Survey Research Methods (2021)
Vol. 15, No. 1, pp. 101-113
doi:10.18148/srm/2021.v15i1.7725
European Survey Research Association

How to Reconstruct a Trend when Survey Questions Have Changed Over Time. Methods for Scale Homogenization Applied to the Case of Life Satisfaction in Japan 1958-2007

Tineke de Jonge
Erasmus Happiness Economics Research Organization
Erasmus University Rotterdam, The Netherlands

Akiko Kamesaka
School of Business Administration
Aoyama Gakuin University, Japan and
Former Visiting Research Fellow
Economic and Social Research Institute (ESRI)
Cabinet Office
Government of Japan

Ruut Veenhoven
Erasmus Happiness Economics Research Organization
Erasmus University Rotterdam, The Netherlands and
Opentia Research Program
North-West University South-Africa

Many trend studies draw on survey data and compare responses to questions on the same topic that has been asked over time. A problem with such studies is that the questions often do not remain identical, due to changes in phrasing and response formats. We present ways to deal with this problem using trend data on life satisfaction in Japan as an illustrative case. Life satisfaction has been measured in the Life in Nation survey in Japan since 1958 and the question used has been changed several times. We looked at three methods published by scholars who tried to reconstruct a main trend in life satisfaction from these broken time series, coming to different conclusions. In this paper we discuss their methods and present two new techniques for dealing with changes in survey questions on the same topic.

Keywords: Trend analysis; response scale homogenization; survey questions; happiness; life satisfaction

1 Introduction

1.1 The problem

Many topics in survey research have been measured over the years. The analysis of trends over such time series is often hampered due to changes in the operationalization of the latent variable[1]. Minor changes in the wording of survey questions or in the wording used to label response options, can influence how respondents interpret a question and cloud as such the view on the time trend. Bjørnskov (2010, p. 43) points out that a discussion about the effect that the way the life satisfaction question is framed has on the answers given, has been ongoing since the 1940s. More recently the discussion on scale effects and other factors such as a change in the mode of surveying or the ordering of survey questions, is still going on, Saris and Gallhofer (2007), Mazaheri and Theuns (2009), OECD (2013), Stone and Mackie (2013), Bais et al. (2019), Bond and Lang (2019).

The trend in life satisfaction in Japan is a prominent case in this discussion. Japan exemplifies a case of late but rapid modernization in the late 19th century, and in the second half of the 20th century the country witnessed unprecedented economic growth. Since the 1960s there has been a growing concern about the cost of economic growth and claims have been made that economic growth does not make us any more satisfied with life (e.g. Easterlin 1974). Responding to this train of thought, several researchers analyzed the time trend in life satisfaction in Japan.

1.2 Life satisfaction in the Japanese Life in Nation Survey

The longest time series on life satisfaction in Japan is the Life in Nation (LIN) survey, for which there is at least one measurement for almost every year since 1958.The question used to measure life satisfaction in the LIN survey and the

Contact information: Tineke de Jonge, P.O.Box 1738, 3000 DR Rotterdam, The Netherlands. (E-mail: dejonge@ese.eur.nl.)

[1] A latent variable is an unobservable variable, which can take any value on a continuum and is normally measured indirectly through observed scores obtained using a discrete scale.

wording used to label the 4 verbal response options have been changed several times during this period.

Suzuki (2009, pp. 84–85) remarks that the labels of the response options of the question on life satisfaction in the LIN survey have significantly changed between 1958 and 2007. In 1964 the labels in the questionnaire were changed from long descriptions which translated into English read "I cannot stand my current life", "My current life is far from satisfactory", "I am not satisfied with my life, but it is not too bad to keep more or less of the current level" and "My life could be better, but on the whole I am satisfied with my current life' to more tersely formulated statements which translated into English read "Extremely dissatisfied", "Fairly dissatisfied", "Rather satisfied, but not sufficiently" and "Sufficiently satisfied". A second change happened in 1992 when the response scale was made more symmetric by changing the labels of the response options to "Dissatisfied", "A little dissatisfied", "Rather satisfied" and "Satisfied", and a Don't-know choice was added. Stevenson and Wolfers (2008) also point to the changes in the survey question for measuring life satisfaction between 1958 and 2007 in terms of the labels of the response options. They (p. 21) make one additional split to Suzuki's time split series on life satisfaction to account for a change in the focus of the LIN survey leading question which was changed from "life at home" to "general life satisfaction" in 1970. Suzuki also mentions this change (p. 8, footnote 2), but he, as a native speaker, is of the opinion that the question employed until 1969 would be interpreted as focusing on your life which makes the effect of a change less relevant.

1.3 Plan of this paper

Using the LIN data, Easterlin (1995), Stevenson and Wolfers (2008), Suzuki (2009) tried to deal with the broken time series in trend analysis and came to different conclusions. Each of them applied a different methodological approach to overcome changes in the operationalization of the latent variable. This implies that we have a methodological problem that must be solved before we can talk authoritatively about the situation in Japan with respect to life satisfaction.

In this paper we will consider several approaches for dealing with discontinuities in time series caused by changes in the operationalization of a latent variable when analyzing trends. We will first describe and discuss the methods applied by respectively Suzuki (2009), Easterlin (1995), Stevenson and Wolfers (2008) to deal with the changes in the question on life satisfaction in the LIN survey. Following on this discussion, we will consider two new techniques for dealing with diversity in survey questions over time on the same topic[2] that can be used to transform the mean life satisfaction values of the different LIN surveys into values on a discrete or continuous numerical scale. We end this paper with a discussion and conclusions.

We restricted the time series on life satisfaction taken from the LIN survey in this paper to the period 1958–2007, as this period was considered in the papers mentioned above.

2 Three Methods Used to Deal with the Broken Time Series on Life Satisfaction from the Japanese LIN Survey

2.1 Method used by Suzuki: Rank Method and dummy variables

Mean life satisfaction in the study by Suzuki (2009) is based on the Rank Method in which the sample mean is calculated as the weighted average of the ranks of the response options using the relative frequencies as weights. The Rank Method is the common practice to calculate a sample mean from responses to questions with verbal response options ranked from high to low, such as "very satisfied', 'pretty satisfied" and "not at all satisfied". These response options are assigned rank numbers, in this example 3 for "very satisfied", 2 for "pretty satisfied" and 1 for "not at all satisfied", which are treated as numerical values. The trend in mean life satisfaction using the Rank Method to the time series from the LIN survey, is given in Figure 1. Note: the time series is split into four time periods in Figure 1, based on the years in which the survey question was changed, we give the linear trend for each of these time periods and for the entire period 1958–2007.

