed either on a LS or on a VAS. Results are promising. First, survey participants did not manifest problems in using the VAS. Second, no differences in data quality are found between the VAS and LS. Third, lower mean and wider spread happiness scores for the VAS were detected. We have been able to demonstrate that the higher variance is not due to the high sensibility of the VAS but to the increased likelihood of participants of scoring close to the boundaries. This finding explains the high frequency LS categories 7 and 8 as a result of too little discriminating power. The former finding, the lower VAS mean, seems particularly interesting for the third part of our paper, namely the correlation analyses. A gender happiness inequality, i.e. women being on average happier than men, which is economically hardly interpretable, but a robust empirical finding, is identified as an artifact of the LS. Women were found to overrate their happiness on the numerically labeled LS by 0.56 points on average. We conclude that the VAS is preferable to the LS. On one hand, the VAS can be theoretically interpreted as a reference continuum for the latent continuous happiness. On the other hand, the VAS overcomes empirically artifacts of numerically labeled LS like distribution distortions.

References

Abdel-Khalek, A.M., 2006, "Measuring Happiness with a Single-Item Scale", *Social Behavior and Personality*, Vol. 34, No.2, 139-150

Andrews F.M. and R.Crandall, 1976, "The Validity of Measures of Self-Reported Well-Being", *Social Indicators Research*, Vol. 3, 1-19

Bouazzaoui, A.B. and E. Mullet, 2006, "Employment and Family as Determinants of anticipated Life Satisfaction: Contrasting European and Maghrebi People's Viewpoints", *Journal of Happiness Studies* Vol.6, 161185

Conti, G. and S. Pudney, 2011, "Survey Design and the Analysis of Satisfaction", *The Review of Economics and Statistics*, Vol. 93, No. 3, 10871093

Cook, C., F. Heath and R.L. Thompson, 2001, " Score reliability in web- or Internet-based surveys: Unnumbered graphic rating scales versus Likert-type scales", *Educational and Psychological Measurement*, 61, 697-706

Couper, M.P., R. Tourangeau, F. G. Conrad and E. Singer, 2006, "Evaluating the Effectiveness of Visual Analog Scales : A Web Experiment", *Social Science Computer Review*, Vol. 24, No. 227

Cummins, R.A., 2001, "Normative Life Satisfaction: Measurement Issues and a Homeostatic Model", *Social Indicators Research*, Vol.64

de Vos, K., 2010, "Representativeness of the LISS-panel 2008, 2009, 2010", http://www.lissdata.nl, last consultation 14.10.2011

Diener, E., 1994, "Assessing Subjective Well-Being: Progress and Opportunities", *Social Indicators Research*, Vol. 31, No. 2, 103

Easterlin, R., 1995,"Will Raising the Incomes of All Increase the Happiness of All?", *Journal of Economic Behavior and Organisation*, Vol. 27, No. 1, 35-48.

Flint, A., A. Raben, J.E. Blundell and A. Astrup, 2000, "Reproducibility, power and validity of visual analogue scales in assessment of appetite sensations in single test meal studies", *International Journal of Obesity*, Vol24, 38-48

Flynn, D., P. van Schaik and A. van Wersch, 2004, "A Comparison of Multi-Item Likert and Visual Analogue Scales for the Assessment of Transactionally Defined Coping Function", *European Journal of Psychological Assessment*, Vol. 20, No. 1, 4958

Frey, B.S. and A. Stutzer, 2002, "The Economics of Happiness", *World Economics*, Vol. 3, No. 1

Funke, F., U.-D. Reips and R. K. Thomas, 2010, "Sliders for the Smart: Type of Rating Scale on the Web Interacts With Educational Level", *Social Science Computer Review*,

Funke, F., U.-D. Reips, forthcoming, "Why Semantic Differentials in Web-Based Research Should be Made From Visual Analogue Scales and Not From 5-Point Scales", *Field Methods*, Vol 24, No. 3

