# DIFFERENT SURVEY QUESTIONS ON THE SAME TOPIC

## How to make responses comparable?

Tineke de Jonge

# DIFFERENT SURVEY QUESTIONS ON THE SAME TOPIC

## How to make responses comparable?

**Tineke de Jonge**

Cover design: © 2015 Robert Oude Wolbers

Printed in The Netherlands by Ipskamp Drukkers B.V., Enschede

ISBN 978-94-6259-670-2

# DIFFERENT SURVEY QUESTIONS ON THE SAME TOPIC

## How to make responses comparable?

**VERSCHILLENDE ENQUÊTEVRAGEN OVER HETZELFDE ONDERWERP**
**Hoe kunnen antwoorden vergelijkbaar gemaakt worden?**

**Proefschrift**

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

**Prof.dr. H.A.P. Pols**

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
donderdag 2 juli 2015 om 11.30 uur

**Jantiena Jacoba de Jonge**
**geboren te Winschoten**

## Promotiecommissie

| | |
|---|---|
| Promotor: | Prof.dr. L.R. Arends |
| Overige leden: | Prof.dr. J. J. van Busschbach |
| | Prof.dr. J.J.G. Schmeets |
| | Prof.dr. P. Theuns |
| Copromotoren: | Prof.dr. R. Veenhoven |
| | Dr.ir. W.M. Kalmijn |

## Keywords

Beta distribution
Comparability
Cultural measurement bias
Demographic categories
Happiness
Interpreters' bias
Numerical rating scales
Parameter estimation
Pooling time series
Reference distribution
Research synthesis
Response scales
Satisfaction with life
Scale equivalence
Scale interpretation
Scale interval method
Scale transformation
Subjective well-being
Survey research
Trend analysis
Verbal rating scales
World Database of Happiness

# DIFFERENT SURVEY QUESTIONS ON THE SAME TOPIC
## How to make responses comparable?

## Table of Contents

# PART 1

## Comparing responses to different survey questions on the same topic
## PROBLEMS AND CONVENTIONAL SOLUTIONS

# 1 Diversity in survey items and the comparability problem

## 1.1 Introduction

Survey research is a major method used in the social sciences and is largely based on standard questions with pre-coded response options called 'response scales' to which respondents answer by picking one of the options. There is little uniformity in survey items[1] used. This difference in items is no problem when surveys are analysed separately, but it limits the comparability of findings gathered in different surveys that used different items for the same topic. This reduces our accumulation of knowledge and calls for techniques to improve the comparability of data.

This diversity in the wordings of questions and in response options also appears in survey research on subjective well-being which took off in the 1970s in the wake of the Social Indicator Revolution. In this context, Andrews and Withey (1976) explored a large set of survey items, among which were questions on the subjective appreciation of one's life as a whole. Many more questions designed to measure subjective well-being have been used since then. To date about a 1,000 different questions on the subjective appreciation of one's life as a whole from some 10,000 studies have been gathered in the collection 'Measures of Happiness' of the World Database of Happiness (Veenhoven, 2015a, 2015b). About half of the differences in questions are in the number and wording of response options, other differences are due to causes such as the reference to time, the visual presentation of the scale or the method of assessment. The World Database of Happiness (WDH) focuses on happiness in the sense of subjective enjoyment of one's life as a whole (Veenhoven 1984). In this definition 'happiness' is synonymous with 'life satisfaction'. This concept of happiness is currently the one most commonly used in the social sciences and it lies at the heart of the WDH (Veenhoven 2011).

One of the aims of happiness researchers is to assess differences in happiness across nations. This requires comparison of data drawn from different surveys containing questions about happiness. In surveys however, different kinds of response scales are used, both verbal scales and numerical scales and these scales also differ in the number of response options available, some including only two options, for example yes or no, and others as many as eleven, for example 0 to 10 numerical scales. As a

[1] We use the term 'item' for a survey question and its corresponding response options.

consequence only a part of the available research can be used. Likewise, another aim of happiness researchers is to compare happiness within countries over time. This also requires equivalent questions and response scales, but since the response scales can change over the years, the number of comparable data will often be inadequate for a valid comparison to be made.

The diversity in survey items is often treated in one of two ways: one way is to abstain from any comparison when rating scales are not fully identical. This means that most of the findings on happiness are incomparable and thus lost for synthetic analysis. The other approach is to ignore the problem, typically by assuming that the ranks of the response options reflect the degree of happiness denoted and treating these numbers as metric values that can be transformed to the same range. This latter approach carries the danger of producing nonsense.

### 1.2 Overview of the diversity in survey items

Most people have a positive perception of their own well-being, at least in the western world. As a result, the distribution of responses to questions on happiness is skewed, with a long tail on the left that represents 'negative' outcomes (Lee, Kim, and Shin, 1982; Diener, and Diener, 1996; Cummins, 2003; Frijters, Johnston, and Shields, 2008; Guven, Senik, and Stichnoth, 2011). Irrespective of the scale used, this skewness has to be kept in mind when interpreting the results of such measurements.

Within the large set of existing measures of happiness, the number of response options and the distinction between verbal and numerical response scales are obvious variations. To meet the skewness of the distribution, in the past verbal scales have been devised that are skewed due to mainly positively formulated response options. An example of such a scale is the one used by Statistics Netherlands to measure satisfaction with life, consisting of the response options 'Extraordinarily satisfied', 'Very satisfied', 'Satisfied', 'Fairly satisfied' and 'Not very satisfied'. Only the latter of these responses is formulated negatively. The idea behind this rather asymmetric scale at the time it was devised was that it would give the possibility for more variation in the responses than if a more symmetric scale was used. The satisfaction with life scale of Statistics Netherlands is a unipolar scale: all response options contain the word 'satisfied'. This differs from a bipolar scale, where in the response options, for example, the word 'dissatisfied' would be used as the opponent of 'satisfied'. Furthermore, a scale does not necessarily need to have a neutral midpoint dividing it into a positive and a negative pole and the end points of different scales may vary

in the extremity of the wording used, for example 'extraordinarily' is more extreme than 'very' but both are subject to the respondents interpretation of the words and this will vary from respondent to respondent, and each variation will influence the response patterns (Cummins and Gullone, 2000).

Most of the variations discussed above hold for both verbal response scales and for numerical response scales. Although numbers are used on a numerical scale to express the respondent's degree of happiness, it is still necessary to use words to describe what the anchor points[2] of the scale mean, and it is this wording which defines whether the scale is conceived unipolar or bipolar. The wording of such descriptions can include the subject of measurement, as in 'dissatisfied' –'satisfied' or leave to the respondents how they interpret the anchor points or extremes of the scale when a formulation is given in terms like 'best possible' – 'worst possible'.

Other variations in numerical scales are the visual orientation, which can be vertical or horizontal, and the labelling of the anchor points that can go from negative to positive, for example -5 to +5, consists only of positive numbers and possibly including zero starting at 0 or 1, or there can be no numbering (Mazaheri and Theuns, 2009). In an experiment done by Schwartz et al (1991) using an 11-point numerical scale with anchor points labelled from 'Not at all successful' to 'Extremely successful' and ranging from -5 to +5 only thirteen per cent of the respondents gave an answer between -5 and 0. When the range changed from 0 to 10, the percentage of answers at the lower end of the scale changed to 34 per cent. A similar result was found by Sangster and colleagues (2001). From this experiment Schwarz and his colleagues concluded that a numerical scale starting at zero suggests the absence or presence of the subject under study, which makes the scale unipolar. If conversely, one half of the scale is negative and the other half is positive, then the positive values are related to the presence of the subject one is interested in, whereas the negative values represent the opposite. Schwartz et al also suggest that scales that are intended to assess the intensity of a single attribute, for example happiness, should follow a zero-to-positive-values format to emphasize that the question pertains to the absence or presence of this specific attribute, rather than the presence of its opposite. This suggestion is underpinned in an elaborated discussion on happiness as a variable in Kalmijn (2010, Ch. 2). In his thesis, Kalmijn

[2] We use the term 'anchor points' for the response options at both ends of a discrete scale. In case of a continuous distribution, we use the term 'extremes' to refer to the boundaries of the continuum that bounds this distribution.

devotes a section to our perspectives on the nature of happiness and satisfaction, the difference between intensity and extensity variables, the polarity of happiness scales and the level of measurement.

The response scale cannot be seen separately from the related leading survey question. The variations in the wording of the questions also lead to numerous different survey items. Furthermore, the time frame a question relates to leads to more variations. For example, the question can refer to satisfaction with life over the life time or just at this moment or in the last four weeks. Moreover, the question can contain a keyword such as the word 'happy' in the question "Are you happy with your life?", where the subject can either be or not be explicitly formulated in the response options, but likewise be formulated as "Do you feel …?" with the keyword only mentioned in the labels of the response options of the related scale.

These are just some examples of the variations in the wording of the questions used in happiness research. Of course there are many more variations one can think of and a comprehensive description of such questions and a discussion on these variations is given in Saris and Gallhofer (2007). Additionally a systematic overview of all the variations in survey items on happiness can be found in the collection 'Measures of Happiness' of the WDH. The measures are classified by six aspects, see Tab. 1, and the survey questions presented in this thesis are coded according to this classification.

**Table 1 Classification of survey questions on happiness in the World Database of Happiness**

| ***Aspect*** | ***Example*** | ***Code*** |
|---|---|---|
| Keyword used | satisfaction with life | O-SL |
| Time reference | currently | c |
| Method of assessment | single question | sq |
| Kind of rating scale | verbal | v |
| Length of rating scale | 4-step | 4 |
| Variant of rating scale | agree - disagree | a, b, ... etc |

We will illustrate the diversity in survey items and the comparability problem by an example based on time series for happiness and life satisfaction in The Netherlands in the next two sections. These sections also serve as an introduction to most of the survey items and time series we use throughout this thesis.

## 1.3 Time series on happiness and life satisfaction in The Netherlands

Happiness and life satisfaction have been measured periodically since the early seventies of the 20th century in The Netherlands. It was common practice to use items for these measurements which had verbal response scales. The present trend however, is to use 10- and 11-point numerical scales with only the anchor points defined by verbal labels to measure subjective well-being (Voorpostel et al, 2009; OECD, 2013; Van Beuningen, Van der Houwen, and Moonen, 2014). An increasing body of research (e.g. OECD, 2013) states that numerical scales are more suitable for measuring such constructs, because the use of numerical scales minimises categorisation effects, improves data analysis and produces more reliable data. Furthermore, numerical scales are especially suitable for telephone interviews, because they are time saving and prevent response-order effects (Scherpenzeel, 1999).

The eldest time series on life satisfaction comes from the Eurobarometer (EB), a series of public opinion surveys conducted in the member states of the European Union regularly on behalf of the European Commission which dates back to 1973. The standard version of the EB has, almost without exception, a spring wave and an autumn wave for every year since then (Schmitt, Scholz, Leim, and Moschner, 2008; European Commission, 2012a, 2012b, and 2013). The number of waves mounted up to 76 waves in the period from 1973 to 2012, using the same 4-point verbal scale item every wave, except for wave 76.2 in the autumn of 2011 in which the item had to be rated on a 10-point numerical scale. The EB was also the first survey used to measure happiness periodically. These measurements started in 1975 and were repeated every year until 1986 with interruptions in 1980 and 1981. There were 14 waves in total for this period, in all of which a 3-point verbal item was used. Happiness was also measured in version 66.3 of the EB in 2006. This was a special version of the EB in order to better understand the social realities of European Union citizens and a 4-point verbal scale was used to measure happiness. The measurement of happiness in this wave however, cannot be considered to be part of a time series, since in none of the other waves of the EB happiness was measured using the same scale.

The measurements with the EB were closely followed by a series of measurements by Statistics Netherlands (CBS) on life satisfaction in 1974 in The Netherlands in the first Life Situation Survey that CBS developed at the request of, and in close collaboration with, The Netherlands Institute for

Social Research (SCP)[3]. Happiness followed in 1977. The 5-point verbal scale items for happiness and life satisfaction were used in changing surveys by both organisations and with different periodicities over a period of almost 40 years (DeJonge, 2009). CBS used the item for happiness in 27 waves and the item for life satisfaction in 28 waves in the period from the first wave until 2010. After having conducted a split-half experiment in 2012, in which a verbal and a numerical scale were used for both topics, CBS decided to change to 10-point numerical response scales (Van Beuningen et al, 2014). In the same time frame, SCP had measured happiness 15 times and life satisfaction 22 times using the verbal scales, but it changed to a 10-point numerical scale for life satisfaction in 2002.

Happiness has also been measured in the Dutch Household Survey (DHS)[4] every year since 1993, resulting in 20 waves by 2012. In the 5 low frequency waves of the World Values Survey (WVS)[5] carried out in The Netherlands since 1981, happiness has been measured using a 4-point verbal scale and life satisfaction is measured using a 10-point numerical scale. Finally, items on happiness and life satisfaction form part of the European Social Survey (ESS)[6] which was fielded for the first time in The Netherlands in 2002 and since then, with a periodicity of 2 years. The ESS uses an 11-point numerical scale for both items and the results from 6 waves are available for the period 2002 to 2012. The aforementioned items constitute all the time series for happiness and life satisfaction in The Netherlands that we had to hand for this thesis. An overview of these items is given, excluding the 4-point verbal scale used in wave 66.3 of the EB, the numerical scale used in wave 76.2 of the EB and the new items with a numerical scale of CBS, in Tab. 2 and 3. In both tables, we have denoted the scale code, in line with the classification given in Tab. 1.

We have not mentioned the 10-point numerical survey items on happiness and life satisfaction from the European Quality of Life Survey[7] (EQLS) of Eurofound in the preceding. This survey examines both the objective circumstances of European citizens' lives and how they feel about those circumstances and their lives in general. The EQLS is a pan-European survey which came into existence in 2003 and is carried out since then every 4 years. Due to this periodicity of 4 years, we had only 2

[3] The frequency distributions of the items from CBS and SCP were obtained by personal communication.
[4] CentERdata - Institute for data collection and research, www.dhsdata.nl
[5] http://www.worldvaluessurvey.org/wvs.jsp
[6] http://www.europeansocialsurvey.org/
[7] http://www.eurofound.europa.eu/surveys/eqls/index.htm

measurements available in the period in which this PhD-research was conducted and therefore did not take into account the results from the EQLS.

The diversity in leading questions and the corresponding response options can clearly be seen from Tab. 2 and 3. Some of the leading questions refer to life as a whole framed, or not, in the present time, others just refer to the life currently lead. The diversity in survey items however, is even more manifest in the number and labelling of the response options. The number of response options of the items listed in the two tables varies from three to eleven and the wording of the response options varies in many ways. This makes some scales unipolar, such as that of the EB item on life satisfaction in which the word 'satisfied' is used in the wording of all response options. Furthermore, this EB item is symmetric but has, in contrast to, for example the DHS item on happiness, no neutral middle option. Other scales have clearly defined anchor points, of which the SCP item on life satisfaction is a good example, using the word completely in the labels of these points, which cannot be said of the asymmetric life satisfaction scale of the CBS item which has a response option at the lower end of the scale with a label that fails to express the lowest degree of satisfaction. These are just some examples of the variations in scales that are revealed when looking at the items in the two tables.