As can be seen from Figure 1, mean life satisfaction in Japan fluctuates between 2.67 and 2.88 when the ranks of the response options are used to calculate the mean, and the trend lines differ slightly between the time periods.

Suzuki (2009)) argues that if the wording used to label the response options of a survey question changes, this could have an effect on the response option a respondent may choose. He states, for example, that it is reasonable to expect that someone who would choose the option with rank 2 when asked a question before 1964, would choose the option with rank 1 after the change in the wording of the LIN question in 1964, and that this would lead to a significant decrease in the measured mean value. Suzuki underpins this argument by presenting a figure with a 3-point moving average in mean life satisfaction resulting from the Rank Method (Suzuki, 2009, p. 85). The main difference between the figure with the 3-point moving average and Figure 1 is that Suzuki limits the range of the vertical axis from 2.4 to 2.9, which puts an emphasize on a drop in mean life satisfaction between 1963 and 1964 of about 0.15.

In response to the change in the labelling of the response options, Suzuki introduced two dummy variables for the periods 1958–1963 and 1964–1991 to indicate the different pe-

[2]The descriptions of these techniques are based on those written by us in previous papers and books, among others (De Jonge, Veenhoven, & Arends, 2014; DeJonge, Veenhoven, & Kalmijn, 2017).

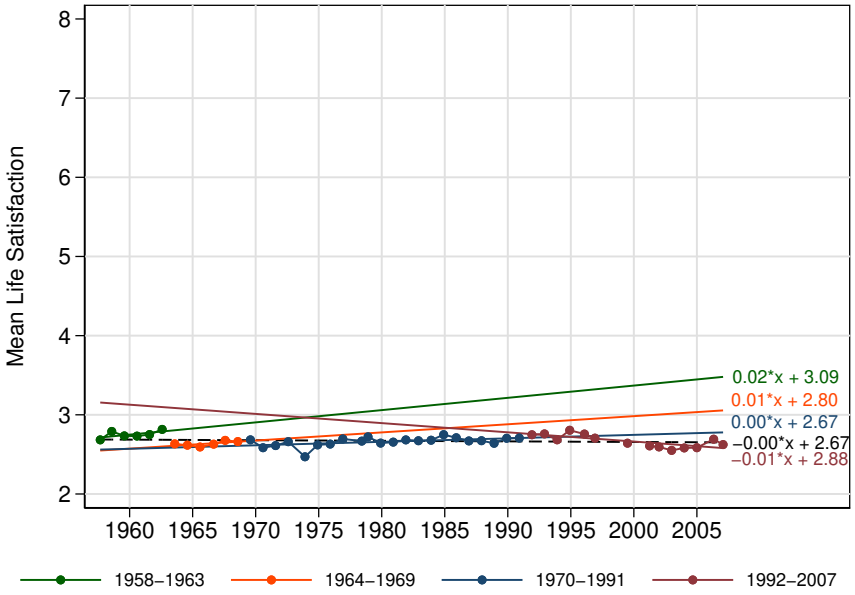

Figure 1. Mean life satisfaction in Japan, 1958–2007, Rank Method

riods of measuring life satisfaction. He then used these dummies as part of the set of independent variables in a regression model to examine the correlation between national income per capita and life satisfaction in Japan, with mean life satisfaction based on the Rank Method as independent variable. Suzuki thus did not try to reconstruct a main trend in life satisfaction from the broken time series taken from the LIN survey, but used dummies in an attempt to correct for the change in the labelling of the response options.

Weaknesses of this approach. *Disadvantages of the Rank Method.* The Rank Method has some serious disadvantages, that confuse the analysis of trends: 1) the range of the scale depends on the number of response options, 2) response options are considered to be equidistant which may not be realistic and 3) the verbal labels of the response options are ignored. This third disadvantage means that when applying the Rank Method in trend analysis, differences in the way the latent variable is operationalized are ignored. Additionally, in the Rank Method the assumption is made that the topic of interest has a discrete distribution within a population whereas it is more realistic to assume that any distribution will be continuous.

Using dummy variables. The dummies included in the models set up by Suzuki were intended to account for the changes in the LIN survey question over time, but in fact they account for all the changes that took place in the period in which one of the questions was used. These dummies thus also represent the economic developments in Japan in the period they cover and the developments taking place in society, such as rising divorce rates and falling fertility rates (The Japan Institute for Labour Policy and Training 2014). All the trends in things other than the economic indicators in a certain period may affect mean life satisfaction, both positively and negatively.

2.2 Method used by Easterlin: mean life satisfaction based on Thurstone Conversion

The five-fold multiplication of real per capita income in Japan in the period from 1958 to 1987 instigated Easterlin (1995) to question whether this development had been accompanied by a rise in the average level of life satisfaction in Japan. Easterlin looked at the time series on mean life satisfaction in Japan based on measurements using the LIN survey that can be found in the World Database of Happiness (Veenhoven, 2020) to answer this question. These values on mean life satisfaction in Japan are the result of a Thurstone Conversion.

Thurstone Conversion is used to cope with the disadvantages of assumed equidistance and the disregard paid to the verbal labels of the response options and is a method where a group of experts is employed to rate the verbal labels of response options on a common numerical scale. For example on a 0–10 scale experts rated "very satisfied" as 9.3, "pretty satisfied" as 6.7 and "not at all satisfied" as 1.3. This method is called Thurstone Conversion, as Jones and Thurstone (1955) were pioneers of this method. In a Thurstone Conversion, the numerical values resulting from the expert rating are assigned as fixed values to the corresponding original verbal response options of a given survey question to obtain a transformed mean. This transformed mean can be cal-

culated as the weighted average of the fixed values on scale 0–10 using the relative frequencies measured in a wave of surveying as weights.

Veenhoven, Ehrhardt, Ho, and de Vries (1993) and 12 coworkers followed the example of Jones and Thurstone to determine and rate the degree of life satisfaction denoted by the verbal labels of commonly used survey questions on a numerical 0 to 10 scale. The questions on life satisfaction from the Japanese LIN survey were also included in this rating exercise, for which the labels of the response options were translated, literally, from Japanese to English. For the LIN questions this resulted in the fixed values given in Table 1.