Fujita, F., E. Diener and E. Sandvik, 1991, "Gender Differences in Negative Affect and Well-Being: The Case for Emotional Intensity", *Journal of Personality and Social Psychology*, Vol.6, No. 3, 427-434

Gerdthama, U.G. and M. Johannesson, 2001, "The relationship between happiness, health, and socioeconomic factors: results based on Swedish microdata", *Journal of Socio-Economics*, Vol. 30, 553557

Goldberger, L.R., 1992, "The Development of Markers for the Big-Five Factor Structure", *Psychological Assessment*, Vol4, No.3, 26-42

Hayes, M. H. S. and D. G. Patterson, 1921, "Experimental development of the graphic rating method", *Psychological Bulletin*, Vol. 18, 98-99

Hofmans, J. and P. Theuns, "On the linearity of predefined and self-anchoring Visual Analogue Scales", *British Journal of Mathematical and Statistical Psychology*, Vol. 61, 401413

Kahnemann, D. and A.B. Krueger, 2006, "Developments in the Measurement of Subjective Well-Being", *Journal of Economic Perspectives*, Vol. 20, No. 1, 324

Knoef M. and K. de Vos, 2009, "The representativeness of LISS", an online probability panel http://www.lissdata.nl, last consultation 14.10.2011

Kreindler, D., A. Levitta, N. Woolridge, and C.J. Lumsdenc, 2003, "Portable moodmapping: the validity and reliability of analog scale displays for mood assessment via hand-held computer", *Psychiatry Research*, Vol.120, 165177

Krueger, B. and D.A. Schkade, "The Reliability of Subjective Well-Being Measures", *Journal of Public Economics*, Vol. 92, 1833-1845

Lalive R. and A. Stutzer, 2010, "Approval of equal rights and gender differences in well-being", *Journal of Population Economics*, Vol. 23, 933962

Lara-Munoz, C., S.P. de Leon, A. R. Feinstein, A. Puentee and C. K. Wells, 2004, "Comparison of Three Rating Scales for Measuring Subjective Phenomena in Clinical Research", *Archives of Medical Research*, Vol. 35, 4348

Larsen, R.J., E. Diener, R.A. Emmons, 1985, "An Evaluation of Subjective Well-Being Measures",*Social Indicators Research*, Vol. 17, No. 1

Likert, R., 1932, "A Technique for the Measurement of Attitudes", *Archives of Psychology*, Vol. 140, 155

Matsubayashi K., S. Kimura,T. Iwasaki, K. Okumiya,T. Hamada,M. Fujisawa, K. Takeuchi,T. Kawamoto and T. Ozawa, 1992, "Application of visual analogue scale of happiness

to elderly Himalayan highlanders", *Nippon Ronen Igakkai Zasshi*, Vol.29, No.(11), 823-828

McCormack, H.M., D.J.L. Horne and S. Sheater, 1988, "Clinical Applications of Visual Analogue Scales: A critical Review", *Psychological Medicine*, Vol. 18, 1007-1019

Oreopoulos, P., 2003, "Do Dropouts Drop Out Too Soon? Evidence from Changes in School-Leaving Laws" Mimeo, University of Toronto, March

Paul-Dauphin, A., F. Guillemin, J.M. Virion and S. Briancon, 1999, "Bias and Precision in Visual Analogue Scales: A Randomized Controlled Trial", *American Journal of Epidemiology*, Vol. 150, No. 10

Price, D.D., F. M. Bush, S. Long and S. W. Harkins, 1994, "A comparison of pain measurement characteristics of mechanical visual analogue and simple numerical rating scales", *Pain*, 56, 217-226

Rosenberg, M., 1965, "Society and the adolescent self-image", Princeton University Press, New Jersey

Saris and Gallhofer, 2007, "Design, Evaluation, and Analysis of Questionnaires for Survey Research", Hoboken, New Jersey

Scherpenzeel, A., "Start of the LISS panel: Sample and recruitment of a probability-based Internet panel", http://www.lissdata.nl, last consultation 14.10.2011