**Table 2 Items used in time series of happiness in the Netherlands, 1975-2012**

| **Survey** | **EB** | **WVS** | **CBS** |
|---|---|---|---|
| **Item code** | **O-HL-c-sq-v-3-ab** | **O-HL-u-sq-v-4-a** | **O-HP-u-sq-v-5-a** |
| **Survey question** | Taking all things together, how would you say things are these days? Would you say you are...? | Taking all things together, would you say you are: | To what extent do you consider yourself a happy person? |
| **Response options** | - Very happy<br>- Pretty happy<br>- Not too happy | - Very happy<br>- Quite happy<br>- Not very happy<br>- Not at all happy | - Very happy<br>- Happy<br>- Neither happy nor unhappy<br>- Not very happy<br>- Unhappy |
| | | | |
| **Survey** | **SCP** | **DHS** | **ESS** |
| **Item code** | **O-HP-u-sq-v-5-a** | **O-HP-u-sq-v-5-d** | **O-HL-u-sq-n-11-a** |
| **Survey question** | To what extent do you consider yourself a happy person? | Taking all together, to what extent do you think of yourself as a happy person? | Taking all things together, how happy would you say you are? |
| **Response options** | - Very happy<br>- Happy<br>- Neither happy nor unhappy<br>- Not very happy<br>- Unhappy | - Very happy<br>- Happy<br>- Neither happy nor unhappy<br>- Unhappy<br>- Very unhappy | 10 Extremely happy<br>.<br>.<br>.<br>0 Extremely unhappy |

**Table 3 Items used in time series of life satisfaction in the Netherlands, 1973-2012**

| **Survey** | **EB** | **CBS** | **SCP** |
|---|---|---|---|
| **Item code** | **O-SLL-u-sq-v-4-b** | **O-SLL-c-sq-v-5-d** | **O-SLL-c-sq-v-5-d** |
| **Survey question** | On the whole how satisfied are you with the life you lead? | To what extent are you satisfied with the life you currently lead? | To what extent are you satisfied with the life you currently lead? |
| **Response options** | - Very satisfied<br>- Fairly satisfied<br>- Not very satisfied<br>- Not at all satisfied | - Extraordinarily satisfied<br>- Very satisfied<br>- Satisfied<br>- Fairly satisfied<br>- Not very satisfied | - Extraordinarily satisfied<br>- Very satisfied<br>- Satisfied<br>- Fairly satisfied<br>- Not very satisfied |
| | | | |
| **Survey** | **SCP** | **WVS** | **ESS** |
| **Item code** | **O-SLL-c-sq-n-10-a** | **O-SLW-c-sq-n-10-aa** | **O-SLW-c-sq-n-11-cd** |
| **Survey question** | How satisfied are you with the life you currently lead? | All things considered, how satisfied are you with your life as a whole these days? | All things considered, how satisfied are you with your life as a whole nowadays? |
| **Response options** | 10 Completely satisfied<br>.<br>.<br>.<br>1 Completely dissatisfied | 10 Satisfied<br>.<br>.<br>.<br>1 Dissatisfied | 10 Extremely satisfied<br>.<br>.<br>.<br>0 Extremely dissatisfied |

## 1.4 The problem of incomparability of time series from different surveys

In survey research it is common practice to assign ranks to the response options of a discrete scale to calculate a sample mean, regardless of the semantics of the wordings used to label the options. The sample mean is accordingly calculated as the weighted average of the ranks of the response options using the relative frequencies as weights. In this common practice it is implicitly assumed that equivalent response options in equivalent scales on different topics are appreciated identically and that the response options are equally distanced. The degree of appreciation assigned to the words by which a response option is labelled, however, heavily depends on the context of the scale just as does the distance between two consecutive response options, see Ch. 5. This notion emphasizes the difficulty of comparing the outcomes of different surveys.

There are a number of other problems for trend analyses than just the ones mentioned above that complicate the comparability of survey outcomes and it is difficult to pool these results into long, consistent time series. Much of this becomes clear if we look at Fig. 1 and 2, in which the time series of the sample means according to the common practice sketched above, denoted the Rank Method, are presented for the items given in Tab. 2 and 3. Since for the EB item, for most years, there is more than one wave of measurements available, we calculated the un-weighted average of the frequency distributions for these items for each year to obtain one sample mean per year. In both figures we have given the number of response options followed by a p and an indication of whether it is a verbal scale or a numerical scale for each scale. For example, 3p-v denotes a 3-point verbal scale and 10p-n a 10-point numerical scale.

As can be seen from both figures it is obvious that there are considerable scale-effects if the Rank Method is applied to calculate a sample mean. From a quick glance, the effect of the difference in labelling may be not as obvious. Intuitively, the differences in sample means between a 4-point scale and a 5-point scale when applying the Rank Method will be more like those shown in Fig. 1, however, this is not the case for the sample means in Fig. 2. This is due to the difference in scales used between the surveys. This difference causes respondents who are satisfied with their life to select a response option with an, on average, relative high rank when they have to rate their life satisfaction on the 4-point EB scale and to select a response option with an, on average, relatively low rank when they have to rate their life satisfaction on the 5-point CBS scale, see Ch. 5.

**Figure 1 Mean happiness in The Netherlands based on the Rank Method**

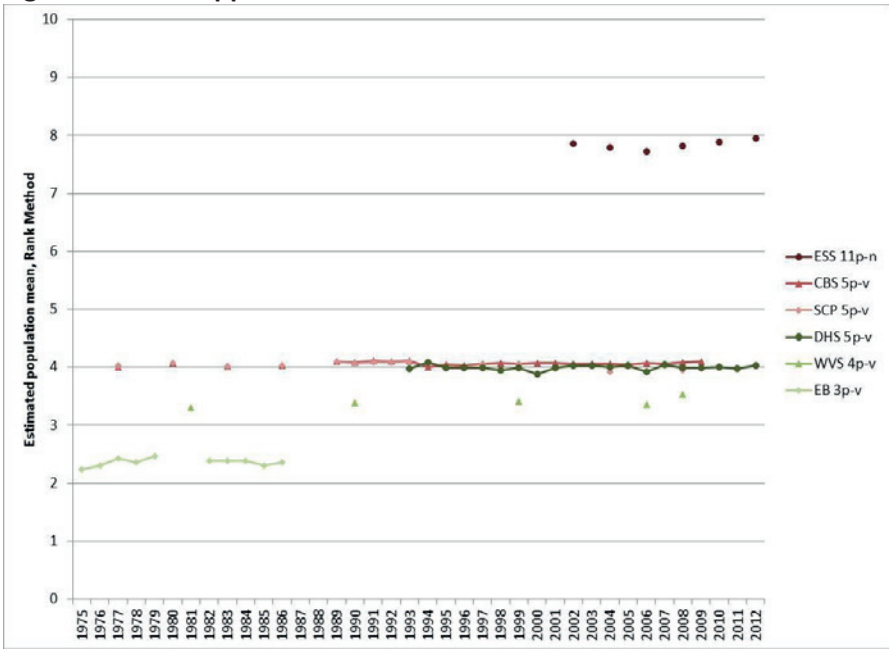
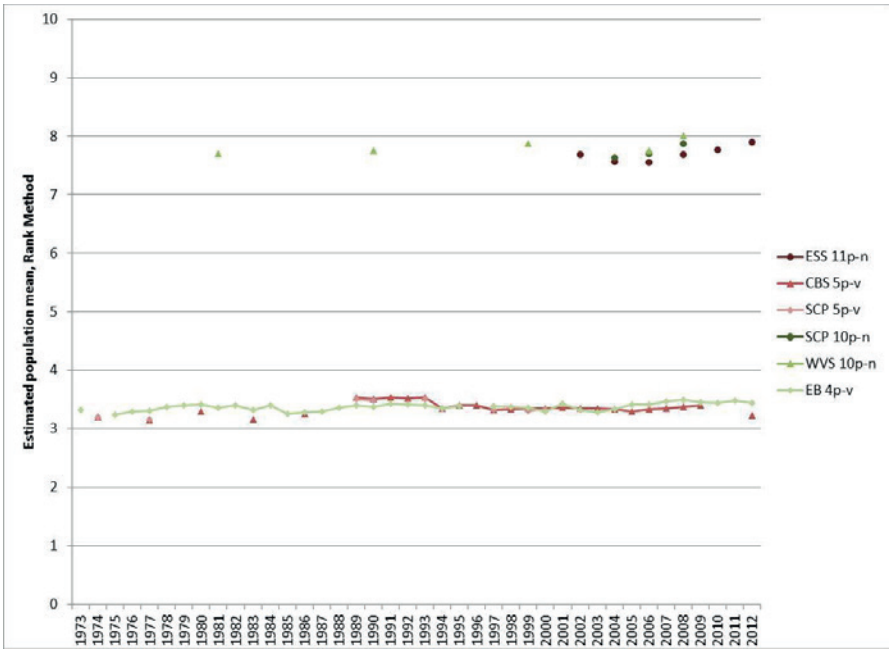



**Figure 2 Mean life satisfaction in The Netherlands based on the Rank Method**



Another fact which has to be taken into account when improving the comparability of time series is the periodicity of the measurements which differs per survey and which has changed over the course of time for some surveys. This may also cause discontinuities, which is clearly visible from Fig. 1 and 2. The above are just some of the issues that have to be addressed when searching for causes of incomparability, but they undoubtedly contribute to our understanding that the pooling of outcomes of different survey items into long consistent time series is not a straightforward exercise.

## 1.5 Conventional scale transformation methods

The limited uniformity of survey items and the other problems that affect the comparability of survey results reduces our ability to accumulate knowledge of such items and limits our analyses of trends. This calls for methods to transform ratings on different scales to attain comparable results and to correct for effects of changes in measurements and other influencing factors. Several methods have been developed to transform ratings on different response scales to a common one, typically a scale ranging from 0-10. Some of these scale transformation methods are applied in the World Database of Happiness, in particular in its collection of 'Happiness in Nations' (Veenhoven, 2015b). In this section we describe two conventional methods and explain why they fall short to overcome the comparability problem caused by the non-uniformity of survey items.

### *1.5.1 Linear Stretch*

A simple and commonly used conventional transformation method is the Linear Stretch Method, an early version of which was already in use almost a century ago (Hull, 1922). The Linear Stretch Method is most applicable for questions that use a numerical response scale. Scales with five or seven response options are typically stretched to obtain a common range for example from 0 to 10. This is done in such a way that the lowest number assigned to a response option is always projected onto 0 and the highest number onto the highest value of the range, and all the intermediate options are given equally distanced numbers in between: for a 5-point verbal scale the transformation to a 0-10 scale according to this method results in {0.0; 2.5; 5.0; 7.5; 10.0}. The sample mean after transformation of

the scale follows from the conventional Weighted Average Approach[8] according to which it is equal to the sum of the - transformed - values of all response options multiplied by their respective measured relative frequencies. When a verbal scale has to be transformed in this way, an initial step is to assign numerical values to verbal response options, typically using consecutive numbers, such as 4 for the happiest option on a 4-step scale and 1 for the least happy option.

The Linear Stretch Method has many serious disadvantages. The two most prominent of these disadvantages are one, the assumption made that the distances between the response options are equal, and two, even more problematically, the assumption that the labelling of the response options is irrelevant to the analysis, though not for the respondent. Despite these disadvantages, the Linear Stretch Method is still applied, for example it is used in the World Database of Happiness for numerical scales with at least seven points to transform them to comparable scales with a 0 to 10 range.

*The Percentage of Scale Maximum*

Another example of where the Linear Stretch Method is applied is in the percentage of scale maximum (%SM) method developed by Cummins (1997, 2003). In this method Likert scale data are transformed to a standard form with a range from 0 to 100. In the %SM-method a score of '0' is given to the lowest scale anchor up to 'n' to represent the highest scale anchor. Any mean score on this scale can subsequently be converted into %SM units by converting the score into a percentage of the scale maximum value as: %SM = (mean score/n)*100.

The %SM-method encounters the same disadvantages as mentioned above for the Linear Stretch Method, since it is based on a simple linear relationship between the response options and disregards their labelling. This can be illustrated by a practical example from an e-mail discussion about the discrepancy between the %SM-scores on satisfaction with life as a whole for Australia and New Zealand[9]. In the Australian Unity Wellbeing Index (AUWI) project, survey 21, there was a %SM-score of 78% (Cummins, 2009). Applying the %SM-method to the results found in the

[8] The Weighted Average Approach is a generalization of the Rank Method, by not requiring that the numbers assigned to the response options are equal to the ranks of these response options to calculate a sample mean.

[9] E-mail discussion of December 2009 between participants to the OECD/ISQLS/ISTAT meeting "Measuring subjective well-being: an opportunity for National Statistical Offices?", Florence, 23-24 July 2009.

New Zealand General Social Survey[10] (NZGSS) 2008, would give a rather lower score of 70%, whereas the discussants would expect the two countries to be virtually identical on such measures. The reason for the discrepancy was found in the differences in scales that were used to compute the results. In the AUWI a 0 to 10 numerical scale was used, whereas the NZGSS used a 4-point verbal scale with response options labelled as (1) dissatisfied/very dissatisfied, (2) neither satisfied nor dissatisfied, (3) satisfied and (4) very satisfied. When the %SM-method was applied these response options were converted to, respectively 0, 33.3, 66.6 and 100, irrespective of their labelling. The remark was made by one of the discussants that 33.3 was perhaps a low score for respondents that state they are neither satisfied nor dissatisfied and that a value of 50 would seem to be fairer.

As a solution it was suggested that the original NZGSS scale could be considered as a 5-point scale with the lowest two categories grouped and given scores of 100, 75, 50 and 12.5. In this way the middle category would get 50 points and the lowest categories, 0 and 25, would be averaged out. Under this method, the average score would equal 77.0, which would be very close to the Australian average, as might be expected. From this practical example, it becomes very clear that, when comparing the results of different surveys, the labelling of the response options cannot be neglected when converting verbal scale outcomes to a common numerical scale.

### *1.5.2 Semantic Judgement of Fixed Word Value*

Several attempts have been made in the course of happiness research to develop better methods to cope with the heterogeneity in response scales. What many of these alternative methods have in common is that they make use of expert ratings (Veenhoven, 1993; Bălţătescu, 2002; Lim, 2008) by getting a group of experts to rate the verbal labels of response options on a common numerical scale.

An early example of such a method is that of Jones and Thurstone (1955) who requested approximately 900 respondents to rate 51 verbal qualifications on a 9-point Likert scale separately. A value on a common interval scale and a standard deviation were calculated for each qualification. The result was a list of the 51 qualifications ordered on the

10 http://www.stats.govt.nz/browse_for_stats/people_and_communities/Households/nzgss_HOTP2008.aspx, Excel tables for NZGSS 2008 HOTP.xls, Table 10 cont.

basis of their value on the common interval scale. In the work done by Lodge (1981) another illustration of this method can be found. We have classified this method as the Semantic Judgement of Fixed Word Value Method, which is also applied in the World Database of Happiness to obtain comparable average scores. Veenhoven (1993) and 12 co-workers rated the degree of happiness denoted by the verbal labels of 29 commonly used survey items on a numerical 0 to 10 scale. For example, the label 'Very happy' was an option in eight of the 29 items and was given a rating varying from 9.2 to 9.4 resulting in an overall mean of 9.3, whereas for the label 'Not very happy' an overall mean of 3.7 was found. To this day, these results are used to transform responses in the WDH for scales for which linear stretching falls short.

The Semantic Judgement of Fixed Word Value Method overcomes the disadvantages of presumed equidistance and the neglecting of the labels that are associated with the Linear Stretch Method. The Semantic Judgement of Fixed Word Value Method, however, also has some weak points. Kalmijn (2010, p. 118) mentions that the fixed values applied in the World Database of Happiness:

- are based on expert judgements that do not necessarily reflect the views of non-expert respondents
- have been rated by Dutch experts on basis of the English version of the questions, thus implicitly assuming that the feelings associated with an item are not affected by its translation from Dutch into English
- do not take into account the phrasing of the lead question, nor the number and the labels of the alternative response options and their position on the scale.

### 1.6 The need for further innovations

The weaknesses of these early transformation methods also appear when the transformed scores are compared to average ratings on 0-10 numerical scales in the same country in the same year (Kalmijn , Arends, and Veenhoven, 2011). All these weaknesses of conventional methods instigated further innovations, which will be discussed in this thesis.

# PART 2

# INNOVATION 1: THE HAPPINESS SCALE INTERVAL STUDY

# 2 The Happiness Scale Interval Study

## 2.1 Introduction to the Happiness Scale Interval Study

In order to counter the shortcomings of the Semantic Judgement of Fixed Word Value Method, Veenhoven (2008) started the Happiness Scale Interval Study (HSIS). This study was set up to look at survey questions on happiness using verbal response options, such as 'Very happy' and 'Pretty happy' with the intent to determine consistently what degrees of happiness are denoted by such terms when based in different questions and languages. In the HSIS persons who are referred to as 'judges' are asked to rate the degree of happiness denoted by each of the verbal response options in the context of the full item. The judges are asked to identify the interval on a 0-10 range that corresponds to a verbal response option such as 'Very happy' using a web-based Scale Interval Recorder (Veenhoven and Hermus, 2006). This method is discussed in detail in (Kalmijn, 2010; Kalmijn et al, 2011) and we have classified it as the Semantic Judgement of Word Value in Context Method (DeJonge, Veenhoven, and Arends, 2014a). In this thesis, however, we will refer to it as the Scale Interval Method.