The effect of these values on the trends in mean life satisfaction is shown in Figure 2.

Despite his expectations, Easterlin had to conclude that the results from the LIN survey as shown in Figure 2 signal that there is no significant correlation between national income and subjective well-being in Japan.

Weaknesses of this approach. The Thurstone Conversion overcomes the disadvantages of assumed equidistance and ignoring of the labels that are associated with the Rank Method and brings the mean values to a numerical value between 0 and 10. A weak point is that the context of the response options is disregarded: "Very satisfied" may, for example, be interpreted differently if it is preceded by the response option "Extremely satisfied" in a 7-point scale, than if it is the top anchor-point option in a 4-point scale or if a different language is used. Furthermore, the Thurstone Conversion, like the Rank Method, implicitly assumes a discrete distribution of life satisfaction in a population.

A last disadvantage of Thurstone Conversion that we mention for the results presented here, is that the assessment was done by only one group of experts on labels of the response options which had been translated literally from Japanese to English. It is unknown whether or not a different group of experts or providing translations by more than one group of experts would lead to the same results or whether the translation from Japanese to English is an influencing factor.

2.3 Method used by Stevenson and Wolfers: mean life satisfaction estimated in an ordered probit analysis

In their study on the relation between life satisfaction and GDP, Stevenson and Wolfers (2008)) reconstructed a main trend in life satisfaction using the results of the LIN survey. They stated that due to the changes in the life satisfaction question, the time series based on each question should be presented separately. They then performed an ordered probit analysis of life satisfaction on time fixed effects and GDP per capita for each of these subseries.

In Ordered Probit response options of a verbal scale are considered to represent intervals on a continuum, where the cut points between the response options of a verbal scale are estimated as parameters used to model the relation between the observed, discrete variable y and the latent, continuous variable y^*. In the case of a 4-point verbal scale, such as the one used to measure life satisfaction in the Japanese LIN survey, three cut points k_1, k_2 and k_3 are estimated assuming that:

$$\begin{aligned} y = 1 \quad &\text{if} -\infty < y^* < k_1 \\ y = 2 \quad &\text{if} \quad k_1 < y^* < k_2 \\ y = 3 \quad &\text{if} \quad k_2 < y^* < k_3 \\ y = 4 \quad &\text{if} \quad k_3 < y^* < \infty \end{aligned}$$

Given this relationship, the ordered probit model is based on the assumption that y_i^* depends linearly on x_i according to the following (Daykin & Moffatt, 2002, p. 160; Cabello, Pareschi, Li, & Vainora, 2018, pp. 14–15):

$$y_i^* = x_i'\beta + \mu_i \quad ,$$

where $i = 1, \ldots, n$ and $\mu_i : N(0, 1)$ In this model the scalar $x'\beta$ consists of independent variables, including dummies if desired. The mean value of the latent variable y^* depends on the values of the estimated cut points and will have a value on the range $[-\infty, \infty]$.

The ordered probit analysis performed by Stevenson and Wolfers led to the results shown in Figure 3.

Stevenson and Wolfers conclude from the data shown in Figure 3 that throughout the period in which Japan moved from poor to affluent, the first three panels, subjective well-being rose with GDP per capita and that, since 1992, the Japanese economy has shown very little growth, and this has come with a sharp fall in average life satisfaction.

In a second step Stevenson and Wolfers combined all the data on life satisfaction and GDP in ordered probit model with life satisfaction per wave as the dependent variable and three dummy variables defining the periods bounded by the changes in surveying, the unemployment rate and the log of GDP per capita. The ordered probit estimation of the model resulted in negative coefficients for the dummies and the unemployment rate and a positive coefficient for GDP per capita. This led to their conclusion, which contrasts that of Easterlin: there has been a strong relationship between life satisfaction and GDP per capita growth in Japan since 1958.

Weaknesses of this approach. *Using dummies.* The same remark about the use of dummies we made in relation to the model of Suzuki (see Section 2.1) can be made for the model set up by Stevenson and Wolfers: these dummies are intended to account for the changes in the survey question, but in fact account for all the changes that took place in the period in which one of the questions was used.

[3] Stevenson and Wolfers used short labels for the response options in Figure 4, instead of the longer formulation which they describe on pp. 21–22 of their paper.

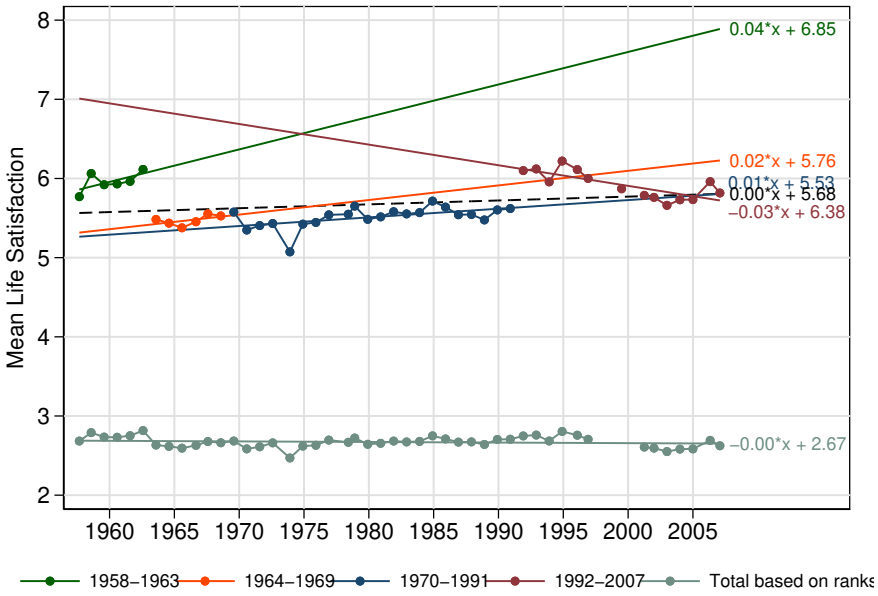

Figure 2. Converted time series of mean life satisfaction in Japan, Thurstone Conversion