Schuman, H. and S. Presser, 1981, "Questions and Answers in Attitudes Surveys Experiments in Question Forms Wording and Context", New York, Academic Press

Srivastava, S., O.P. John, S.D. Gosling, and J. Potter, 2003, "Development of personality in early and middle adulthood: Set like plaster or persistent change?", *Journal of Personality and Social Psychology*, Vol. 84, 1041-1053

Stevenson, B. and J. Wolfers, 2009, "The Paradox of Declining Female Happiness", NBER Working Paper 14969

Stone, A.A., J.E. Schwartz, J.E. Brodericka and A. Deaton, 2010, "A Snapshot of the Age Distribution of Psychological Well-being in the United States", PNAS Paper

Treiblmaier, H., P. Filzmoser, 2009, "Benefits from using continuous rating scales in online survey research" *Technische Universitt Wien, Forschungsbericht*

van Praag, B.M.S.and A. Ferrer-i-carbonell, 2004, *Happiness Quantified: A Satisfaction Calculus Approach*, Oxford University Press , New York

Weng, L.-J., 2004, "Impact of The Number of Response Categories and Anchor Labels on Coefficient Alpha and Test-Retest Reliability", *Educational and Psychological Measurement*, Vol. 64, 956972

Winkelmann, L. and R. Winkelmann, 1998, "Why Are the Unemployed So Unhappy? Evidence from Panel Data", *Economica*, Vol. 65, No. 257, 1-15

Wood, W., N. Rhodes and M. Whelan, 1989, "Sex Differences in Positive Well-Being: A Consideration of Emotional Style and Marital Status", *Psychological Bulletin*, Vol. 106, No. 2, 249-264

Tables

Table 1: Test for Randomization - March Sample

	LIKERT Scale		VAS		Mean Equality
	Observations	Mean	Observations	Mean	T-Test P-Value
Proportion male	2537	0.46	2505	0.47	0.54
Net monthly income (EUR)	2423	1526.98	2359	1499.51	0.81
Age	2537	49.71	2505	49.96	0.61
Number of hh-members	2537	2.62	2505	2.62	0.93
Number of hh-kids	2537	0.85	2505	0.84	0.85
Houseownership	2537	0.73	2505	0.72	0.66
Proportion out of laborforce	2537	0.36	2505	0.38	0.05
Proportion unemployed	2537	0.03	2505	0.03	0.49
Proportion working	2537	0.53	2505	0.50	0.08
Proportion with secondary education	2537	0.44	2505	0.44	0.80
Propordion with vocational education	2537	0.45	2505	0.44	0.60
Proportion married	2537	0.57	2505	0.58	0.33
Proportion separated	2537	0.09	2505	0.09	0.81
Proportion foreigner	2483	0.13	2439	0.11	0.01

Table 2: Test for Time and Order Effect

	Coefficient	Standard Error
April dummy	-0.02	0.04
VAS·April dummy	0.01	0.08
Experiment 2^{nd} dummy	-0.05	0.04
VAS·Experiment 2^{nd} dummy	-0.03	0.06
VAS dummy	-0.44***	0.06

· Paired sample: $N_p = 8548$
· Clustered s.e. at individual level; $***p < 0.01, **p < 0.05, *p < 0.1$
· Experiment 2^{nd} equals 1 if during the 2 hours preceding the lifesatisfaciton questionnaire the background variable questionnaire was answered.

Table 3: Construct Validity of Rating Scales

	VAS	LS 10	LS 11
Extraversion	0.19	0.21	0.22
Agreeablness	0.08	0.09	0.11
Consciousness	0.17	0.16	0.19
Emotional stability	0.43	0.40	0.43
Oppeness to experience	0.03	0.04	0.05
Self-esteem	0.41	0.40	0.43