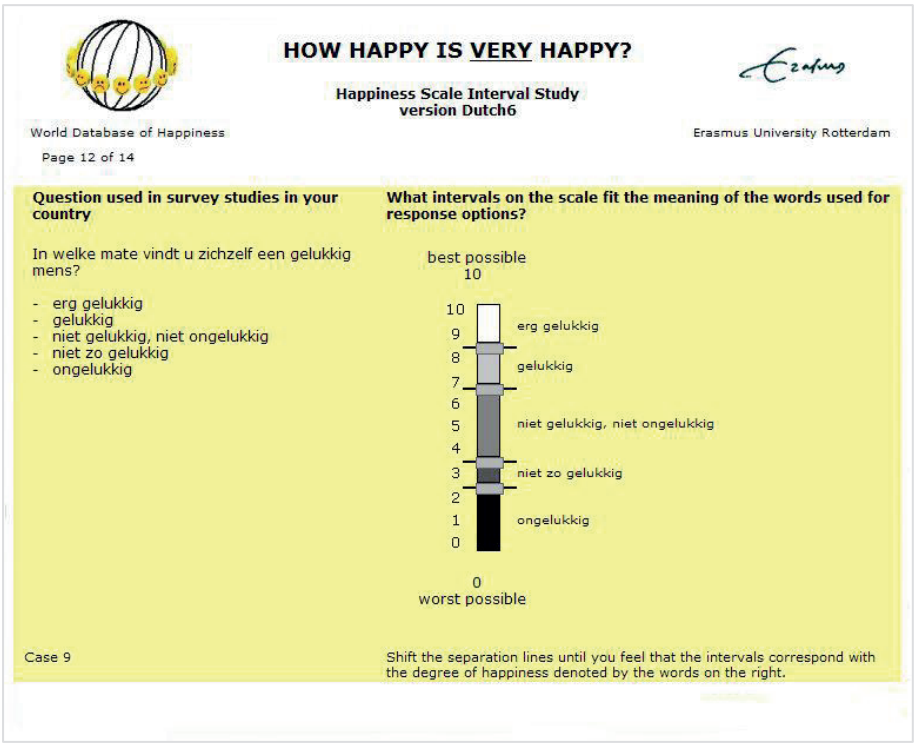
### *2.1.1 The Scale Interval Recorder*

A series of survey items is presented on a computer screen to the judges. Items are presented one by one on the left side of the screen and each item presented consists of a question and its corresponding verbal response scale with options given in the judges' mother tongue. A screen shot of the Scale Interval Recorder from a study presented to Dutch judges is given in Fig. 3. On the right side of the screen a vertical bar scale is displayed with a number of small horizontal sliders on it, the number of which is equal to the number of response options minus one. The judges have to shift the sliders until they feel that the intervals on the vertical bar correspond to the meaning of the words as used for the verbal response options. Note, the response options that are displayed next to the bar move simultaneously with the sliders to the level of the mid-interval value of each interval.

Looking at Fig. 3 it can be seen that the extremes of the numerical bar scale are labelled 'Worst possible' and 'Best possible'. In the terminology of Saris and Gallhofer (2007) these labels are called 'fixed reference points'. What worst and what best means, is left to the interpretation of the judges. The labelling of the extremes is thus semi-abstract which makes them applicable to all questions presented to the judges and independent of the subject of an individual question. An

additional advantage of this semi-abstract labelling is that the judgement is not influenced by the extremity of the wording used for the labels of both extremes of the continuum.

**Figure 3 Screenshot of the Scale Interval Recorder**




### *2.1.2 Difference with conventional methods for scale transformation*

The approach to scale transformation used in the HSIS differs essentially from that used in the Linear Stretch Method and the Semantic Judgement of Fixed Word Value Method, as the response options in the primary scale are not considered to be discrete points, but to be intervals each representing a part of the continuum from 0 to 10 where the perception of happiness can be found. This complies with the view of Kalmijn (2010), who considers happiness to be a latent continuous variable that underlies the survey questions being studied. Moreover, in the Happiness Scale Interval Study each response option is judged in the context of the other response options of the scale and this approach is illustrative of the Scale Interval Method.

## 2.2 Three scale transformation methods applied to empirical data

To illustrate how the three methods, Linear Stretch, Semantic Judgement of Fixed Word Value and Scale Interval, are used we selected two of the survey items on life satisfaction we presented in Tab. 3, the first of which is the item from Statistics Netherlands (CBS) and the second of which is the item from the Eurobarometer (EB). The CBS item has an asymmetric response scale with five options. The EB item has a symmetric response scale without a neutral midpoint and four options. The items are summarized in Tab. 4 which also includes the frequency distributions for this data for 2008.

**Table 4 Survey items on satisfaction with life used in The Netherlands in two surveys**

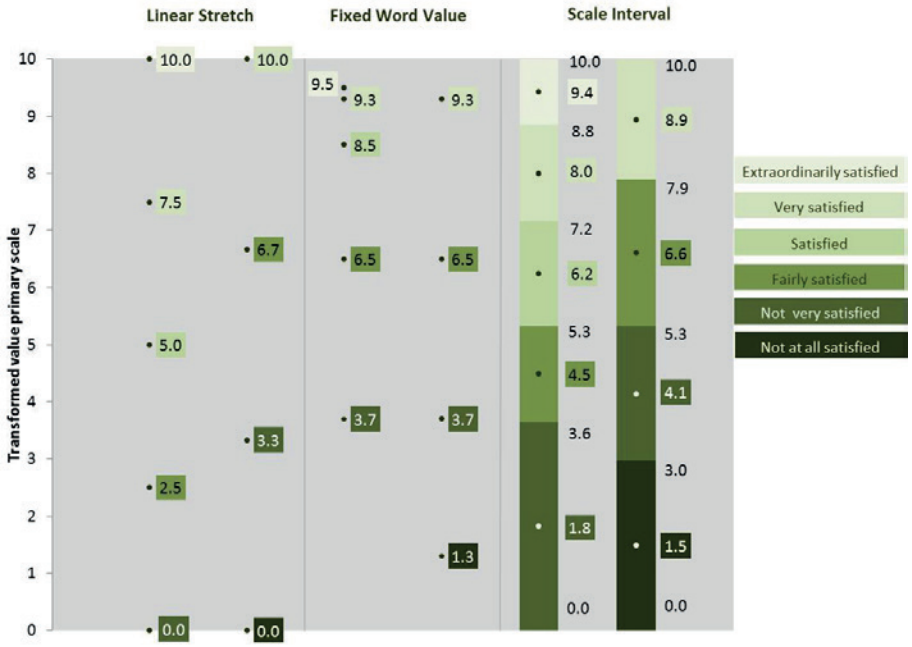
| *Item code Survey* | *Question* | *Response options* | *Frequencies 2008* |
|---|---|---|---|
| O-SLL-c-sq-v-5-d CBS | To what extent are you satisfied with the life you currently lead? | - Extraordinarily satisfied<br>- Very satisfied<br>- Satisfied<br>- Fairly satisfied<br>- Not very satisfied | 8.4%<br>35.5%<br>45.1%<br>7.6%<br>3.4% |
| O-SLL-u-sq-v-4-b EB | On the whole how satisfied are you with the life you lead? | - Very satisfied<br>- Fairly satisfied<br>- Not very satisfied<br>- Not at all satisfied | 51.5%<br>44.8%<br>3.1%<br>0.6% |

The two items presented in Tab. 4 together comprise six response options, three of which are included in both items. The transformation of the response scales of the items to a scale from 0 to 10 according to each of the three transformation methods is depicted in Fig. 4.

From Fig. 4 it can be seen that in the Linear Stretch Method the anchor points of both primary scales are pinned to 0 and 10 and that all the other response options are equally spaced in between. When the Linear Stretch Method is applied the response option 'Fairly satisfied' of the 5-point scale is assigned the transformed value 2.5, whereas this option for the 4-point scale gets a transformed value of 6.7. This large difference between the values 2.5 and 6.7 is elucidatory for the fact that the wordings

of the response options are neglected when the Linear Stretch Method is applied.

**Figure 4 Comparison of transformations using three methods**




If the Semantic Judgement of Fixed Word Value Method is applied the results are entirely different. The value of a label such as 'Fairly satisfied' is fixed in this method and equal to 6.5 according to the Dutch experts, however, from Fig. 4 it can also be revealed that the Semantic Judgement of Fixed Word Value Method treats each response option as isolated from the number and the wording of the other options and thus does not take into account the context of the scale.

As can be seen in the Scale Interval Method the assumption of equal distances between response options and the idea that a fixed value applies to a label of a response option irrespective of the labelling of the other options are abandoned. If we consider the response option 'Fairly satisfied' once more, we can see that this option is assigned the interval 3.6 to 5.3 for the 5-point scale, with a mid-interval value equal to 4.5 and a length of 1.7. For the 4-point scale the interval for this option ranges from 5.3 to 7.9, with a mid-interval value of 6.6 and a length of 2.6.

## 2.3 The three transformation methods and the comparability problem

At the start of the Happiness Scale Interval Study in the Semantic Judgement of Word Value in Context Method a sample mean after transformation of the primary scale followed from the conventional Weighted Average Approach according to which it is equal to the sum of the mid-interval values of all intervals of the transformed scale, each of which is weighted with its corresponding relative frequency. This is analogous to how this is done in the Linear Stretch Method and the Semantic Judgement of Fixed Word Value Method. A comparison of the results obtained using these three scale transformation methods is shown in Tab. 5.

**Table 5 Transformed means obtained using different transformation methods (frequencies 2008)**

| *Item code Survey* | *Linear Stretch* | *Semantic Judgement of Fixed Word Value* | *Scale Interval* |
|---|---|---|---|
| O-SLL-c-sq-v-5-d CBS | 5.9 | 8.6 | 6.9 |
| O-SLL-u-sq-v-4-b EB | 8.2 | 7.8 | 7.7 |

The survey items from CBS and the EB address more or less the same topic and mainly differ in the response scales. The results for each item are assumed to be representative for the Dutch population and therefore one would expect that given that a transformation method is applied, the transformed means for 2008 would be equal. This is clearly not the case. The difference of 2.3 between the transformed means of 5.9 and 8.2 based on the Linear Stretch Method is most striking. It is obvious from these results that none of the three transformation methods offer a solution to the comparability problem.

## 2.4 Discussion

The labels of the response options will not be interpreted in the same way by all respondents. Some people may consider the labels of all the response options of the CBS scale to be positively formulated, whereas others may

interpret the two options at the lower part of this scale as negative expressions of satisfaction with life. Some people may believe one cannot be less satisfied than 'Not at all satisfied' and will consider this option to be the null point of the EB scale, while others may believe things can be worse and assign an interval of positive length to this option. Interpretation of semantic intervals will vary from person to person for all kinds of reasons such as personality, cultural context or the context of the scale (Hazelrigg and Hardy, 2000). As a consequence, in the Happiness Scale Interval Study items are assessed by a group of judges. This results in a report of the average value and the variance for each boundary between two response options. This implies that the results should be considered as representative for the population the judges belong to and are not applicable for subgroups with specific characteristics.

All of the three transformation methods have in common that a discrete primary scale is transformed into a discrete secondary scale and that the sample mean is calculated on the basis of all ratings of this 'secondary scale'. This sample mean is adopted as the estimator of the mean happiness value of the happiness distribution in the population. The variance and standard deviation of the latter distribution are estimated accordingly. The three methods also have in common that they do not offer a solution to the comparability problem as we have shown in Tab. 5.

Solving the comparability problem requires another approach, which we will go into from Ch. 6 on. The Scale Interval Method, however, offers some interesting applications which allow a view on the size of the comparability problem. We will address these applications in Ch. 3 to 5.

# 3 Use of Happiness Scale Interval Studies in this thesis

## 3.1 Research questions addressed in this thesis using HSIS-results

The main aim of the Happiness Scale Interval Study (HSIS) is to improve the comparison of happiness across nations. Therefore, the items included are restricted to those that have been applied in studies of general populations in nations. Since it came into existence the HSIS has been conducted in a number of countries. A complete list of all items ever considered since the start of the study in 2005 can be found on the website of the World Database of Happiness, in the section 'Scale Interval Study'[11].

HSIS-results from studies conducted in The Netherlands, Spain and Chile were very useful to be used to address two of the research questions formulated and answered in this thesis.

- Can response scales which appear to be equivalent also be considered to be equivalent when interpreting and mutually comparing survey results? We will go into this question in Ch. 4.
- To what extent does the meaning attached to identical response options differ when used in the context of non-identical response scales? We have worked this out in Ch. 5.

In this section we go into the selection of the survey items we used to address the two research questions and the recruitment of judges for the studies in which these items were included.

## 3.2 HSIS-studies used and the selection of survey items

Statistics Netherlands and the Erasmus University Rotterdam conducted a HSIS-study in 2010 which covered a total of twenty survey items which were equally distributed on the sub-studies dutch6 and dutch7. A division of the results into dutch6 and dutch7 was not relevant for the analyses. The initial division into two studies was mainly done to prevent there being too many questions to judge within one study and dutch6 and dutch7 were seen to be essentially part of the same study for the purpose of the analyses. The Dutch HSIS-study was set up to address a number of research questions among which the research questions formulated above. Eighteen of the twenty items in the study were taken from past and recent national surveys fielded in The Netherlands. Two additional items were included aimed at addressing the first of the two research questions. The first of these additional items is based on an item on life satisfaction included in

[11] The direct link to this section is: http://worlddatabaseofhappiness.eur.nl/scalestudy/scale_fp.htm.

the study, but with the keyword 'satisfied' replaced by the keyword 'happiness'. Similarly, the second additional item is based on an item on happiness in the study, but with the keyword 'happiness' replaced by the keyword 'satisfied'. By the addition of these two extra items, it was guaranteed that at least two pairs of items in the study meet the conditions for the comparison of survey results, see Sec. 4.2. A third pair which meets these conditions was found among the other sixteen items included in the study. A complete overview of all the twenty survey questions used in the two sub-studies of this Dutch study can be found in the 'study list' on the website of the HSIS.

The results of the Dutch study would only be valid for the Dutch population. Since the HSIS has been conducted in a number of countries it would be interesting to compare the findings from all these studies to investigate the influence of culture and language on the interpretation of verbal response scales. In other languages and cultures the results may be different from what we found for The Netherlands. For this thesis however, we only studied results from HSIS-studies as above conducted in other countries for the question about equivalent response scales. For this purpose we looked for pairs of items on happiness and life satisfaction that meet the conditions formulated in Sec. 4.2 in other HSIS-studies conducted previously in other countries. None of these other studies though, were set up with the research question of Ch. 4 in mind and we could find no more than two other pairs of items meeting the conditions of Sec. 4.2. Both of these pairs were included in HSIS-studies that have been conducted in Spain in 2009 and in Chile in 2007. One of the pairs of items from these studies in Spanish is equivalent to the first pair included in the Dutch study. Although we found no more than two other pairs of items, since these pairs were included in both the Spanish study and the Chilean study, it was possible not only to compare whether differences in results can be attributed to a difference in the language used for the assessment of the scales, Dutch versus Spanish, but also whether culture plays an influential role in possible differences between the results for Spain and Chile.

An overview of the original Dutch and Spanish wording of the selected items for this thesis and their translation into English and preceded by the code that has been assigned to them in the collection 'Measures of Happiness' of the World Database of Happiness, is given in appendix A.

### 3.3 Recruitment of judges and their representativeness

Items from HSIS-studies are in general assessed by students recruited from the university which conducts the study. An objection to their employment as research participants in social science research is that it is doubtful whether they are representative of the general population because they are on average younger, better educated, to tertiary level, and come from more privileged backgrounds than most people in their respective populations (Cummins 2003). The participation of Statistics Netherlands in the Dutch HSIS-study however, made it possible to recruit not only judges from the students of the Erasmus University Rotterdam but also from employees of Statistics Netherlands and employees of the Netherlands Institute for Social Research.

The judges recruited from the two groups of employees are probably also not representative of the general population, but they at least represent a different group than the judges recruited from the students. Although there were some small differences between students and employees in the mean values they assigned to the boundaries between response options, as can be seen in Tab. B.1 of appendix B, the conclusions that can be drawn from the judgements of both groups are equivalent. This gives confidence that the conclusions based on the Dutch HSIS-study will be valid for the general population. Nevertheless, to ascertain that the outcomes are fully valid for the general population a Happiness Scale Interval Study is required with a group of judges that represents the general population, although this may be difficult to organise.

The judges for the study conducted in Spain were recruited among students from the University of Granada in Spain. The judges for the Chilean study were recruited among students from the Universidad Catolica del Norte of Antofagasta in Chile. Although for these studies no other groups of judges were recruited, we adopt the conclusion drawn from the Dutch study that the outcomes for the general population will not go in a totally different direction than those for the judges recruited for the studies.

We want to stress that the Scale Interval Recorder is an instrument to assess how people interpret words in a common language. We believe that the appreciation that groups of people have of their own happiness may on average differ from that in the general population, but that groups do not differ substantially in the meaning they assign to words in a common language.

The employees in the Dutch study could participate in both the studies dutch6 and dutch7. As a result the number of participants in each study was rather large, amounting to 392 judges for dutch6 and 359 for dutch7. The number of judges that participated in the Spanish study was with a total of more than 400 participants even higher. In contrast to this, the number of judges in the Chilean study with valid assessments was rather low and for the items on happiness accounted to slightly less than 25 judges.