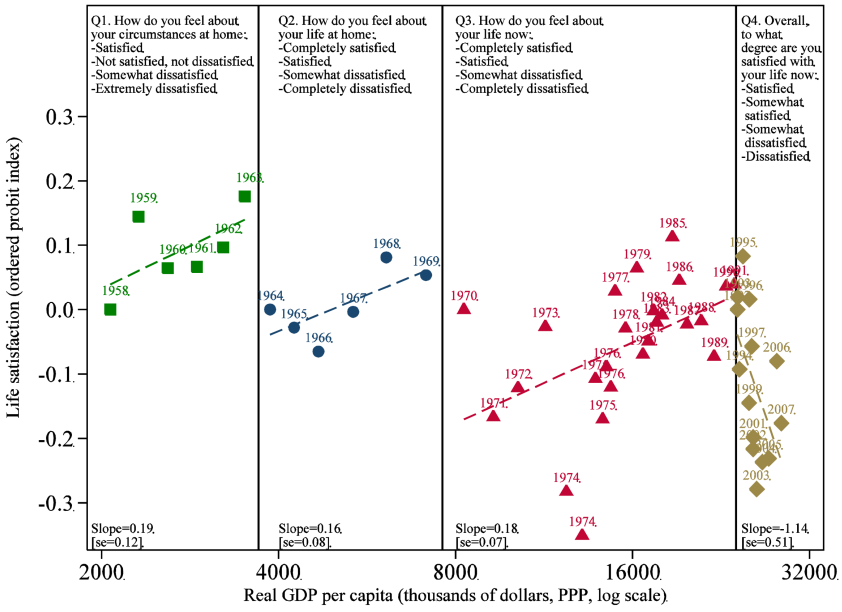

Figure 3. Life satisfaction and GDP per Capita over time in Japan (Note: original Figure 18, Stevenson and Wolfers (2008, p. 69). Source: Life in Nation Survey, 1958–2007)[3]

Table 1
Fixed values response options LIN questions

Rank response option	1958–1963	1964–1991	1992 to date
4	8.5	7.0	8.5
3	6.9	6.5	6.8
2	4.0	4.0	4.0
1	1.1	1.2	2.9

Distribution of the latent variable in ordered probit. In ordered probit it is assumed that the latent variable is normally distributed within a population (Stevenson & Wolfers, 2008, appendix-1; Daykin & Moffatt, 2002, p. 161). We doubt whether this is a realistic assumption in general, as in the western world most people have a positive perception of their own well-being which makes the distribution of responses to questions on life satisfaction skewed, with a long tail on the left that represents 'negative" outcomes (Diener & Diener, 1996; Cummins, 2003; Frijters, Johnston, & Shields, 2008).

Estimation of the "cut points" in ordered probit models. In the model of Stevenson and Wolfers latent, continuous variable y^* is latent life satisfaction and the scalar $x'\beta$ consists of log GDP per capita, the unemployment rate and the dummies for the different questions used in the Japanese LIN survey for measuring life satisfaction. As we understand, parameters have to be estimated in ordered probit on both sides of the equation to make the model fit: the coefficients of the variables on the right side of the equation and the cut points on the left side of the equation. This gives the impression that the cut points depend on the values of the independent variables. The paper by Stevenson and Wolfers, however, does not submit substantiating evidence on this. This requires further investigation, which goes beyond the scope of this paper.

Interpretation of the values of the cut points. According to Daykin and Moffatt (2002, p. 162), the values of the cut-points estimated in ordered probit models are rarely given interpretation. They mention two suggestions for the interpretation of cut points. Firstly, they expect that the dispersion of the cut points depends on the extent to which most people strongly agree or strongly disagree with a statement. Secondly, they believe that the cut points adjust according to the wording of a statement: they expect that the more difficult the wording is to understand, the more the middle cut points might be apart from each other. In our opinion, however, the cut points should be interpreted as the transition points between two consecutive response options, and therefore it is doubtful that they can be dependent on the independent variables in the model. We wonder whether the strong growth in life satisfaction reported by Stevenson and Wolfers is mainly to be attributed to this interdependency between the cut points and the independent variables in ordered probit, which forces life satisfaction in the model to follow the developments in log GDP per capita and the unemployment rate linearly. It would be a nice exercise to apply ordered probit to life satisfaction and other indicators than log GDP per capita and the unemployment rate, to see if this would give the same cut points. If so, this would be an indication that the cut points can be considered as the transition points between two consecutive response options. If the exercise results in different cut points, this would be an indication that they depend on the independent variables used in the model.

Value range of the cut points. In addition to the above, the cut points estimated in ordered probit are hard to interpret, as they can take all values on range $[-\infty, \infty]$. It would be easier to interpret these points if they are considered to be the transition points between response options on the continuum from 0 and 10.

3 Two New Methods for Scale Homogenization

In our opinion when a latent variable is measured using different survey questions, before drawing any conclusions the results should be homogenized before a trend can be identified. Coming forward to this request, we will present two recent scale homogenization methods: the Scale Interval Method and the Reference Distribution Method.

3.1 The Scale Interval Method

In the Scale Interval Method native language speaking judges assess the points on a bounded continuum from 0 to 10, in which verbal response options, in their language, for a given response scale transit from one option to another. This is done using the web-based Scale Interval Recorder (Veenhoven & Hermus, 2006).

The Scale Interval Method was developed to tackle the shortcomings of Thurstone Conversion (Veenhoven, 2008). In the Scale Interval Method, the response options in the measurement scale are not considered to be discrete points, but assumed to be bounded intervals that each represent a part of the continuum from 0 to 10 where the value of the latent variable can be found. The boundaries of these intervals are assessed in the context of the response scale using native speakers as judges of the points on a bounded continuum from 0 to 10 at which verbal response options for a given response scale transit from one option to another. This assessment can be done using the web-based Scale Interval Recorder (Veenhoven & Hermus, 2006). See Figure 4 for

an illustration of this method using an example of a question from the trial version[4].

Using this recorder, a series of survey questions is presented on a computer screen to judges. Questions are presented sequentially on the left side of the screen and each question presented consists of a leading question and its corresponding verbal response scale with options given in the judges" mother tongue. The judges have to shift sliders on the vertical bar scale on the right side of the survey question until they feel that the intervals represented on the vertical bar correspond to the meanings of the words as used for the verbal response options. A possible result of this shifting of the sliders is given on the right of Figure 4.