Table 4: Moments of Happiness by Rating Scale - March and April

	Likert Scale Moments	VAS Moments	Equality of Mean T-Test (P-Value)	Equality of Variances Levine's Test (P-Value)	Discretized VAS Moments	Equality of Mean T-Test (P-Value)	Equality of Variances Levine's Test (P-Value)
March wave	7.16	6.70	0.00		6.92	0.00	
	(1.22)	(1.53)		0.00	(1.69)		0.00
April wave	7.14	6.70	0.00		6.91	0.00	
	(1.18)	(1.49)		0.00	(1.65)		0.00
Paired sample	7.16	6.71	0.00		6.91	0.00	
	(1.19)	(1.50)		0.00	(1.66)		0.00

· Standard deviations in parantheses

Table 5: Explaining Score Locations by Linear Probability Models

	Average Effect on Probability of		
Explanatory Variable	$7 > s_i > 8$	$s_i \in \{1, 2, 3\}$	$s_i = 9$
VAS	0.216***	0.026***	0.082***
	(0.009)	(0.003)	(0.006)

· Full Sample: $N_p = 7114$, clustered standard errors
· Other explanatory variables: order of question type, order of questionnaires, gender, lnet-inc, age, age2, number of person in hh, number of kids in hh, cohabitation with partner, houseownership, employment- and marital status, education level, origin.

Table 6: Happiness Regressions by Rating Scales

	Likert Happiness		VAS Happiness	
Explanatory Variables	Coefficient	S.E.	Coefficient	S.E.
Male	-0.162***	0.042	-0.012	0.054
Log of Monthly Net Income (EUR)	0.131***	0.035	0.157***	0.045
Age	-0.042***	0.008	-0.045***	0.010
Age^2 $\cdot 10^{-2}$	0.044***	0.008	0.045***	0.010
Number of hh-members	-0.028	0.101	-0.163	0.128
Number of hh-kids	-0.014	0.103	0.131	0.131
Cohabiting	0.286***	0.111	0.397***	0.140
Houseownership	0.267***	0.047	0.362***	0.059
In workforce	0.113**	0.055	0.163**	0.070
Unemployment	-0.138	0.121	0.021	0.153
Secondary Education	-0.004	0.072	-0.013	0.091
Vocational Education	-0.026	0.074	-0.091	0.094
Married	0.338***	0.064	0.468***	0.081
Separated	-0.053	0.074	-0.134	0.093
Foreigner	-0.167***	0.061	-0.180**	0.077
Experiment 2^{nd}	0.008	0.046	-0.073	0.058
April dummy	-0.028	0.038	0.010	0.048
Constant	6.651	0.298	6.074	0.379

· $N_{lik} = N_{vas} = 3557$,
· $***p < 0.01, **p < 0.05, *p < 0.1$

Table 7: Explaining Differences in Happiness Scores

Explanatory Variables	Coefficients	S.E.
Male	0.148***	(0.041)
Log of Monthly Net Income (EUR)	0.026	(0.030)
Age	-0.003	(0.008)
Age^2 $\cdot 10^{-4}$	-0.148	(0.834)
Number of hh-members	-0.134	(0.105)
Number of hh-kids	0.144	(0.106)
Cohabiting	0.111	(0.116)
Houseownership	0.095**	(0.047)
In workforce	-0.050	(0.056)
Unemployment	0.158	(0.140)
Secondary Education	-0.011	(0.072)
Vocational Education	-0.066	(0.073)
Married	0.132**	(0.064)
Separated	-0.079	(0.081)
Foreigner	-0.012	(0.065)
Experiment 2^{nd} March	0.005	(0.044)
Experiment 2^{nd} April	-0.062	(0.036)
VAS·March	-0.021	(0.036)
Constant	-0.560	(0.285)

· $y_i = s_{i,vas} - s_{i,lik} \in [-9, 9]$
· $N_p/2 = 3557$
· $***p < 0.01, **p < 0.05, *p < 0.1$
· Explanatory variables due not change for any individual from wave 1 to wave 2, except the dummy indicating if background variables were updated before the experiment (Experiment $2nd$)

Table 8: Explaining Women's Behaviour by Linear Probability Models

	Average Effect on Probability of		
Explanatory Variable	$s_{i,vas} - s_{i,lik} < -1$	$s_{i,vas} - s_{i,lik} \in [-1, 1]$	$s_{i,vas} - s_{i,lik} > 1$
Male	-0.063***	0.049***	0.014*
	(0.016)	(0.017)	(0.008)

· Full Sample: $N_p = 7114$, clustered standard errors
· Other explanatory variables: question type, order of question type, gender, lnetinc, age, age2, number of person in hh, number of kids in hh, cohabitation with partner, houseownership, employment- and marital status, education level, origin.