# 4 Equivalence of rating scales using different keywords

## 4.1 The keywords 'happiness' and 'satisfaction with life'

The World Database of Happiness (WDH) focuses on happiness in the sense of subjective enjoyment of one's life as a whole (Veenhoven, 1984). In this definition 'happiness' is synonymous with 'life satisfaction'. This definition is supported by a semantic analysis conducted by Storm, Jones, and Storm (1996) who found that their English-speaking Canadian participants interpreted 'satisfaction' as a kind of 'happiness', but also 'happiness' as a kind of 'satisfaction', which implies that the two terms share a large part of the semantic space. Other researchers however, for various reasons conclude that these terms do tap different aspects of subjective well-being. Among them are McKennel and Andrews (1980) and Fischer (2009) who suggest that satisfaction is strongly related to the cognitive/rational component of a subjective evaluation of one's life, while happiness is more related to the affective/emotional component, and Saris and Andreenkova (2001) who believe that the relationship between the two terms varies with the cultural and linguistic environment in which it is studied. The latter fits with the conclusion of Wierzbicka (2004) that adjectives in some languages are restricted for exceptional states, for example 'felice' in Italian and 'счастливый' (transcribed as 'sčastlivyj') in Russian, whereas the equivalent in another language, such as 'happy' in English, is weaker and needs an additional word such as 'very' to strengthen the meaning. In addition to this, we believe that groups of people may on average differ from the general population in how they appreciate their own happiness, but that they do not differ substantially in the meaning they assign to words in a common language.

### *4.1.1 Surveys that measure 'happiness' or 'satisfaction with life' in an equivalent way*

To determine whether or not happiness and satisfaction with life do tap different aspects of subjective well-being, an obvious step is to measure both topics in one survey and to compare the response patterns. The comparability of such a study however, largely depends on the variations in survey items of which the difference in the keywords used is just one. Each variation will influence the response patterns and therefore an approach to verify that a difference in response patterns has to be attributed to the difference in the keywords used, needs to exclude other variations in the

items. In the case of happiness and life satisfaction, this would require a study that includes items for both terms which have equivalent response scales and only differ in the keyword used. The method we present in this thesis focuses on response scales with verbal labels for all response options and thus excludes response scales with no verbal labels or only verbally labelled anchor points. In the collection 'Measures of Happiness' of the WDH only a few studies can be found that meet both these requirements. These eighteen studies are presented in Tab. 6.

**Table 6 Survey studies that involve questions on both 'happiness' and 'life satisfaction'**

| The Leisure Development Centre (1980) (13 studies) | | Ventegodt (1996) (1 study) | |
|---|---|---|---|
| **How happy do you feel as you live now? Please choose one item on this card that comes closest to your feeling.** | **Overall, how satisfied are you with your present life.....?** | **How happy are you now?** | **How satisfied are you with your life now?** |
| 1 very unhappy | 1 very dissatisfied | 1 very unhappy | 1 very dissatisfied |
| 2 fairly unhappy | 2 fairly dissatisfied | 2 fairly unhappy | 2 fairly dissatisfied |
| 3 neither happy nor unhappy | 3 neither satisfied nor dissatisfied | 3 neither happy nor unhappy | 3 neither satisfied nor dissatisfied |
| 4 fairly happy | 4 fairly satisfied | 4 fairly happy | 4 fairly satisfied |
| 5 very happy | 5 very satisfied | 5 very happy | 5 very satisfied |

| Michalos and Zumbo (1999, 2003); Michalos (2003); Michalos and Orlando (2006) (4 studies) | |
|---|---|
| **Considering your life as a whole, how happy would you say you are?** | **How satisfied are you with your life as a whole?** |
| 1 very unhappy | 1 very dissatisfied |
| 2 somewhat unhappy | 2 somewhat dissatisfied |
| 3 a little unhappy | 3 a little dissatisfied |
| 4 about evenly balanced | 4 about evenly balanced |
| 5 a little happy | 5 a little satisfied |
| 6 somewhat happy | 6 somewhat satisfied |
| 7 very happy | 7 very satisfied |

Source: World Database of Happiness, collection Measures of Happiness (Veenhoven 2015b)

The means for the ratings on happiness and life satisfaction have been reported in all the eighteen studies. In fourteen out of the eighteen pairs people rate their happiness slightly higher than their satisfaction with life.

In half of the cases, all from the studies done by the Leisure Development Centre (1980), the difference in reported means for happiness and life satisfaction was less than 0.1 point. In the WDH gamma correlations between happiness and satisfaction with life were calculated for all the studies of the Leisure Development Centre. These gamma correlations fluctuate between 0.62 and 0.88. Michalos and Orlando (2006) and Ventegodt (1996) reported Pearson correlation coefficients for the relationship between happiness and life satisfaction of 0.73 and 0.72 respectively. There is no clear relationship between these correlation coefficients and the differences between the means of happiness and life satisfaction.

In most of the studies contained in Tab. 6, happiness and life satisfaction have been related to satisfaction with a number of life domains, such as family, social relations, income, education, health, politics, et cetera. Although some studies show a small difference between mean happiness and mean satisfaction with life, both topics may have a rather different relation to the satisfaction with one of those life domains: in France for example, the differences in means in the study of the Leisure Development Centre was only 0.02. The gamma correlation between happiness and the level of education however, turned out to be 0.21, whereas the gamma correlation of this life domain with satisfaction with life was no higher than 0.04.

Recapitulating, several studies that use equivalent response scales for happiness and satisfaction with life reveal different response patterns for these topics, quantified by a difference in mean or a difference in correlation with satisfaction with other life domains and in some studies a correlation coefficient for the relation between happiness and life satisfaction. Based on this, we assume that happiness and satisfaction with life are likely to tap different aspects of subjective well-being, although according to Saris and Andreenkova (2001) this may depend on the cultural and linguistic environment in which they are studied.

### *4.1.2 The problem*

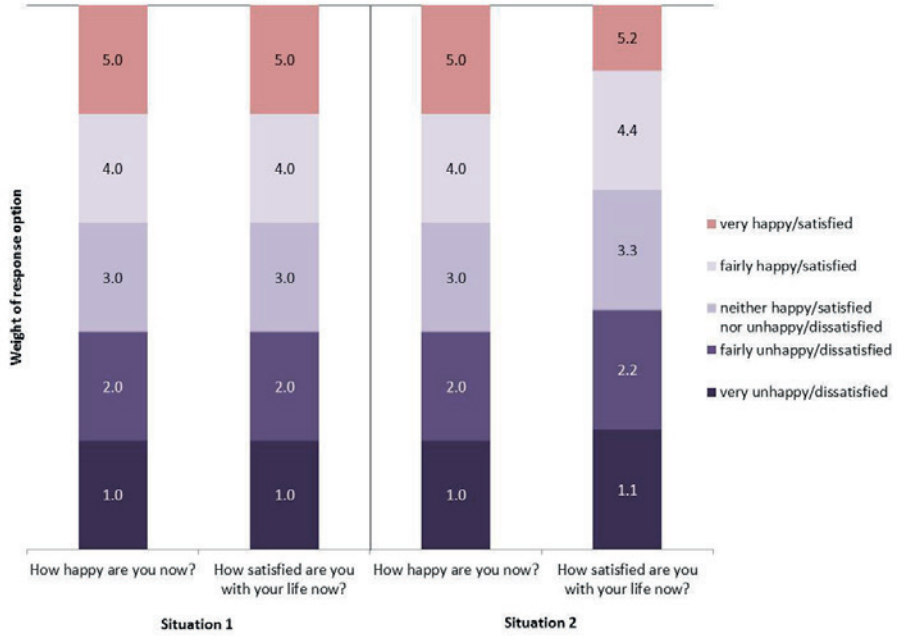
The assumption that happiness and life satisfaction tap different aspects of subjective well-being, gives rise to the question “Can response scales which appear to be equivalent also be considered to be equivalent when interpreting and mutually comparing survey results?”

The means and correlations, as discussed in Sec. 4.1.1 are, according to common practice, all based on the ranks of the response options, disregarding the degree of appreciation assigned to each response

option. In this common practice it is implicitly assumed that equivalent response options in equivalent scales on different topics are appreciated equivalently. Yet, this may contradict reality. We have made this tangible in Fig. 5 with an example making use of the two 5-point scales from the studies of the Leisure Development Centre and Ventegodt presented in Tab. 6.

On the left side of Fig. 5 we visualized the situation in which the corresponding rank is assigned to each response option. Now assume that the response options for life satisfaction represent a slightly different degree of appreciation than the response options for happiness for example as is depicted on the right side of Fig. 5. In the hypothetical example in situation 2 the ratings assigned to the response options are no longer equal to their respective ranks but to 1.1, 2.2, 3.3, 4.4 and 5.2.

**Figure 5 Degree of appreciation of response options**




For a fair comparison of the means for happiness and life satisfaction it seems reasonable to take into account the differences in degree of appreciation of the response options of each scale used. We will illustrate this using the frequency distributions for happiness {0.7%, 3.1%, 31.4%, 44.3%, 20.5%} and life satisfaction {1.0%, 6.8%, 15.3%, 48.4%, 28.5%} ordered in ascending order of the ranks of the response options taken from

the study by Ventegodt (1996). According to common practice, these frequency distributions would result in a mean value of 3.8 for happiness and a mean value of 4.0 for life satisfaction. If we used the ratings for life satisfaction from situation 2 instead of the ranks of the response options, the mean value for life satisfaction would be equal to 4.2. In comparison to common practice, the difference between the means of happiness and satisfaction with life has been doubled from 0.2 to 0.4 by taking the degree of appreciation of the response options into account.

This example shows that when interpreting and mutually comparing survey results, it should be taken into account if a different keyword in an otherwise equivalent response scale comes with a different degree of appreciation represented by each response option. Therefore, to be sure that when comparing survey outcomes the results will not be muddled due to incomparability of the scales or differences in the interpretation of each scale by the respondents, the interpretation of the scales by respondents has to be examined and discussed carefully beforehand.

In this chapter we focus on a new method that is suitable to do such an examination.

## 4.2 The Scale Interval Recorder as an instrument to compare the degree of appreciation expressed by equivalent response options

The Scale Interval Recorder was initially developed to compare the degree of appreciation expressed to, identical or non-identical, response options in different response scales on the same topic, for example happiness or life satisfaction. The instrument however, can also be used to investigate whether equivalent response options in equivalent scales that only differ in the keywords used, denote an equal degree of appreciation. For this purpose, we require that a pair of survey items meets the following conditions:

- the question of each item must correspond to the topic, in this thesis happiness or satisfaction with life, used for the labels of the response options
- the questions posed in both items must refer to the same time frame, for example life as a whole, the past four weeks or at this moment
- the response scales of both items must be equivalent, having the same number of response options, which by themselves may only differ in the topic they refer to

Once one or more pairs of items that fulfil these conditions have been selected, the Scale Interval Recorder can be used for a group of judges to assess the response scales of these items.

We explain the method for testing the equivalence of rating scales using different keywords which we described in Sec. 4.1.2 using a mutual comparison of the assessment of the scales of the three pairs of items presented in the Dutch HSIS-study we discussed in Sec 3.2 and of the two pairs of items assessed by Spanish-speaking judges we also discussed in Sec. 3.2.

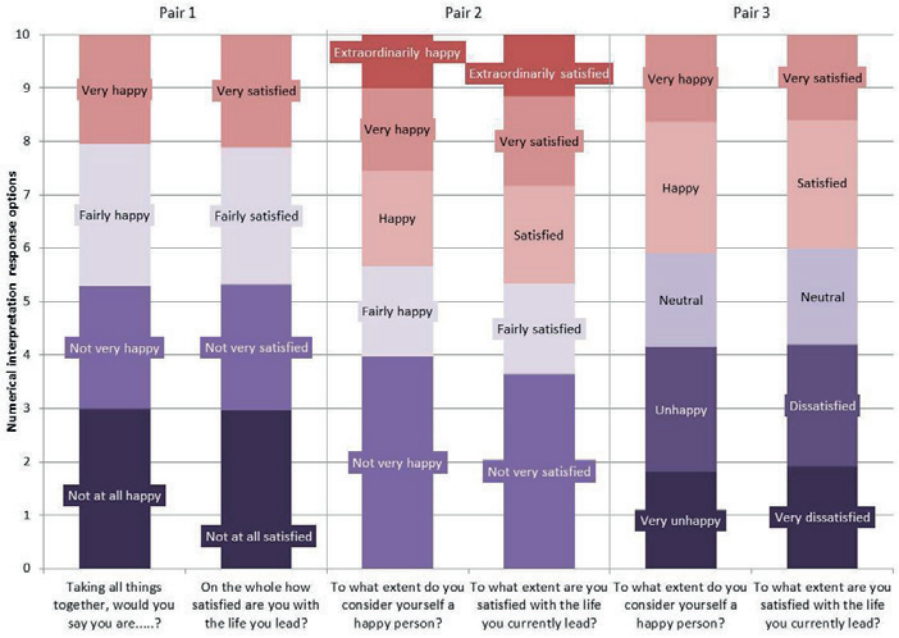
We first considered the questions in Dutch and compared both level and dispersion of the assessments of the boundaries between the response options. Next we compared the assessment made by the Dutch judges to the assessment made by the Spanish-speaking judges. Note: an overview of the wording of the items in Dutch/Spanish and their English equivalents, and the number of judges that assessed each item is given in Tab. A.2 and A.3 of appendix A.

### 4.3 Differences in degree of appreciation of response options labelled in Dutch

The assessments by Dutch judges of scale intervals for the three pairs of items are depicted in Fig. 6 by juxtaposing the average assessments for the response options in each pair. The left item in each pair addresses happiness and the right item life satisfaction

Only the scales of the last pair in Fig. 6 are symmetric and have an explicit neutral midpoint. For the first and third pair of items the degree of appreciation assigned to happiness is equal to that of satisfaction with life. This equality is not so strict for the second pair of items in Fig. 6 which have very asymmetric scales, but, although the boundaries between the response options for this pair of items not fully coincide, the difference in degree of appreciation for this pair is still not noteworthy. This difference could even be designated as negligible when compared to the differences in degree of appreciation of similar response options in dissimilar scales, for example the response option ‘Fairly satisfied’ in the 4-point scale of the first pair is appreciated very differently compared to the same option in the 5-point scale of the second pair. In other words, the meaning given by respondents to the labels of response options is not fixed, but depends on the number of options to choose from and the labels attached to the other options. Altogether, this makes it plausible that for the Dutch population degrees of happiness are comparable to degrees of life satisfaction.

***Figure 6 Comparison of the interpretation of verbal response scales by Dutch judges***




In Tab. 7 the average assessments and the standard errors of their estimated mean values are given of the upper boundaries for the three pairs of items depicted in Fig. 6. The options have been numbered from worst to best, for the first pair of items option 1 thus refers to 'Not at all' happy/satisfied and option 4 to 'Very' happy/satisfied. For the two other pairs, the response option that represents the best situation is numbered 5.

The averages in Tab. 7, in combination with the very small standard errors, underpin the conclusion that Dutch judges on average appreciate the scales for happiness and satisfaction with life equally. This does not mean that the perception Dutch people have of how happy they are, is identical to how they judge their own satisfaction with life. It only points out that the comparisons of the perceptions that Dutch people have of their own happiness and satisfaction with life are not disturbed by a difference in appreciation of the scales.