The assessments of all native speaking judges are averaged for each transition point between two consecutive response options, to define the boundaries of these intervals. The lower boundary of the response option labelled by the worst option is, by definition, always equal to 0. The upper boundary of the response option labelled by the best option by definition is always equal to 10.

An important assumption of the Scale Interval Method is that the distribution of the latent variable is continuous within a population and can well be approximated by a beta distribution that is bounded by the domain from 0 to 10, has shape parameters α and β and a mean on the 0–10 continuum that is equal to $10 * \alpha/((\alpha+\beta))$ (Kalmijn, 2010; Kalmijn, Arends, & Veenhoven, 2011). For a more elaborate description of this Continuum Approach we refer to Kalmijn (2010), and, DeJonge et al. (2017, Ch. 7). An estimated population mean in the Scale Interval Method for a given survey question and wave of surveying can be calculated on the basis of the shape parameters of the beta distribution that best fits the measured cumulative frequency distribution and the upper boundaries of intervals corresponding to the response options obtained from the assessment.

The approach to scale transformation used in the Scale Interval Method differs essentially from that used in Thurstone Conversion, as the response options in the primary scale are not considered to be discrete points, but to be intervals that each represent a part of the continuum from 0 to 10 where the value of the latent variable can be found. In the Scale Interval Method, the discrete points of the primary scale are converted to a series of connected intervals that together span the entire continuum from 0 to 10. Moreover, in the Scale Interval Method, the context of the scale is taken into account: a response option with a given label may represent a different intervals, depending on the position on the scale it is used in and the labels of the other response options in that scale.

An application of the Scale Interval Method will be used to illustrate the trend in life satisfaction in Japan between 1958 and 2007 in the next section.

Application to the time series on life satisfaction in Japan from the LIN survey. The boundaries between the response options of questions used for measuring life satisfaction in Japan, were assessed in the Scale Interval Study covered by the Japanese-6 and Japanese-7[5] studies. The assessments were done by almost 200 students drawn mainly from the School of Business administration of the Aoyama Gakuin University in Tokyo. Three of the four questions from the Japanese LIN survey used to measure life satisfaction in Japan since 1958 were included in the study, for the question used to measure life satisfaction from 1964–1969 we applied the boundaries obtained for the period 1970–1991, analogous to how the entire period from 1964–1991 was treated for the Thurstone Conversion. The results of the assessment are given in Table 2.

As can be seen from Table 2, there are small differences between the boundaries assessed for each period of time. Given the intervals representing the response options of the LIN question in each period, we estimated the parameters $\alpha(t, p)$ and $\beta(t, p)$ of the best fit beta distribution for each year t in period p. We used these parameters to estimate the mean life satisfaction in the population in each year in the period 1958–2007 as $\frac{10*\alpha(t,p)}{(\alpha(t,p)+\beta(t,p))}$. The result of these estimates is given in Figure 5.

The converted means based on the application of the Scale Interval Method fluctuate between a somewhat lower range than when applying Linear Stretch or Thurstone Conversion to the LIN question and had a minimum value 4.61 and maximum value of 5.62. On the basis of the Scale Interval Method we once again had to conclude that life satisfaction in Japan had not, on average, substantially changed in the period 1958–2007.

Weakness of this approach. The converted means of similar questions from different surveys fielded among representative samples of a population must be equivalent. The latter is a logical consequence of the fact that if the data from a representative sample are used to draw conclusions, these conclusions are representative for the population from which the sample is taken (D'Exelle, 2014). We noticed in earlier research, however, that against our expectations, this is not always the case if the Scale Interval Method is applied (De Jonge et al., 2014, p. 287; DeJonge et al., 2017, pp. 78–80). We developed the Reference Distribution Method, to deal with this weakness of the Scale Interval Method.

3.2 The Reference Distribution Method

The Reference Distribution Method builds heavily on the Scale Interval Method, but takes into account that for a given year and a given population, the means after transformation for similar questions from different surveys fielded among representative samples of a population must be equivalent.

[4] http://www.risbo.org/fsw/english-trial/

[5] See the study list at https://worlddatabaseofhappiness.eur.nl/scalestudy/scale_fp.htm.

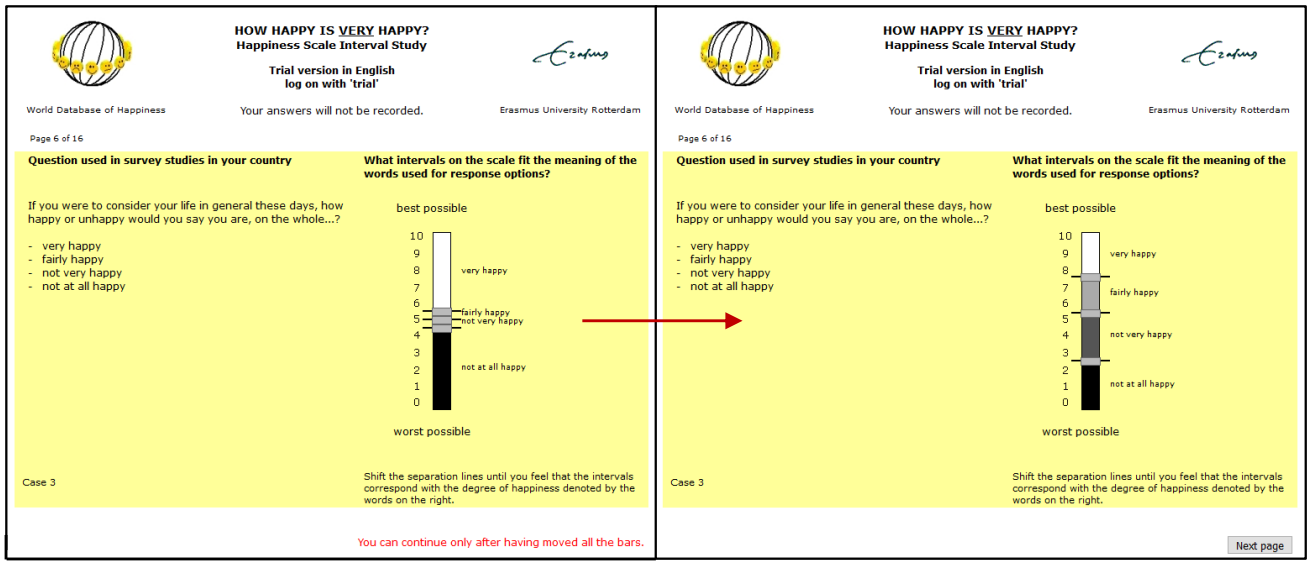

Figure 4. Example Scale Interval Recorder, before and after assessment

Table 2
Assessed boundaries between response options LIN questions

Rank response option	1958–1963	1964–1991	1992 to date
4	10.00	10.00	10.00
3	7.80	7.81	7.59
2	4.53	4.63	4.43
1	1.81	1.81	1.86