Graphs

Figure 1: Data Structure: Stocks and Flows

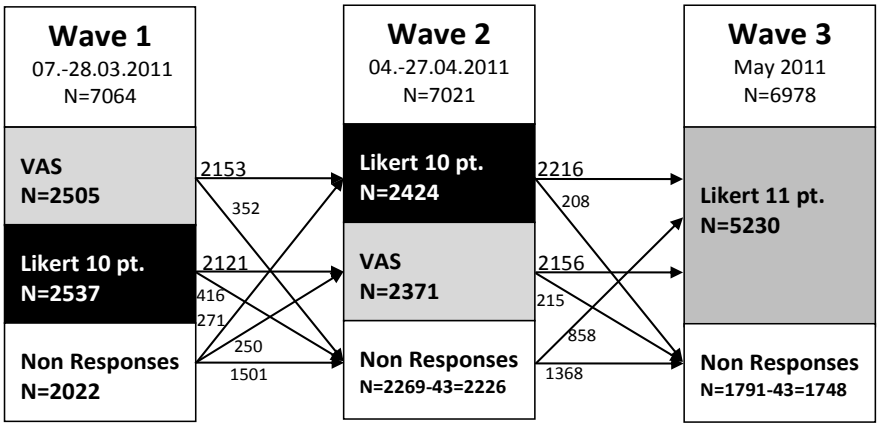

Figure 2: Question Design

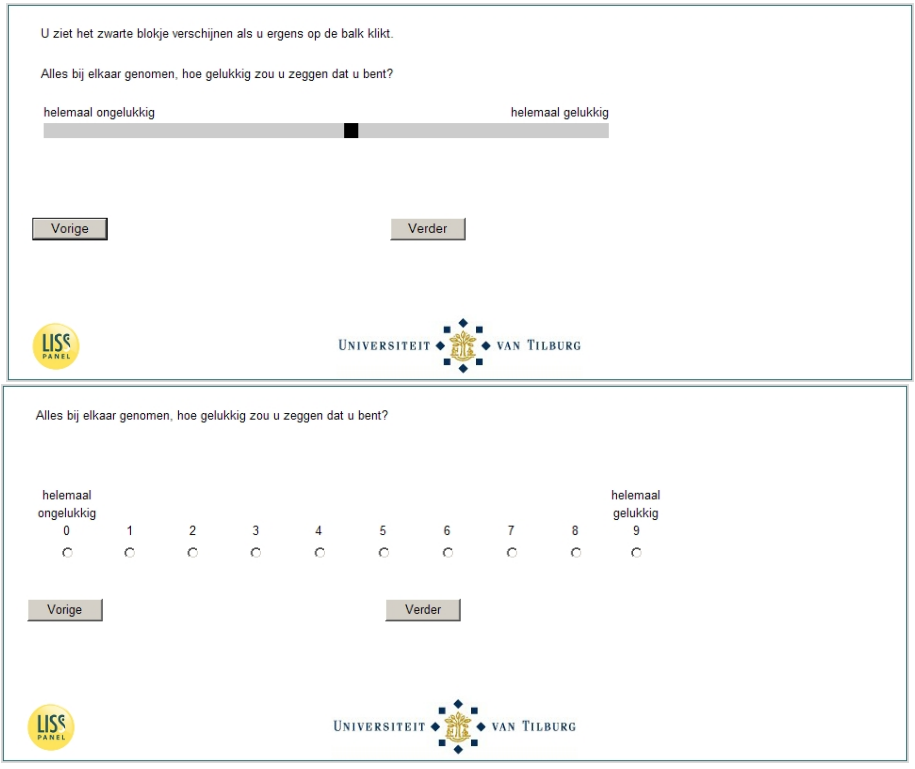