***Table 7 Average and standard error assessment of upper boundaries by Dutch judges***

| Response option | Pair 1 | | | | Pair 2 | | | |
|---|---|---|---|---|---|---|---|---|
| | Happiness | | Satisfaction | | Happiness | | Satisfaction | |
| | Average | Std. err | Average | Std. err | Average | Std. err | Average | Std. err |
| 5 | | | | | 10.0 | - | 10.0 | - |
| 4 | 10.0 | - | 10.0 | - | 9.0 | 0.04 | 8.8 | 0.04 |
| 3 | 7.9 | 0.05 | 7.9 | 0.04 | 7.5 | 0.05 | 7.2 | 0.06 |
| 2 | 5.3 | 0.06 | 5.3 | 0.06 | 5.7 | 0.07 | 5.3 | 0.07 |
| 1 | 3.0 | 0.08 | 3.0 | 0.08 | 4.0 | 0.09 | 3.6 | 0.09 |

| Response option | Pair 3 | | | |
|---|---|---|---|---|
| | Happiness | | Satisfaction | |
| | Average | Std. err | Average | Std. err |
| 5 | 10.0 | - | 10.0 | - |
| 4 | 8.4 | 0.04 | 8.4 | 0.04 |
| 3 | 5.9 | 0.05 | 6.0 | 0.05 |
| 2 | 4.1 | 0.06 | 4.2 | 0.06 |
| 1 | 1.8 | 0.05 | 1.9 | 0.06 |

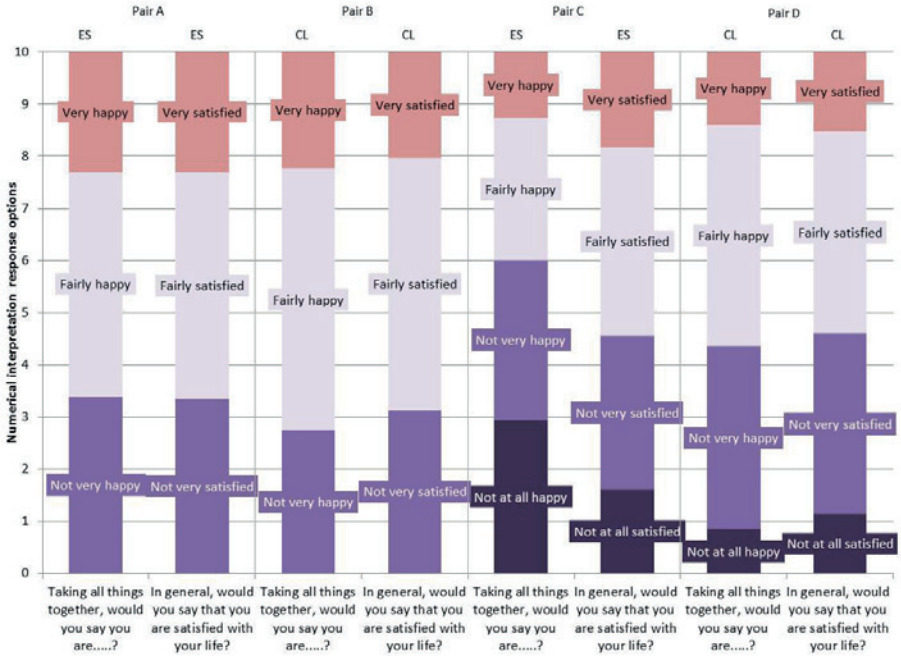
### 4.4 Comparison of results for options labelled in Dutch versus options labelled in Spanish

Although Dutch judges appreciate the scales for happiness and satisfaction with life synonymously, this may be different for native speakers of other languages. To shed some light on this we looked at two pairs of items from studies in Spanish which were assessed by Spanish-speaking judges from Spain and Chile. The result of their assessment is presented in Fig. 7. Pair D in Fig. 7 is equivalent to pair 1 presented in Fig. 6. The difference between these two pairs is in the item for life satisfaction. The question in the Spanish version of the life satisfaction item in pair 1 in Fig. 6 is “En general, ¿qué tan satisfecho está con el tipo de vida que lleva?” which can be considered as equivalent to the life satisfaction question of the item in pair B in Fig. 7 which is formulated as “En términos generales, ¿diría Ud. que está satisfecho/a con su vida?”. The problem however, is entailed in the labelling of the response options. Instead of ‘Bastante satisfecho/a’, as is used for one of the response options of the item in Fig. 7, the label for the item in Fig. 6 used in the Spanish studies is ‘Más bien satisfecho’. The difference between ‘bastante’ and ‘más bien’ is too large to consider these labels as equivalent, also because these labels are attached to response

options which are not at the end of the scale. Therefore we had to turn to the item for life satisfaction presented in Fig. 7. One could argue that 'nada' and 'para nada' are also not fully equivalent, but since these words in the labels of the response options are at the lower end of the scale, this is in view of the small observed frequencies of minor importance.

Unlike in the assessments by Dutch judges, noticeable differences can be observed in the interpretations of the Spanish-speaking judges of equivalent response scales for happiness and satisfaction with life. The judges from Spain interpret the response options of the 3-point scale items identically, which cannot be said for the assessments by the judges from Chile. Differences in the degree of appreciation of response labels for happiness and life satisfaction are, however, apparent in the assessments of the 4-point scale by judges from Spain and Chile. Moreover, differences are also observed among the boundaries between response options and the intervals produced by the two Spanish-speaking groups. This is likely to be due, at least to some extent, to cultural differences and the fact that Spanish and the Chilean Spanish-based languages have drifted apart when speaking colloquially.

***Figure 7 The assessment of similar scales by Spanish and Chilean judges***



Given the results shown in Fig. 7, the impression rises that Spanish people rate themselves easier as being satisfied with life than as being happy, in the sense that for pair C the bar for being happy is set higher than for being satisfied. These results differ from the assessments by the Chilean judges which give a more or less opposite impression. Markedly, the difference in degree of appreciation between response options for happiness and life satisfaction by Chilean judges are not visible in the upper part of the scale. Furthermore it is worth noting that in contrast to the assessment of the judges from Spain, the judges from Chile assign a larger degree of appreciation to 'Not at all satisfied' than to 'Not at all happy'. The assessments by the Chilean judges point in the same direction for both pairs of items. In both pairs, the degree of appreciation assigned to the response options at the lower end of the scale is higher for life satisfaction than for happiness.

In both Spain and Chile, the language used in the studies was Spanish. From Fig. 7 it can be noticed that despite the 'equality' in languages, the differences in judgement are remarkable, which to our opinion is not just due to the unrepresentativeness of the judges or the fact that the Spanish spoken in Spain is different from the Spanish spoken in Chile. It is more likely that the differences have to be attributed to the differing cultural environments in Spain and in Chile.

From a comparison between languages it follows that while Dutch judges assign intervals of similar length to equivalent response options, this similarity does not fully return in the assessments of Spanish-speaking judges. Concerning the results for Spain, the standard errors of the estimated mean values of the assessments of the upper boundaries of the response options are very small, just like the standard errors reported in Tab. 8 for The Netherlands. From that we conclude that, unlike the Dutch judges, the Spanish-speaking judges from Spain, though not in general, do on average appreciate the scales for happiness and satisfaction with life differently. We conclude the latter also for Chile, but with some restraint, given the large standard errors resulting from the assessments by the Spanish-speaking judges from that part of the world, which are highly related to the low number of judges, especially for the scales for happiness which were assessed by slightly less than 25 judges.

In sum, Dutch judges appreciate the response scales for equivalent items on happiness and life satisfaction equally, but these findings are not mirrored in the results of the assessments by Spanish-speaking judges.

***Table 8 Average and standard error assessment of upper boundaries by judges from Spain and Chile***

| Response option | Pair A | | | | Pair B | | | |
|---|---|---|---|---|---|---|---|---|
| | *Spain* | | | | *Chile* | | | |
| | Happiness | | Satisfaction | | Happiness | | Satisfaction | |
| | Average | Std. err | Average | Std. err | Average | Std. err | Average | Std. err |
| 3 | 10.0 | - | 10.0 | - | 10.0 | - | 10.0 | - |
| 2 | 7.7 | 0.08 | 7.7 | 0.07 | 7.8 | 0.38 | 8.0 | 0.23 |
| 1 | 3.4 | 0.10 | 3.3 | 0.10 | 2.7 | 0.37 | 3.1 | 0.35 |

| Response option | Pair C | | | | Pair D | | | |
|---|---|---|---|---|---|---|---|---|
| | *Spain* | | | | *Chile* | | | |
| | Happiness | | Satisfaction | | Happiness | | Satisfaction | |
| | Average | Std. err | Average | Std. err | Average | Std. err | Average | Std. err |
| 4 | 10.0 | - | 10.0 | - | 10.0 | - | 10.0 | - |
| 3 | 8.7 | 0.06 | 8.2 | 0.07 | 8.6 | 0.27 | 8.5 | 0.15 |
| 2 | 6.0 | 0.08 | 4.6 | 0.09 | 4.4 | 0.31 | 4.6 | 0.17 |
| 1 | 2.9 | 0.10 | 1.6 | 0.08 | 0.8 | 0.26 | 1.1 | 0.15 |

## 4.5 Discussion

In this chapter we have demonstrated how the Scale Interval Recorder can be used to explore whether equivalent response options of equivalent scales that only differ in the keywords used denote an equal degree of appreciation or not. We gave an idea of how this exploration can be done by illustrating the method for the topics happiness and satisfaction with life. In this section we will discuss some points of interest in relation to the topic of this chapter.

### *4.5.1 Methodological consideration*

Current and past practice show that in many studies in which topics such as happiness and satisfaction with life are related to other aspects of life, the scales of the items that are associated are dissimilar. In order to be convinced that results will not be muddled due to incomparability of the scales or differences in the interpretation of each scale by the respondents, in future research the scales of items that will be compared, have to be examined and discussed carefully beforehand. The Scale Interval Recorder is a useful instrument to do this since it offers the opportunity to have the response scales that are used in a study be assessed by judges and to compare the outcomes.

### *4.5.2 Limitations*

The results presented in this thesis for the interpretation of survey items by native Dutch- or Spanish-speaking judges are based on only a limited number of pairs of items on happiness and satisfaction with life and thus do not represent the large variation in survey items in full. The conclusions drawn are only valid for the type of items that have been mutually compared. To generalize these findings, further research is needed that involves investigating the equivalence of scales for an extended set of pairs of survey items that better represents the large variation in survey items.

We acknowledge that the group judges we used to illustrate the method may not be fully representative for this latter purpose, which especially holds for the results obtained for Chile. The judgement of the items on happiness was done by slightly less than 25 judges from Chile and it is obvious that a low number of judges may lead to an imprecision of the results if these different sample sizes are not taken into account in the appropriate way.

### *4.5.3 Implication of the method*

When searching for pairs of items with equivalent response options for happiness and life satisfaction we found that the occurrence of such equivalency in one language does not mean that this equivalency also exists in the corresponding pair labelled in another language. This is a by-product of our research, one from which we conclude that the translation of the labels of response options may come with, what we call, an interpreters' bias that reduces the comparability of survey responses between language groups.

### *4.5.4 Advice for further research*

If only one pair of items has to be assessed, it is likely that the assessment of the scale of the second item is influenced by the assessment of the scale of the first item, if they are presented successively to the judges. An option to prevent that form of bias from happening is to split the group of judges in two equally representative groups and to have each group assess one of the items or both items but present them in a different order. Another option is to include the pair of items in a series of items that are part of a Scale Interval Study as was the case for the pairs of items we present in this thesis.

If more pairs of items have to be assessed, it seems to make sense for the same reason as mentioned in the previous point, to split the group of judges in two and to have each group assess one item of each pair. If the

group of judges is split to have each group assess one item of each pair, another choice is to have each group assess only the items on the same topic or items for both topics.

### 4.6 Conclusion

Equivalent response scales do not always elicit well comparable responses across topics and languages, for example ‘Very happy’ and ‘Very satisfied’ denote the same degree in Dutch language, but not in Spanish. Comparative analysis requires therefore that comparability of response scales is assessed in advance. The Scale Interval Recorder can be used for that purpose.

# 5 'Very Happy' is not always equally happy

## 5.1 The same keyword in different contexts

In survey research response options are often labelled verbally, in the case of happiness research with terms such as 'Very happy' or 'Fairly happy'. Such response scales differ in the number and wording of response options. The following questions are an example of different questions on the same topic.

A question on happiness in the periodical Dutch Household Survey reads:

*To what extent do you consider yourself a happy person?*
- *Very happy*
- *Happy*
- *Neither happy nor unhappy*
- *Unhappy*
- *Very unhappy*

The same topic is also measured using questions with a different set of response options, such as this question in the International Social Survey Program:

*If you were to consider your life in general, how happy or unhappy would you say you are, on the whole?*
- *Completely happy*
- *Very happy*
- *Fairly happy*
- *Neither happy nor unhappy*
- *Fairly unhappy*
- *Very unhappy*
- *Completely unhappy*

Both questions offer the option 'Very happy', but do these options denote the same degree of happiness? Probably not. The difference between 'Very happy' and the next option is likely to be larger in the first case, 'happy', than in the second, 'Fairly happy'. Likewise 'Very happy' is likely to denote a higher degree of happiness in the first case, where it is presented as the highest option, than in the second case where 'Very happy' comes after 'Completely happy'.

### *5.1.1 Research question*

The difference in the phrasing of questions is no problem when surveys are analysed separately, but, as we have stated in Ch. 1, it limits the comparability of findings gathered in different surveys that used different questions. This begs the question of how serious the comparability problem really is. If differences are marginal, we can continue in the second way. If not, we must either abstain from comparison at all or develop better methods for scale transformation. Since we cannot address all comparability issues, we limit ourselves to the comparability of responses to identical response options that figure in non-identical scales, such as the option 'Very happy' in the two questions above. Does 'Very happy' mean just as much happiness in these cases? More formally formulated our research question reads: To what extent does the meaning attached to identical response options differ when used in the context of non-identical response scales?

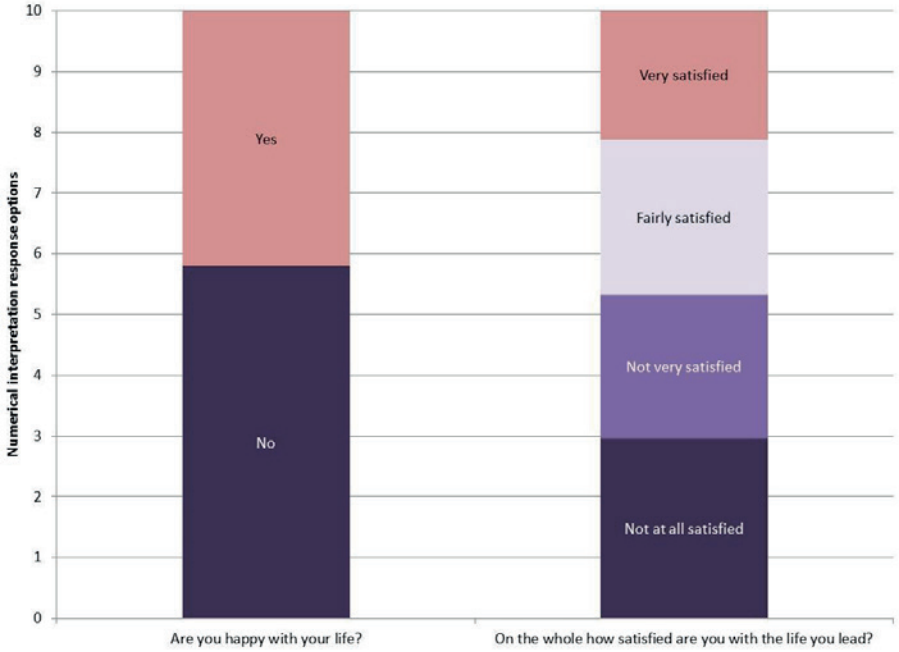
### 5.1.2 The keywords 'happiness' and 'satisfaction with life' and the degree of appreciation

One of the findings in Ch. 4 is that Dutch judges assign an equal degree of appreciation to equivalent response options for happiness and life satisfaction This finding does not imply that happiness and satisfaction with life are the same construct in The Netherlands. Despite the equivalence in degrees of appreciation assigned to response options, Dutch respondents may rate their feeling of 'happiness' somewhat differently from how they appreciate their 'satisfaction with life'. The equivalence in degrees of appreciation merely means that a comparison of ratings for happiness and satisfaction with life is not disturbed by a difference in the degrees of appreciation assigned to the response options in equivalent scales. Given the results for the Dutch-speaking judges we described in Ch. 4, we feel it is justified not to distinguish between questions that use the term 'happiness' or those that use 'satisfaction with life' for the results presented in this chapter.

## 5.2 The meaning of 'happy' and 'satisfied' in the context of the response scale

We started with the question "Are you happy with your life?", with response options 'Yes' and 'No'. The interpretation of these options and that of an item with four response options are presented in Fig. 8.

**Figure 8 Numerical interpretation of verbal scales on happiness and life satisfaction**

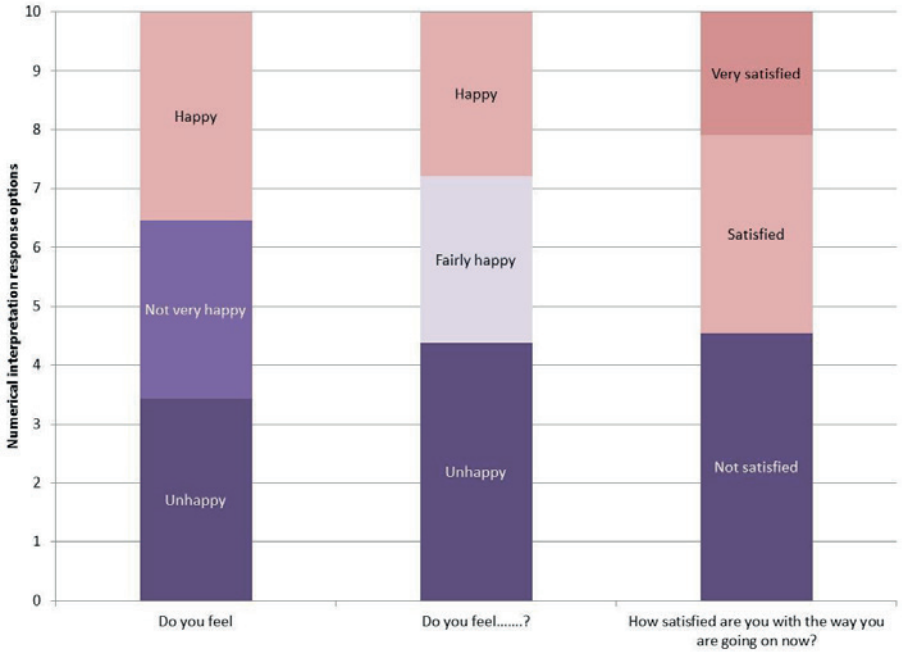



If only the options 'Yes' and 'No' are offered, the continuum is not partitioned into two intervals of equal length, the split is made at the value 5.8. This result for The Netherlands may partly be influenced by the fact that in the Dutch school system tests are graded on a numerical scale from 1 to 10, where a grade below 5.5 means that you have failed the test.