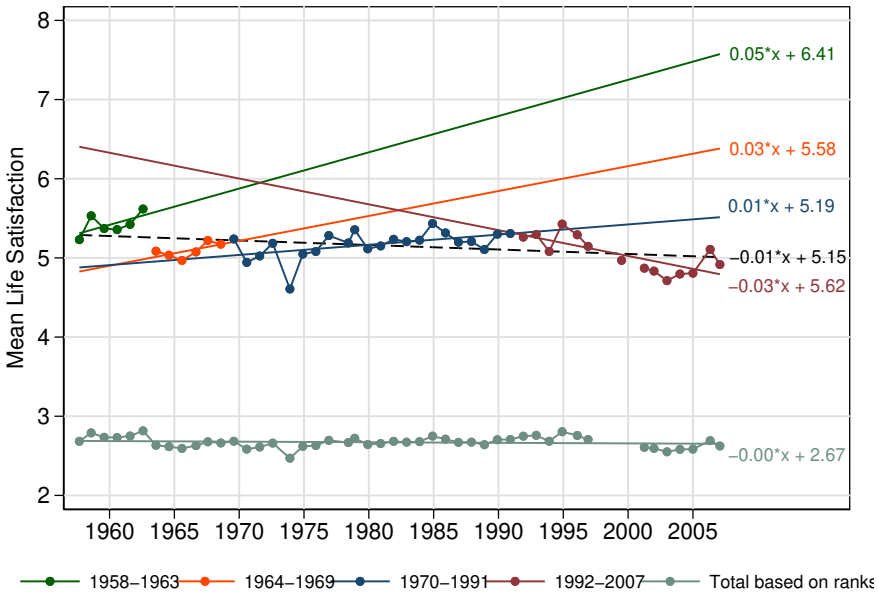

Figure 5. Converted time series of mean life satisfaction in Japan, Scale Interval Method

In the Reference Distribution Method, a reference distribution is used to make survey results from different surveys or those measured with the same survey but using different survey questions in different periods of time comparable. To explain what a reference distribution is, we make use of the question on life satisfaction taken from the World Values Survey (WVS), which consists of the leading question "All things considered, how satisfied are you with your life as-a-whole these days?" and 10 numerical response options. The cumulative frequency distribution for this question measured in Japan in 2005, as vertical bars equidistantly distributed[6] over the 0 to 10 continuum is shown in Figure 6. The curve shown in Figure 6 is the beta distribution that according to the Continuum Approach best fits the upper boundaries of the response options and cumulative frequency distribution of the WVS question in 2005 for Japan.

The shape parameters of this best fit beta distribution are equal to $\alpha = 3.98$ and $\beta = 2.20$ and give an estimated population of 6.45.

The best fit beta distribution to the WVS question can now be used as a reference distribution for deriving the boundaries between response options for other survey questions used for measuring life satisfaction in Japan in 2005. This is illustrated in Figure 7 for the question from the LIN survey used since 1992 and the measured frequency distribution in 2005. The cumulative frequency distribution for the LIN question in 2005 is given as a stacked bar on the left of Figure 7.

The boundaries between the response options are equal to the values on the 0–10 continuum where the cumulative frequency of the LIN question is equal to that of the reference distribution, which is 4.04 for the option "Dissatisfied", 6.03 for the option "Somewhat satisfied", 8.84 for the option "Fairly satisfied" and by default 10 for the option "Satisfied".

The converted population mean for the LIN question in 2005 is the same as that found for the reference distribution and equal to 6.45. The reference boundaries of the response options of the LIN question found in this way are kept fixed and are used to estimate the beta distributions that best fit the cumulative frequencies of other waves in which the same question has been employed. The estimated mean on the 0 to 10 continuum for each of these waves is equal to the mean of the corresponding best fit beta distribution.

It is plausible that the changes in the question on life satisfaction in the LIN survey between 1958 and 2007 have influenced the position of the boundaries between the response options. This position has therefore to be reconsidered for each change and (presumably) determined anew. These new boundaries must be equal to the values on the 0–10 continuum where the cumulative frequency of changed survey question is equal to the beta distribution that best fits the existing boundaries and the frequency distribution of the survey results in the year prior or equal to that in which the question was changed.

Application to the time series on life satisfaction in Japan from the LIN survey. We will now inspect what the transformed trends in life satisfaction in Japan for the period 1958–2007 are if the Reference Distribution is applied. We only used one wave of data per year, this, however, did not have a noteworthy effect on the results.

Previously we explained how we applied the Reference Distribution Method, using the 2005 Japanese wave of the WVS to provide a reference distribution which we used to derive the boundaries between the response options of the Japanese LIN question used between 1992 and 2007. We have also described how we calculated the converted means for each wave in this period. We then used the best fit beta distribution found for the 1992 Japanese LIN wave, as a reference distribution to derive the boundaries between the response options of the question used between 1970 and 1991. These boundaries are equal to the values where the cumulative frequency distribution of the 1991 Japanese LIN wave measured with the LIN question is equal to this new reference distribution. Note: this is a suboptimal solution, since no double measurement was done in 1992, i.e. one measurement with the question on life satisfaction used between 1970 and 1991 and one with the question on life satisfaction used from 1992 onwards. The estimated population mean for 1991 is thus equal to the estimated population mean in 1992.

Using the fixed boundaries derived for the question used in the Japanese LIN survey in the period 1970 to 1991, we estimated best fit beta distributions for all the other waves in this period. The best fit beta distribution found for the 1970 wave was then used as a reference distribution to derive the boundaries between the response options for the question on life satisfaction employed in the LIN survey from 1964 to 1969. Similarly, the best fit beta distribution found for the 1964 LIN wave was used as a reference distribution to derive the boundaries between the response options for the question on life satisfaction employed from 1958 to 1963.

The boundaries found for all periods using the Reference Distribution Method are given in Table 3. The boundaries differ slightly per period, which might be due to the difference in the wordings used for the question and the labels of the response options in each period.