Figure 3: Evaluation of Question Design

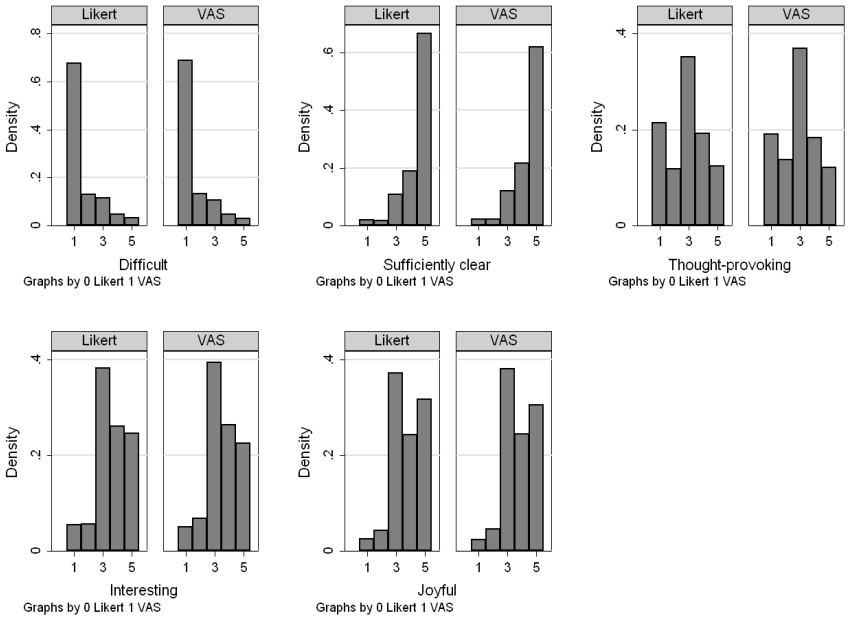

Figure 4: Score Densities - March and April

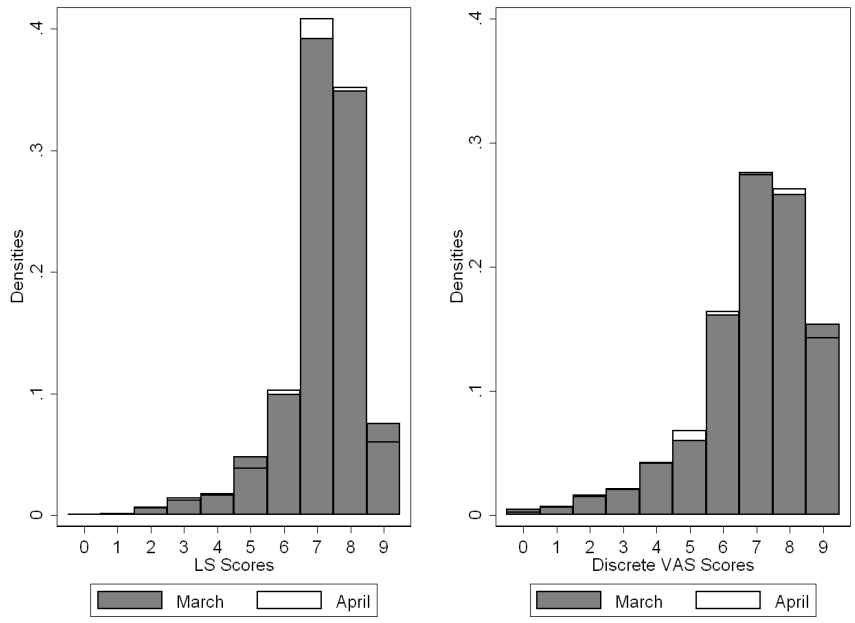

Figure 5: Score Densities - March