The 4-point scale item in Fig. 8 is illustrative for the difficulty of comparing survey results for different items. The words 'Yes' and 'No' in the 2-point scale item can be replaced by 'Satisfied' and 'Not satisfied', however, these qualifications of the degree of happiness do not return as such in the 4-point scale. In this latter scale the wording chosen to express the degree of satisfaction is stronger or weaker than 'Satisfied' and 'Not satisfied'. The most notable of the results for the 4-point scale is that, although the wording chosen for the most negative option cannot be superseded by something that is more extreme, the average length of the interval it has been assigned by the judges is rather large. The reason could be that the preceding response option is formulated rather moderately.

More insight into the effect the wording used for labelling response options has on how a scale is interpreted can be gained by comparing the results for the three differently labelled 3-point scales shown in Fig. 9.

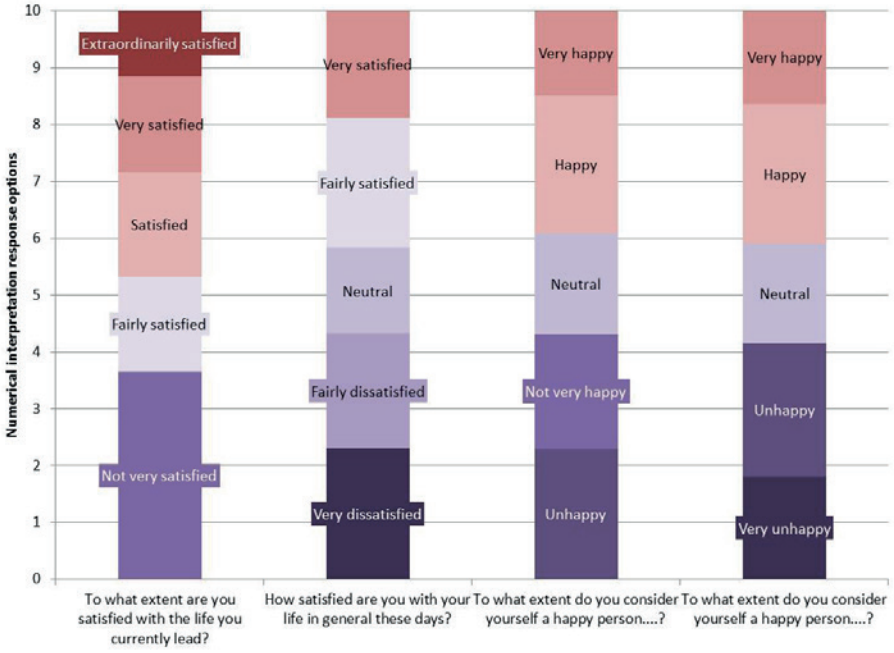
**Figure 9 Numerical interpretation of verbal scales on happiness and life satisfaction (continued 1)**




For the first two items in Fig. 9, which have equally labelled anchor points, the wording chosen for the label of the middle option can be seen to be crucial for the interpretation of the scale. A negative formulation comes at the cost of the interval for the lower anchor point and a positive formulation at the cost of interval for the upper anchor point. A more extremely labelled anchor point, as in the third item, turns out to reduce the valuation of the word 'Satisfied' to a much lower position on the continuum. In this scale 'Satisfied' no longer acts as the upper part of the continuum, the lower bound of this option has dropped below 5. The percentages of happy people measured with a response option labelled 'Happy' are clearly shown in Fig. 9, but it is obvious that the different scales cannot be compared in a straightforward manner. The meaning of the label 'Happy' shows a large contextual dependency on the composition of the response scale.

This becomes even more obvious from Fig. 10 in which two items with asymmetric response scales, one on satisfaction with life and one on happiness, are presented next to, more or less, symmetric variants. Comparing the two items on satisfaction with life, it is striking to see that the response option 'Fairly satisfied' in the asymmetric scale is positioned in the lower part of the continuum, whereas in the symmetric scale this is a degree of satisfaction with life on the other, non-overlapping part of the spectrum. In addition it is remarkable that, although the asymmetric scale consists of five response options, the option 'Not very satisfied' covers more than one third of the continuum, probably due to the absence of a truly negatively labelled option.

**Figure 10 Numerical interpretation of verbal scales on happiness and life satisfaction (continued 2)**




Looking at the items on happiness in Fig. 10, it can be noted that the upper part of the scales is equal for both items. The interpretation of this scale shows only minor differences between the two items. The distortion comes in the lower part of the scale for the third item which is not symmetric compared to the upper part. The word 'Unhappy' in the asymmetric scale seems to denote something different than the same word in the symmetric

scale. Furthermore it is noteworthy, from Fig. 10 that the last three items all have a neutral response option in the middle. This neutral option is interpreted as positioned in the middle of the continuum, where the length of the interval seems to depend on the wording of the surrounding response options. A weaker labelling of the directly neighbouring response options seems to trim down the length of the interval of the neutral option.

### 5.3 The effect of the wording chosen for the anchor points of the response scale

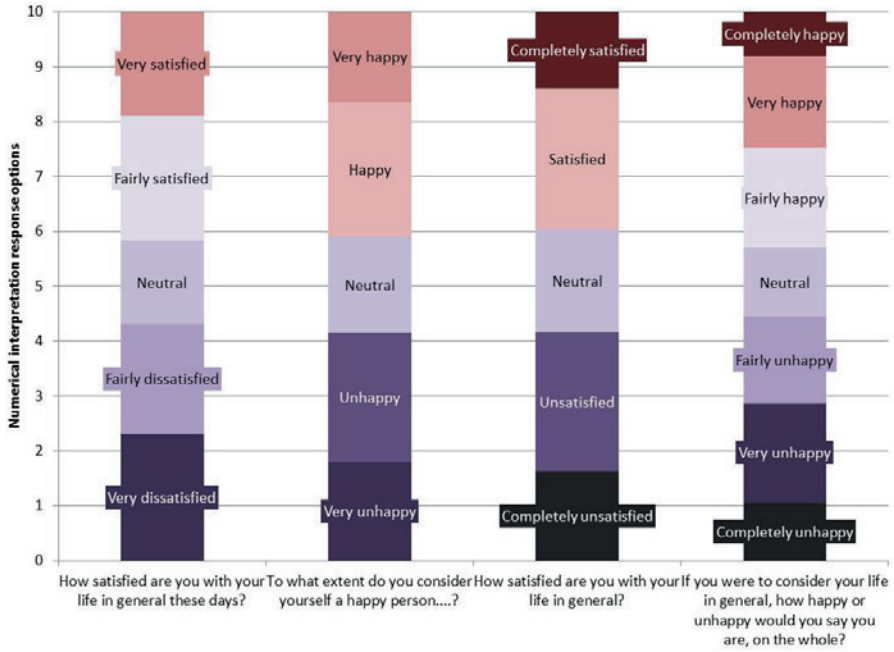
The wordings chosen for the anchor points of the response scales are especially interesting. The extremes of the continuous bar scale on which judges have to rate the verbal response options are labelled 'Worst possible' and 'Best possible'. This scale is all inclusive offering room to every degree of satisfaction with life or happiness one can think of, it can even be combined with both questions on happiness and satisfaction with life, since the topic is not explicitly part of the wording of the extremes. These characteristics are not obvious for a verbal response scale, as is demonstrated by the examples given above. Most often the anchor points are expressed in terms of the topic they relate to and, although they may differ in the intensity they express, they do not always make a scale all inclusive.

The labelling of the anchor points of a response scale is depicted in Fig. 11. The two items with a symmetric scale depicted in Fig. 10, return in Fig. 11 together with two items that also have a symmetric scale, but where the word 'Completely' is used instead of 'Very' for the labelling of the anchor points. Comparison of the first two items shows that the intervals for the response options labelled with a 'Completely' are smaller than when they are labelled with a 'Very'. This can be attributed to the fact that more than completely is not possible, whereas one can be 'very satisfied' but still not be 'completely satisfied'. The word 'Completely' as used for the items in Fig. 10 make the scales that they belong to all inclusive with respect to the topic they refer to.

Comparing the anchor points of the third and fourth item in Fig. 11, it is clear that the word 'Completely' does not replace the word 'Very' in the scale of the fourth item, but it has been introduced to label an extra response option on both sides. As a result this last item offers respondents the largest choice of options. The intervals assigned to these options however, seem to be somewhat compressed compared to the 5-point scales. The space reserved for the anchor points of this 7-point scale is rather limited compared to the other response options. One could question

whether a 7-point verbal scale has an added value over a 5-point verbal scale.

**Figure 11 The effect of labelling the anchor points on the interpretation of response options**

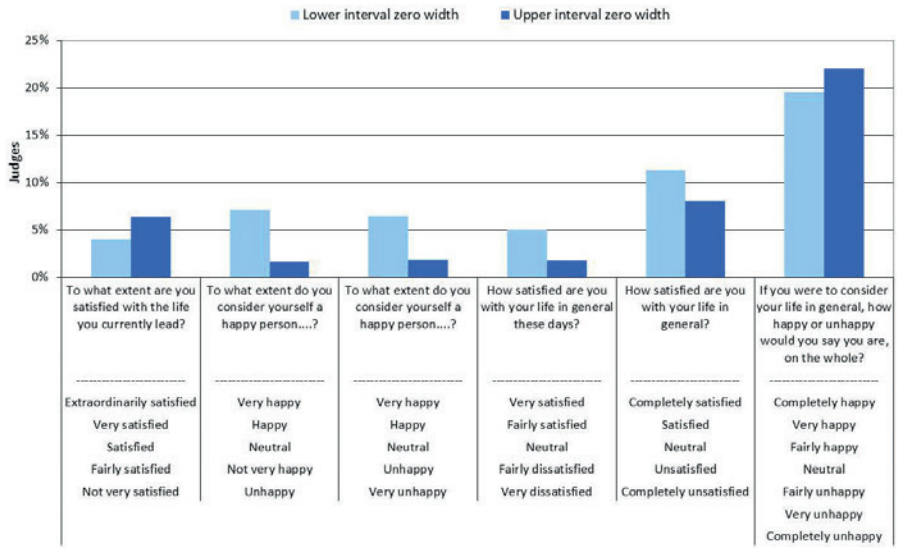



Response options formulated with wordings such as ‘Completely’ or ‘Not at all’ may tempt judges to assign a zero-width interval to them by choosing the upper and lower bound both equal to one of the extremes of the continuum (Kalmijn, 2010, p. 147 sqq). These labels all express some kind of limit that cannot be exceeded however, for the selected items with less than five response options, these zero-width intervals occur very rarely. For the items with at least five response options the percentage of judges that assigned a zero-width to the anchor points is displayed in Fig. 12.

Although the formulation ‘Extraordinarily’ is rather extreme, strikingly it does not lure judges to assign a zero-width interval to one of the anchor points of the verbal scale. Yet the percentage of zero-width intervals for the first item in Fig. 12 differs from that for the other 5-point items in that the largest percentage belongs to the upper anchor point of the scale. Maybe the percentage of zero-width judgements for the lower response option of the second item is more important. The percentage is not very

high, but the response option it applies to is not extreme. The zero-width interval for this item might be attributed to the fact that the range from 'Very happy' to 'Not very happy'[12] can be considered to be a complete, though not all inclusive scale making the response option labelled 'Unhappy' an outside class. With respect to this, the composition of the scale belonging to the third item, starting with 'Very unhappy' and ending with 'Very happy' seems to be more logical. Most notable however, are the percentages for the two items with the words 'Completely' in the labelling of their response options. There are no superlatives for the response options with these labels. For the 7-point item the zero-width percentages make it plausible that the formulation of the anchor points for this scale have no added value. From this it may be concluded either that a 5-point scale would do, what would amount to a scale as for the fourth item, or that a different and a less extreme wording should be chosen for the response options labelled 'Very unhappy' and 'Very happy' in the 7-point scale under the condition that a verbal response scale would be preferred over a numerical response scale.

**Figure 12 Percentage of anchor points to which a zero-width interval has been assigned**




[12] It must be noted here, that 'Not so happy' is somewhat closer to the original Dutch phrasing for this response option. The word 'Very' instead of 'So' is chosen however, for the translation for reasons of comparison with other scales, but this does not change the conclusion.

## 5.4 The effect of the number and wording of response options on the central tendency

Although the outcomes of the study as presented in the previous sections are probably not very surprising, and as one could have expected, it is interesting to see what the effect of these differences in the number and wording of response options has on the central tendency of the scores on happiness and satisfaction with life. A glimpse on this can be found in the scores on satisfaction with life and happiness taken from the Permanent Survey on Living Conditions of Statistics Netherlands[13]. In this survey the scores are based on the first and third items presented in Fig. 10.
The response scale for the item on satisfaction with life is an exceptional case, because it is asymmetric with only one negative response option and options formulated in extreme terms at the positive end of the response scale. Due to this, the interpretation of the response options 'Satisfied' and 'Fairly satisfied' seems to be forced to end in a too low part of the numerical scale. The response option 'Fairly satisfied' is judged to be less positive than the neutral options of the other items shown in Fig. 10. The scale of the item on happiness is more or less symmetric, with a neutral option in the middle, but with the 'extreme' for the lower end labelled 'Unhappy' making it an outside class as discussed in Sec. 5.2.

Over the years the scores for these items in the Netherlands have been stable but they show a remarkable difference in the central tendency of the response on happiness and satisfaction with life. Two thirds of the population claims it is 'Happy' whereas less than half of the population believes it is 'Satisfied'. Conversely, over forty per cent of the Dutch seems to be 'Very satisfied' or even 'Extraordinarily satisfied', which is about twice as much as the share of people indicating they are 'Very happy'. It is very likely that the answer given by a respondent is influenced by the position of the response option on the scale. Someone who is satisfied with his of her life would consider a response option, even if it is labelled 'Satisfied', in the middle of the scale not in concordance with this, this could tempt these persons to choose the option 'Very satisfied', since this is the position on the scale that is more in harmony with their perception of satisfaction with life. In practice this difference in outcomes for happiness and satisfaction with life is often overcome by dichotomizing them and for this purpose the population that scored 'Very happy' or 'Happy' on the question about happiness is declared to be happy. For satisfaction with life the scores on

[13] http://statline.cbs.nl/StatWeb/publication/?DM=SLEN&PA=60027ENG&D1=43-52&D2=0&D3=a&LA=EN&HDR=T&STB=G1,G2&VW=T

the response options ‘Extraordinarily satisfied’, ‘Very satisfied’ and ‘Satisfied’ are combined in a category of people that are assumed to be satisfied with their life. An example of this is given in DeJonge, Hupkens, and Bruggink (2009). As a result of the dichotomization, the share of happy people nearly equals the share of satisfied people. Although this brings us to the observed difference, one could question whether dichotomization is really justified and if doing so causes us to lose a lot of information.

## 5.5 Discussion

A comparison of survey items with as few as 3-point on the scales is already sufficient to see that the number and wording of the response options does matter. If the anchor points of a 3-point scale are labelled in a purely oppositional manner, than it very much depends on the wording chosen to label the response option in the middle where the boundaries between this and the two anchor points are positioned. A negative formulation comes at the cost of the interval for the lower anchor point and a positive formulation will trim the length of the interval for the upper anchor point.

In a symmetric scale, independent of the number of response options, a neutrally labelled middle option will be positioned in the centre of the scale. The length of the interval assigned to it will depend on the intensity expressed by the wording used to label the surrounding response options. If the wording is moderate, like in ‘Fairly happy’, the length of the interval will be smaller, than when a more pronounced formulation is applied such as ‘Happy’ or ‘Very happy’.

Scales that have been deliberately devised to be asymmetric to bring variation in the responses turn out to have an unintended side-effect. These scales may tempt respondents to choose a response option that is positioned on the scale in accordance with their perception even if this is not fully in conformity with the label attached to it. As a result the central tendency of the measurements reflects the asymmetry of the scale instead of the subjective well-being of the respondents.

Sometimes the wording chosen for the labels of the anchor points on a scale invite judges to assign a zero-width interval to them, by choosing the upper and lower bound equal to one of the extremes of the continuum. This holds especially for wording such as ‘completely’ for which there are no superlatives. If the adjacent response options also have extremely formulated labels, judges are encouraged even more to assign a zero-width interval to the anchor points of the scale, which would reduce their added value. From this it might be concluded that, for verbal scales, anchor points labelled using the word ‘completely’ are more or less redundant if the

adjacent options are also expressed in extreme terms, however, to make a numerical scale all inclusive it is preferable to label the anchor points using 'completely' instead of 'very' and even better to use the labels 'Worst possible' and 'Best possible'.