The converted time series for mean life satisfaction in Japan, measured using the LIN survey, and following from this procedure is given in Figure 8. Note: we have included a measure of the trend in life satisfaction in Japan between 1958–2007 based on the Rank Method for comparison.

The converted means shown in Figure 8 fluctuate between 6.20 and 6.99 when applying the Reference Distribution Method using the 2005 wave of the WVS as a seed. The trend in the converted time series for the entire period from

[6] For numerical scales we fix the upper boundaries of the response options on the 0 to 10 continuum equidistantly (Kalmijn, 2013).

Figure 6. Application of the Continuum Approach to derive a reference distribution

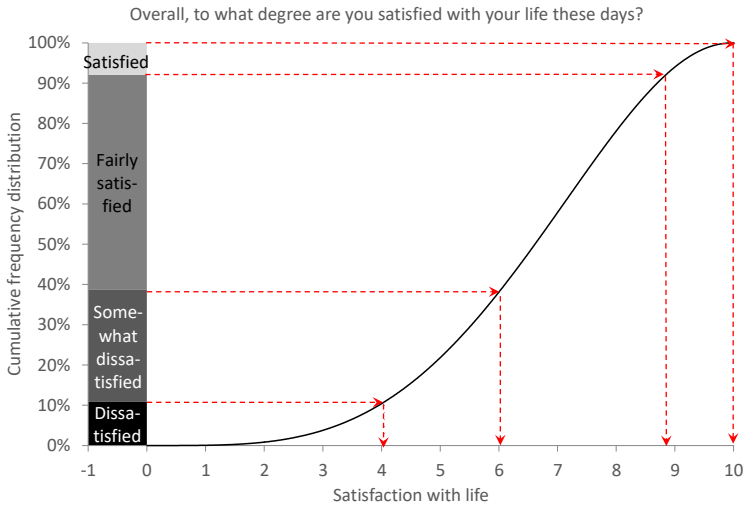

Figure 7. Illustration of the Reference Distribution Method

Table 3
Boundaries between response options LIN questions based on the Reference Distribution Method, starting with the 2005 wave of the WVS

Rank response option	1958–1963	1964–1969	1970–1991	1992 to date
4	10.00	10.00	10.00	10.00
3	8.30	9.07	9.05	8.84
2	6.23	6.28	6.18	6.03
1	3.94	3.67	3.64	4.04

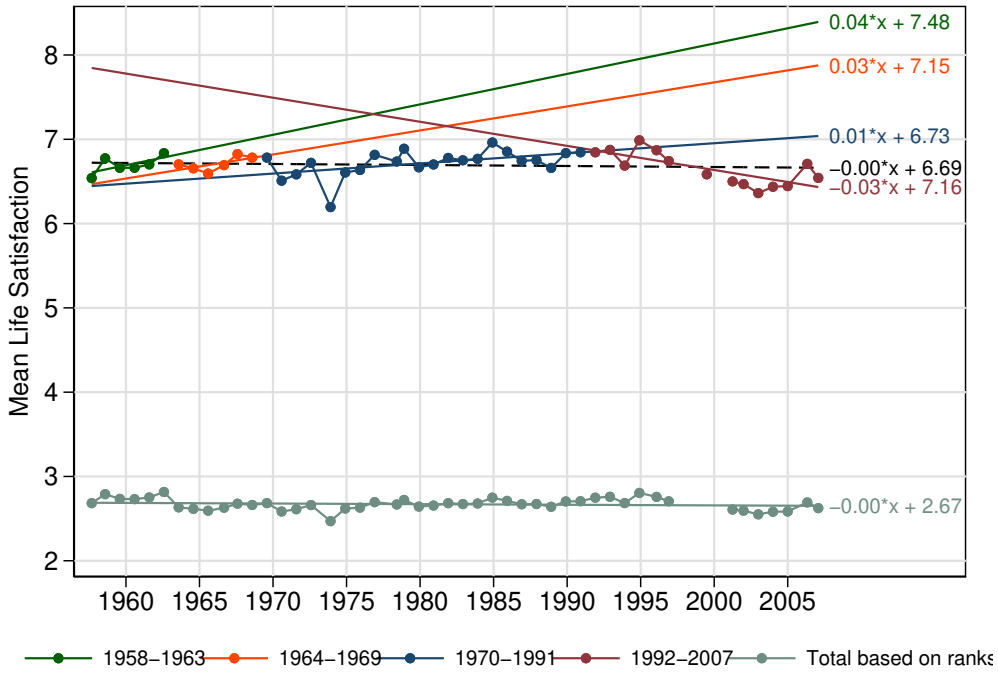

Figure 8. Converted time series of mean life satisfaction in Japan, Reference Distribution Method, initial reference distribution based on the 2005 wave of the WVS

1958–2007 has a zero slope. This led us to the conclusion that life satisfaction in Japan has not, on average, changed between 1958 and 2007.

4 Discussion

The use of verbal response scales. We have shown how different approaches to deal with changes in the operationalization of a latent variable may lead to very different conclusions about the evolution of a trend, using the case of life satisfaction in Japan as an illustration. The different conclusions are in in this case largely due to differences in how the verbal labels of the response options used over time are numerically interpreted. The use of verbal labels has recently given rise to a discussion on whether valid conclusion can be drawn based on the use of verbal response scales (Bond & Lang, 2019; Kaiser & Vendrik, 2019).

Less pronounced, but in line with this discussion, is the promotion by several researchers of using numerical response scales instead of verbal response scales. Veenhoven (2017, p. 80) uses the argument that respondents may differ more in their interpretations of words than in their interpretation of numbers. Sangster, Willits, Saltiel, Lorenz, and Rockwood (2001) claim that the use of a numerical scale with equidistant markings suggests the idea of equal intervals for the resulting scale more clearly than would be possible with any verbal scale and that using numerical response scales simplify and reduce errors in data preparation. Similar arguments are used by Scherpenzeel (1999) who advocates the use of 11–point numerical scales. Also, the OECD (2013) recommends to label options in a response scales of evaluative measures with numerical, rather than with verbal, labels.

Other scale effects and factors of influence than changes in the wording. Discontinuities in trends caused by changes in the wording of the labels of the response option, are to a great deal typical for survey questions using a verbal response scale. In numerical scales, most often only the anchor points of the scale are given a verbal label. We already made notice of other scale effects and influential factors such as a change in the mode of surveying or the ordering of survey questions in Section 1.1. These scale effects and influential factors may also apply to numerical scales. The Reference Distribution Method, however, is also suited to deal with discontinuities caused by these non-verbal scale related factors.

5 Conclusions

Changes in wording of survey questions on the same topic can obscure the view on the trend over time. Despite the doubts about the use of verbal scales and the recommendations to change to using numerical scales, we have to deal with measurements that have been done in the past. To incorporate the results of these past measurements in trend analyses, we need to convert these results to a common, numerical

range. The currently best available method we know to do that is the Reference Distribution Method. When no reference data is available, the Scale Interval Method is the second best option. Both methods lead to the conclusion that life satisfaction in Japan did not change in the period from 1958 to 2007.

6 Acknowledgements

This work was supported by the Japanese Society for the Promotion of Science (JSPS) KAKENHI "Grant Number 16H03640". We would like to thank Susumu Kuwahara and Ryoichi Watanabe of the Cabinet Office, Government of Japan, for their comments and support. The authors also thank Tije Euvermans, methodologist at the Erasmus University Rotterdam, for his comments and Miranda Aldham-Breary, senior volunteer at the WDH, for editing the English text.

References

Bais, F., Schouten, B., Lugtig, P., Toepoel, V., Arends-Tòth, J., Douhou, S., ... Vis, C. (2019). Can survey item characteristics relevant to measurement error be coded reliably? a case study on 11 dutch general population surveys. *Sociological Methods & Research*, *48*(2), 263–295.

Bjørnskov, C. (2010). How comparable are the gallup world poll life satisfaction data? *Journal of happiness Studies*, *11*(1), 41–60.

Bond, T. N., & Lang, K. (2019). The sad truth about happiness scales. *Journal of Political Economy*, *127*(4), 1629–1640.

Cabello, M., Pareschi, F., Li, Y., & Vainora, J. (2018). Case 2: About the transmission channels of unemployment rate on individual happiness. econometric game 2018.

Cummins, R. A. (2003). Normative life satisfaction: Measurement issues and a homeostatic model. *Social indicators research*, *64*(2), 225–256.

D'Exelle, B. (2014). Representative sample. doi:10.1007/978-94-007-0753-5_2476

Daykin, A. R., & Moffatt, P. G. (2002). Analyzing ordered responses: A review of the ordered probit model. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *1*(3), 157–166.

De Jonge, T., Veenhoven, R., & Arends, L. (2014). Homogenizing responses to different survey questions on the same topic: Proposal of a scale homogenization method using a reference distribution. *Social Indicators Research*, *117*(1), 275–300.

DeJonge, T., Veenhoven, R., & Kalmijn, W. (2017). Diversity in survey questions on the same topic. *Social Indicators Research Series*, *68*.

Diener, E., & Diener, C. (1996). Most people are happy. *Psychological science*, *7*(3), 181–185.

Easterlin, R. A. (1995). Will raising the incomes of all increase the happiness of all? *Journal of Economic Behavior & Organization*, *27*(1), 35–47.

Frijters, P., Johnston, D. W., & Shields, M. A. (2008). Happiness dynamics with quarterly life event data.

Jones, L. V., & Thurstone, L. L. (1955). The psychophysics of semantics: An experimental investigation. *Journal of Applied Psychology*, *39*(1), 31.

Kaiser, C., & Vendrik, M. (2019). How threatening are transformations of reported happiness to subjective wellbeing research? doi:10.31235/osf.io/gzt7a

Kalmijn, W. (2010). *Quantification of happiness inequality*.

Kalmijn, W. (2013). From discrete 1 to 10 towards continuous 0 to 10: The continuum approach to estimating the distribution of happiness in a nation. *Social Indicators Research*, *110*(2), 549–557.

Kalmijn, W., Arends, L., & Veenhoven, R. (2011). Happiness scale interval study. methodological considerations. *Social Indicators Research*, *102*(3), 497–515.

Mazaheri, M., & Theuns, P. (2009). Effects of varying response formats on self-ratings of life-satisfaction. *Social Indicators Research*, *90*(3), 381.

OECD. (2013). *Oecd guidelines on measuring subjective well-being*. OECD publishing.

Sangster, R. L., Willits, F. K., Saltiel, J., Lorenz, F. O., & Rockwood, T. H. (2001). The effect of numerical labels on response scales. Retrieved from https://www.bls.gov/osmr/research-papers/2001/pdf/st010120.pdf

Saris, W. E., & Gallhofer, I. (2007). Estimation of the effects of measurement characteristics on the quality of survey questions. In *Survey research methods* (Vol. 1, pp. 29–43).

Scherpenzeel, A. (1999). Why use 11–point scales? documentation of the swiss household panel. Retrieved from https://forscenter.ch/wp-content/uploads/2018/10/varia_11pointscales.pdf

Stevenson, B., & Wolfers, J. (2008). *Economic growth and subjective well-being: Reassessing the easterlin paradox*. National Bureau of Economic Research.

Stone, A. A., & Mackie, C. E. (2013). *Subjective well-being: Measuring happiness, suffering, and other dimensions of experience.* National Academies Press.

Suzuki, K. (2009). Are they frigid to the economic development? reconsideration of the economic effect on subjective well-being in japan. *Social Indicators Research*, *92*(1), 81–89.

Veenhoven, R. (2008). The international scale interval study.

Veenhoven, R. (2017). Measures of happiness: Which to choose? In *Metrics of subjective well-being: Limits and improvements* (pp. 65–84). Springer.

Veenhoven, R. (2020). World database of happiness: A findings archive. In *Handbook on wellbeing, happiness and the environment*. Edward Elgar Publishing.

Veenhoven, R., Ehrhardt, J., Ho, M. S. D., & de Vries, A. (1993). *Happiness in nations: Subjective appreciation of life in 56 nations 1946–1992*. Erasmus University Rotterdam.

Veenhoven, R., & Hermus, P. (2006). Scale interval recorder: Tool for assessing relative weights of verbal response options on survey questions. Web survey program. Erasmus University Rotterdam, Department of Social Sciences & Risbo Contract Research, The Netherlands. Retrieved from https://worlddatabaseofhappiness.eur.nl/scalestudy/scale%5C_fp.htm.Veenhoven