In the Happiness Scale Interval Study it is assumed that there is no conflict between the personal perception of happiness of a judge and his or her assessment of the response scales (Kalmijn 2010, p. 179). According to Kalmijn, the only justification for retaining this assumption is that it has never been investigated. In addition to this, it is worth noting that, even if the assumption is true, a verbal response scale does not necessarily offer response options that meet the perception of respondents well, they may force them to choose between two less than optimal alternatives. The least inappropriate option may be ranked in a counterintuitive position by a respondent in between the other response options. As a consequence, the boundaries derived from the assessments by judges may not correspond to how the response options are selected in practice by respondents. To illustrate this, take the item with four response options that we presented in Fig. 8, that offers the response options 'Fairly satisfied' and 'Very satisfied'. Respondents who are satisfied with their life have to choose between an option that either underestimates or overestimates their perception of satisfaction with life.

The study presented in this thesis does not lead to an answer as to which scale is the best to use. This however, is not what the study was aimed at. For example from the study by Schwarz et al (1991) it is already obvious that the choice of numbering of closed-ended numerical scales and the labelling of their anchor points affects the distribution to survey answers. What the Scale Interval Study contributes in addition to Schwartz's study, is that it focuses on verbal response scales and provides a systematic way to express the degree of appreciation denoted by each response options on a continuum from 0 to 10. What our study very clearly revealed is that the degree of happiness denoted by verbal response options, such as 'Happy' or 'Unhappy', is strongly affected by the number of options presented, the wording of these options and sometimes also the position of an option in the scale. Hence findings on the same topic obtained using different verbal response scales cannot be compared well.

## 5.6 Conclusion

The degree of happiness denoted by verbal response options, such as 'Happy' or 'Unhappy' is strongly affected by the construction of the scale, which is among others reflected in the number of options presented, the

wording of these options and the position of an option in the scale. Hence findings on the same topic obtained using different response scales cannot be compared. Conventional methods for scale transformation fail to overcome the differences in degree of happiness or satisfaction with life denoted by the different response options used in different questionnaires. More advanced scale transformation methods or other alternatives are needed before the findings of such studies can be used effectively for research synthesis.

# PART 3

# INNOVATION 2: THE CONTINUUM APPROACH

# 6 The Continuum Approach

## 6.1 Happiness: a discretely or continuously distributed variable?

We have shown that neither the two conventional scale transformation methods, Linear Stretch and the Semantic Judgement of Fixed Word Value Method, nor the Scale Interval Method offer a solution to the comparability problem by means of Tab. 5 in Sec. 2.3. This has to be attributed to the fact that if a discrete primary scale is transformed by one of these methods, the resulting secondary scale is still discrete.

The use of discrete scales in survey research is often practically motivated, for example in several modes of surveying it is easier to ask a respondent to make a choice from a limited number of options than to have them point out an exact individual value on a continuous scale that corresponds to their perception. Respondents are asked to answer a closed question with a limited number of response options which together make a survey item. The response scales, both verbal and numerical, vary in the number of response options available, some including only two options, for example 'Yes' or 'No', and others eleven, for example the integer numbers from 0 to 10. A more valid approach, as Kalmijn (2010, Ch. VI) argues, is to consider the existence of a latent continuous variable underlying the survey variable, the distribution of which is estimated using the survey item and the response to it.

The use of discrete scales explains the variety of response scales that has developed over time. This variety limits the comparability of answers to survey questions in general and using happiness as an exemplary topic. Kalmijn (2010, Ch. VI) developed the Continuum Approach to tackle this comparability problem in combination with the notion that happiness is to be treated as a continuous variable.

## 6.2 Outline of the Continuum Approach applied to happiness

The Continuum Approach postulates a latent happiness variable in the population, which is continuous over the interval [0, 10]. In the case of happiness, a beta distribution is the most appropriate to use in the Continuum Approach, due to at least three interesting properties it has (Kalmijn et al, 2011, pp. 509-510):

(i) it is a continuous distribution, which makes it suitable as a model for the continuous latent happiness variable in the population

(ii) the random variable has a two-sided bounded domain, which makes it suitable for happiness as it is measured using two-sided bounded primary scales

(iii) the distribution has two shape parameters, which makes beta distributions cover a wide class of different distribution shapes, including skew distributions, both positive and negative.

A fourth property which we mention here is that a beta distribution is independent of the arbitrary choice of an item by the institute that conducts the survey. We do not know any other distributions with these properties. More generally known alternatives as the normal distribution and the logistic distribution are less suitable than the beta distribution, among other things because their domain is infinite, they are bell-shaped and symmetric around their mean (Kalmijn, 2012), whereas happiness has clearly skew distributions (Lee et al, 1982; Cummins, 2003; Frijters et al, 2008; Guven et al, 2011).

The family of beta distributions consists of a series of distributions each member of which being characterized by two shape parameters, α and β.

A beta distribution can be expressed using the complete beta function:

(1) $$B(\alpha,\beta) := \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

where the parameters α and β are positive real numbers.

Given the formula (Eq. 1) the probability density function of the beta distribution on the continuum from 0 to 10 can be written as:

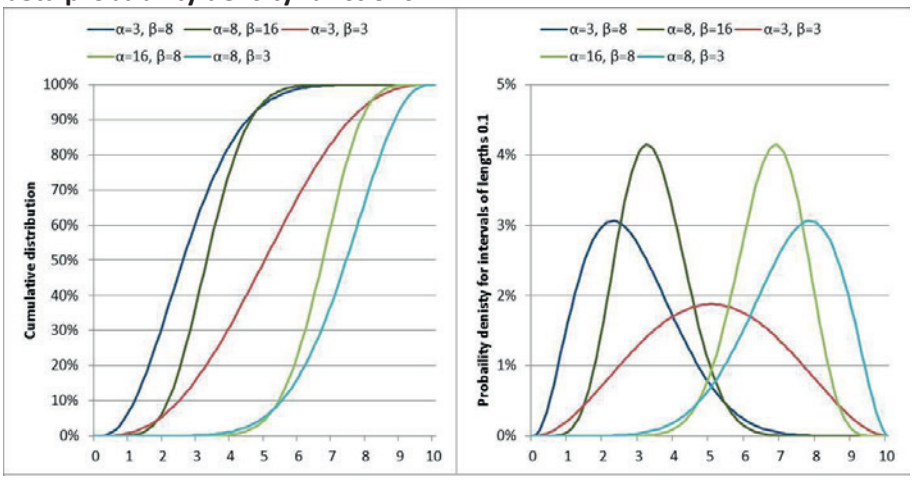
(2) $$f(x|\alpha,\beta) := \begin{cases} [10B(\alpha,\beta)]^{-1}x^{\alpha-1}(10-x)^{\beta-1} & \text{for } x \in [0,10] \\ 0 & \text{otherwise} \end{cases}$$

The mean $\mu$ of a beta distribution with parameters α and β on the continuum from 0 to 10 is equal to:

(3) $$\mu = 10\frac{\alpha}{\alpha+\beta}$$

To make this less abstract we give some examples of the probability density functions and the cumulative distribution functions for different values of $\alpha$ and β in Fig. 13.

**Fig. 13 Examples of the cumulative beta distribution functions and the beta probability density functions**




If $\alpha < \beta$, the probability density function is skewed to the right, if $\alpha > \beta$ the function is skewed to the left and if both parameters are equal the function is symmetric about x=5, the abscissa x being the happiness value on a 0 to 10 continuum, and the larger the values of $\alpha$ and $\beta$, the more peaked the density curve and the steeper the cumulative distribution curve.

A starting point for the Continuum Approach to happiness is provided by the cumulative frequencies of measured happiness on a discrete primary scale and the values on the continuum from 0 to 10 at which respondents change their judgement from one to the adjacent response option on this primary scale, for example from 'Happy' to 'Very happy'. On basis of the cumulative frequencies and the values on the continuum of the boundaries between the response options of the primary scale, the shape parameters $\alpha$ and $\beta$ of the best fitting beta distribution are estimated in the Continuum Approach as maximum likelihood estimators. This estimation procedure is described into more detail in Kalmijn (2010, p. 160 sqq). There is always a perfect fit in the case of a primary scale with three response options. If the number of response options is restricted to only two, then there is no single solution: the number of perfectly fitting beta distributions is infinite, and use of the Continuum Approach is therefore invalidated. In the case of at least four response options, then in general there will be no perfectly fitting beta distribution and the best fitting solution should be taken. Those who are interested in the methodological considerations of the Continuum Approach can find more information about it in Kalmijn (2010, Ch. VI) and Kalmijn et al (2011).

## 6.3 Combination of the Continuum Approach with the Scale Interval Method

The two verbal scale items taken from the Statistics Netherlands (CBS) and the Eurobarometer (EB) surveys which were shown in Tab. 4 of Sec. 2.2 are convenient to demonstrate the application of the Continuum Approach to the upper boundaries of the response options found by using the Scale Interval Method. The upper boundaries for these two items were presented in Fig. 4 and, as a reminder, are also given in Tab. 9.

**Table 9 Upper boundaries satisfaction with life scales based on the Scale Interval Method**

| *Item code Survey* | *Question* | *Response options* | *Upper boundaries* |
|---|---|---|---|
| O-SLL-c-sq-v-5-d CBS | To what extent are you satisfied with the life you currently lead? | - Extraordinarily satisfied<br>- Very satisfied<br>- Satisfied<br>- Fairly satisfied<br>- Not very satisfied | 10.0<br>8.8<br>7.2<br>5.3<br>3.6 |
| O-SLL-u-sq-v-4-b EB | On the whole how satisfied are you with the life you lead? | - Very satisfied<br>- Fairly satisfied<br>- Not very satisfied<br>- Not at all satisfied | 10.0<br>7.9<br>5.3<br>3.0 |

We used the upper boundaries given in Tab. 9 and the cumulative frequencies based on the frequency distribution in 2008 for each item as were given in Tab. 4 as input for the application of the Continuum Approach. The result for the CBS item is given in Fig. 14 and the result for the EB item in Fig. 15. The positions of the vertical lines in both figures on the continuum from 0 to 10 are equal to the upper boundaries in Tab. 9. The length of each vertical line is equal to the cumulative frequency in 2008 for the response option it belongs to. The continuous line in each figure is the beta distribution that fits best to these cumulative frequencies and the given upper boundaries.

**Figure 14 Application of the Continuum Approach to the CBS item**

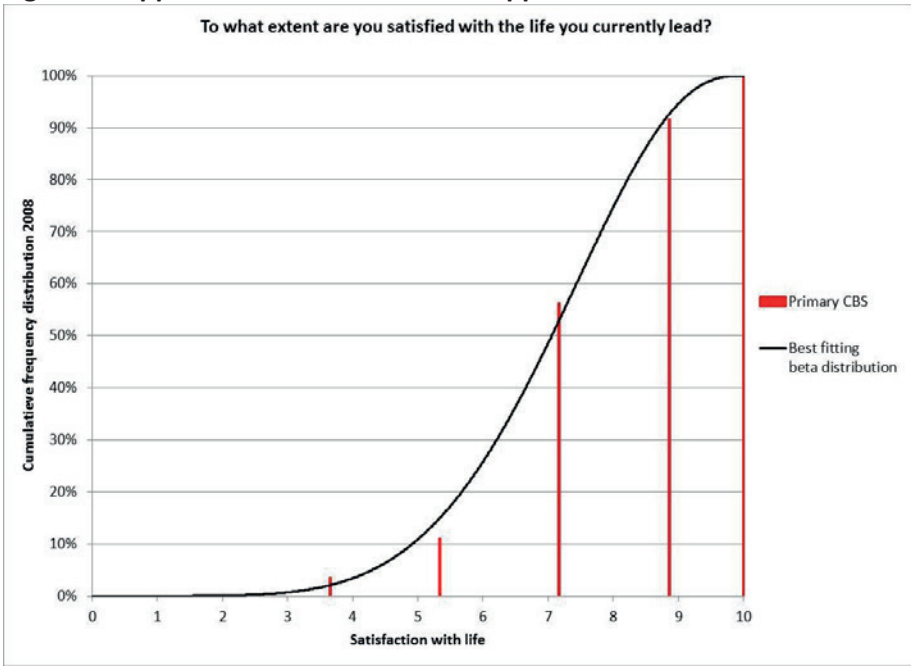
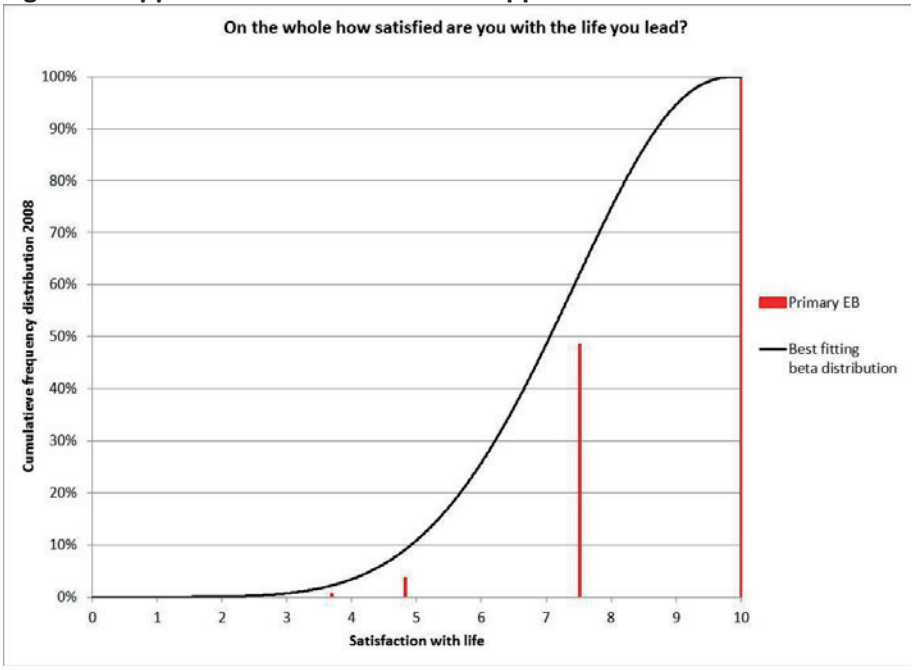



**Figure 15 Application of the Continuum Approach to the EB item**



It can be noticed from Fig. 14 and 15 that the fit of the best fitting beta distribution estimated for the CBS item is better than that of the best fitting beta distribution estimated for the EB item. The vertical distance of the best fitting beta distribution to the cumulative frequencies measured on the primary scales is much larger for the EB scale than for the CBS scale.

### 6.4 The Continuum Approach and discrete numerical scales

All accepted items of happiness are gathered in the collection 'Measures of Happiness' of the World Database of Happiness (Veenhoven, 2015b). About half of these are single questions which have to be rated on a numerical scale with ten or eleven response options. If the number of options of these numerical scales is less than ten, then in three out of four cases the number is equal to seven. There are also response scales which are not labelled with numbers or text, for example merely consisting of a series of boxes such as

□□□□□□□□□□

When applying the Continuum Approach these scales are treated as quasi numerical by assigning a rank to each option.

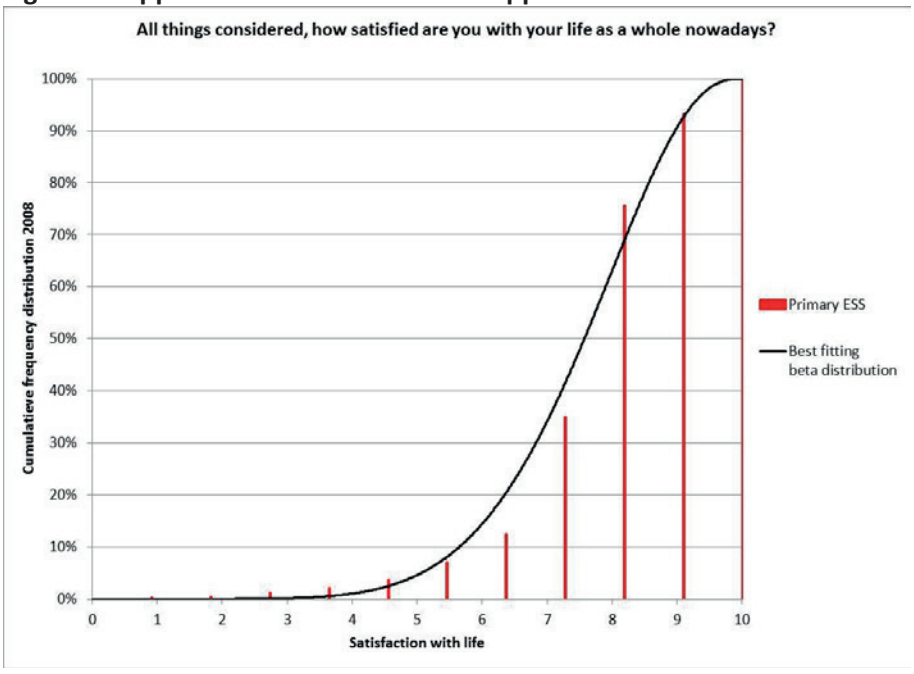
The use of numerical scales is in line with the assumption made by Voorpostel et al (2009) that attitudes fall along a single, latent continuum and that the larger the number of points on a response scale, the better it represents this underlying, latent continuum and the more accurately it reflects the variation. Voorpostel et al also state that 'the larger the number of points, the more powerful the scale is in discriminating, but at a certain point respondents become unable to make fine distinctions and thus round off'. In addition, Saris and Gallhofer (2007, pp. 118-119) mention that respondents who are asked to give an answer on a magnitude scale with fixed reference points, which is comparable to a bounded continuum, have a tendency to prefer numbers which can be divided by five, leading to peaked response distributions. They state that this does not happen if line production scales[14] are used, but that, due to practical considerations when using other modes of surveying, continuous scales may have a future once computer-assisted interviewing becomes more popular.

If the Continuum Approach is to be applied to survey items with numerical scales, a pragmatic choice to is to assume that the upper boundaries of the response options on the 0 to 10 continuum are equally distanced (Kalmijn, 2013).

[14] For line production scales respondents are asked to draw a line with a length that expresses the ratio of their judgement to the length of a standard line that is used as a reference judgement for a certain topic.

We applied the Continuum Approach to the life satisfaction item taken from the European Social Survey (ESS), which we introduced in Sec. 1.3, with the leading question: All things considered, how satisfied are you with your life as a whole nowadays? The answer has to be rated on an 11-point numerical scale from 0 to 10 with the anchor points labelled 'Extremely unsatisfied' and 'Extremely satisfied'. We fixed eleven equidistant upper boundaries, one for each response option, starting at 0.91 for the response option at the lower end of the scale and ending at 10.0 for the option at the upper end of the scale as is depicted in Fig. 16. We have drawn the cumulative frequency for each response option of the discrete primary ESS scale as a vertical bar at the position of the equidistant boundaries on the horizontal axis of Fig. 16. These cumulative frequencies are 0.3%, 0.5%, 1.1%, 2.0%, 3.7%, 7.1%, 12.4%, 34.9%, 75.5%, 93.2% and 100.0%. The curve in Fig. 16 is the beta distribution that according to the Continuum Approach fits best to the boundaries and cumulative frequencies distribution of the ESS item in 2008.

**Figure 16 Application of the Continuum Approach to the ESS item**

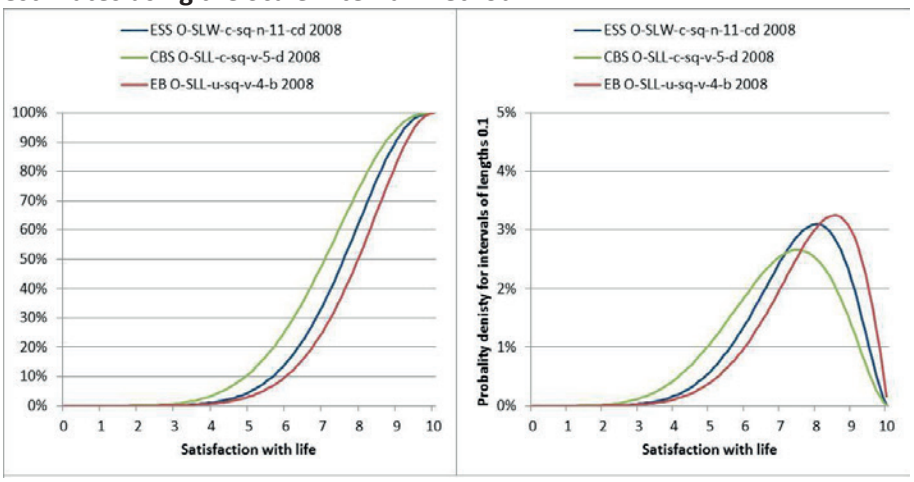



The parameters of the beta distribution in Fig. 16 are $\alpha = 7.92$ and $\beta = 2.76$, which, according to Eq. 3 corresponds to a mean of 7.4.

## 6.5 Comparison of the estimated means using different methods

We have combined the best fitting beta distributions for the items taken from the ESS, the CBS and the EB surveys in two graphs which are depicted in Fig. 17. The left graph shows the cumulative distribution function, the density function is shown on the right.

**Figure 17 Distribution of life satisfaction in the Netherlands in 2008: estimates using the Scale Interval Method**




As stressed in Sec. 2.3, since the results for all three items were based on survey responses made in 2008 to similar items, one would expect the three curves to more or less coincide. This is obviously not the case. Compared to the distribution for the ESS item, the distribution for the EB item is too skewed to the left and that for the CBS item too skewed to the right. For the EB item this can be explained by the fact that the primary scale offers the response options 'Fairly satisfied' and 'Very satisfied'. Respondents who are satisfied with their life thus have to choose between an option that either underestimates or overestimates their perception of satisfaction with life. Apparently a majority of the satisfied respondents tend to prefer the option 'Very satisfied' over the option 'Fairly satisfied', which pushes the beta distribution to the right. The explanation for the CBS item lies in the strong asymmetry of the primary scale in which four of the five options are formulated positively. As a consequence the option 'Satisfied' in the primary verbal scale is positioned in the middle of the scale, which may not be in accordance with the position a satisfied respondent would expect its position to be. Furthermore, as can be seen from Fig. 4 of Sec. 2.2 , the judges valued the position of the option

'Satisfied' in this context rather low on the continuum. As a consequence, the beta distribution for the EB item falls to the left of the distribution for the ESS item.

The estimated population means according to Continuum Approach and the different methods described previously are presented in Tab. 10. The Semantic Judgement of Fixed Word Value Method does not allow the calculation of a mean for the numerical ESS item, since the response options of this item do not contain any words of which the values can be fixed. Based on the discussion of the construction of the primary scales of the CBS item, however, we can conclude that a mean after scale transformation of 8.6 is far too high to be realistic. We would not expect the mean to be substantially higher than the mean estimated on basis of the Scale Interval Method for the EB item.

**Table 10 Means in 2008 according to different estimation methods**

| *Item code Survey* | *Linear Stretch* | *Fixed Word Value* | *Scale Interval Method (Weighted Average Approach)* | *Scale Interval Method (Continuum Approach)* |
|---|---|---|---|---|
| O-SLW-c-sq-n-11-cd ESS | 7.7 | - | 7.5 | 7.4 |
| O-SLL-c-sq-v-5-d CBS | 5.9 | 8.6 | 6.9 | 6.9 |
| O-SLL-u-sq-v-4-b EB | 8.2 | 7.8 | 7.7 | 7.7 |

Of all methods the means obtained using the joined Scale Interval Method and the Continuum Approach come closest to the estimated mean for the ESS item, yet they still leave a large gap in between the estimated means of this ESS item and are far from identical. We have also noticed these differences in outcomes for other survey items, although these showed smaller deviations of the means after transformation to that of the reference item than is the case for the items taken from the Eurobarometer and CBS surveys. In the remainder of this thesis when we talk about the Scale Interval Method we imply it is combined with the Continuum Approach.

Since the results for the items taken from the Eurobarometer and CBS surveys were the worst compared to other items we looked at, these two items were chosen as illustrative examples to show that an additional step has to be added to the Scale Interval Method to solve the comparability problem. Nevertheless we could conclude that the Scale Interval Method in general shows a smoother pattern of results than either the Linear Stretch Method or the Semantic Judgement of Fixed Word Value Method. The Scale Interval Method alleviates many of the shortcomings of the two older methods. Moreover in contrast to the older methods, the Scale Interval Method does do justice to the continuous nature of the latent variables that are assumed to underlie the survey questions being studied.

# PART 4

# INNOVATION 3: THE REFERENCE DISTRIBUTION METHOD

# 7 The Reference Distribution Method

## 7.1 Using a Reference Distribution to derive boundaries between response options

The observed differences for all methods in estimated distribution means between items as discussed in Ch. 6 were a trigger to devise a method in which a reference distribution is used to 'tune' responses to other questions on the same topic across surveys.

The Reference Distribution Method for making happiness data comparable builds heavily on the Scale Interval Method. Basically the two methods are identical except that in the Reference Distribution Method the boundaries between the response options of the primary scale are derived from a reference distribution instead of from assessments by judges by means of the Scale Interval Recorder.

With the Reference Distribution Method an attempt is made to deal with the fact that, for a given year and a given population, one would expect the estimated distribution means for similar questions about happiness asked in different representative surveys to be approximately the same irrespective of the primary response scales used: yet as we have shown in the preceding sections, this is not the case when using the methods described in those sections. We have explained that this is a by-product of the fact that the verbal scales used in for example the Eurobarometer and CBS items do not necessarily offer response options that meet the perception of respondents well, which forces them to choose between two less than optimal alternatives. The least inappropriate option may be ranked in a counterintuitive position in between the other response options. As a consequence, the boundaries derived from the assessments made by native language speaking judges may not correspond to how the response options are selected in practice by respondents.

To find a solution to this problem a different angle of approach is needed (Dijkgraaf, 2008). Instead of taking verbal scales that have to be transformed as the point of departure, the beta distribution that fits best to the survey results of a deliberately chosen item in a given year is used as the reference distribution to start the transformation of other scales. Preferably, this reference distribution is based on survey results measured on a continuum from 0 to 10. In general survey results measured on a continuous scale will not be available. As a second best solution a representative survey item with a numerical scale should be selected and used to estimate the best fitting beta distribution that can serve as the

reference distribution. If the Continuum Approach is used to derive a reference distribution based on survey results measured on a discrete scale, this scale should preferably be numerical with ten to eleven response options for the reasons we described in Sec. 6.4. If however, only verbal scales are available for a type of item that all consist of a similar question but vary in scale, one of these items has to be selected as a basis for the reference beta distribution. The Scale Interval Recorder can be deployed to obtain the values of the boundaries between the response options for this selected item. Combined with the frequency distribution for the selected item in a reference year the parameters of the best fitting beta distribution can be estimated and used as the reference distribution.

Once a reference distribution is available, its cumulative distribution function can be used to derive the boundaries between the response options on a continuum from 0 to 10 for any other survey item addressing a similar question, but with a different scale, that has been fielded in the same year as the reference distribution. These boundaries follow straightforwardly from the cumulative distribution of the reference distribution and the cumulative frequencies for the response options in the primary scale: the boundary between response option *i* and response option *i+1* is equal to the point on a continuum from 0 to 10 where the value of the cumulative distribution of the reference distribution is equal to the sum of the frequencies corresponding to the response options 1 up to and including *i* in the primary scale.

The boundaries between the response options are thus determined as the points where the cumulative frequency of the scale in the reference year matches with the reference distribution. We refer to the boundaries thus found as reference boundaries, since the reference distribution fits perfectly to these boundaries and the cumulative frequency distribution of each other scale in the reference year. The mean of the reference distribution is therefore an estimate of the mean on the 0-10 continuum for each of these other scales in the reference year.

## 7.2 Illustration of the application of the Reference Distribution Method

In Sec. 6.4 we have described how we applied the Continuum Approach to derive a best fitting beta distribution to the 11-point numerical scale and the cumulative frequency distribution in 2008 for the life satisfaction item taken from the European Social Survey (ESS). We will use this beta fitting beta distribution, which we presented in Fig. 16, as a reference for the illustration of the application of the Reference Distribution Method to the

life satisfaction item from CBS introduced in Ch. 1. We recall from Tab. 4 that the frequency distribution of the responses to this item in 2008 in The Netherlands is:

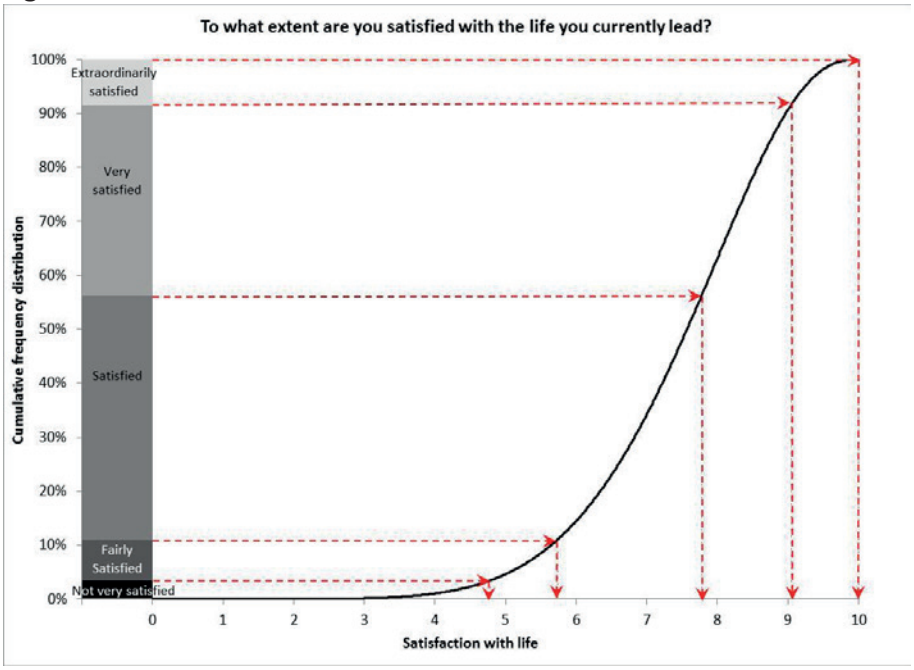
- Extraordinarily satisfied 8.4%
- Very satisfied 35.5%
- Satisfied 45.1%
- Fairly satisfied 7.6%
- Not very satisfied 3.4%

Using the Reference Distribution Method, the procedure to determine the reference boundaries between the response options of the CBS item on the continuum from 0 to 10 is as follows and as depicted in Fig. 18.

- We start with the cumulative frequency distribution of the CBS item for which we want to determine where the boundaries between the response options are positioned on the continuum from 0 to 10.This cumulative frequency distribution is shown as a stacked bar on the left side of Fig. 18.
- The reference distribution derived from the ESS is depicted to the right side of this stacked bar, plotted against the 0 to 10 continuum which is represented by the horizontal axis.
- A horizontal line is drawn from the cumulative frequency displayed in the stacked bar on the left side of Fig. 18 for each response option of the response scale, to the point where it touches the reference distribution. At this point the value of the reference distribution is equal to the cumulative distribution on the scale of the CBS item.
- From this point down, a vertical line is drawn to the 0 to 10 continuum on the horizontal axis. The value at which the vertical line touches the horizontal axis is the position of the reference boundary of the corresponding response option.

Following this procedure, the reference boundaries for the response options of the CBS item on life satisfaction on the 0 to 10 continuum are, consecutively, 4.8, 5.7, 7.8, 9.0 and 10.0. With the reference boundaries at these positions, the reference distribution perfectly fits to the cumulative frequency distribution of the CBS item in 2008. The mean 7.4 of the reference distribution is thus also an estimate of the population mean for 2008 wave of the CBS item.

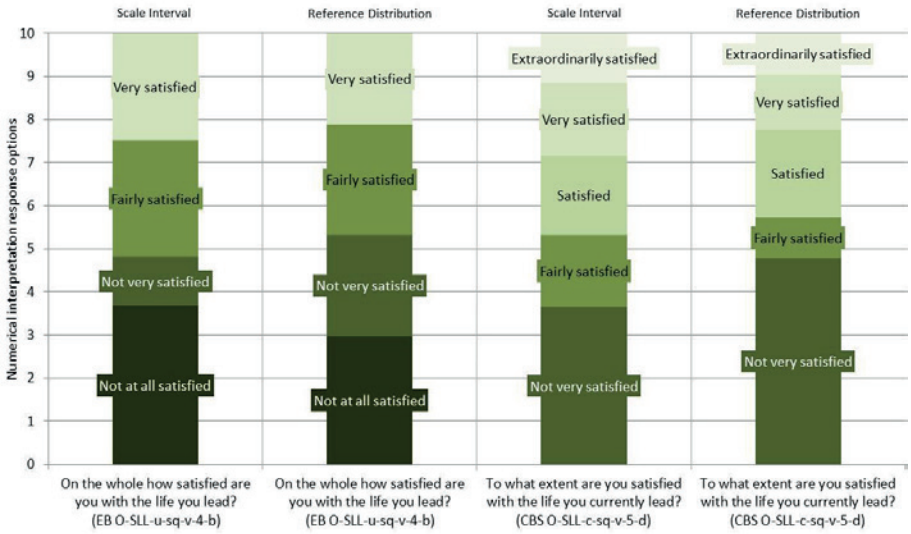
**Figure 18 Illustration of the Reference Distribution Method**




The reference boundaries found in this way can in their turn be used as input for the application of the Continuum Approach to the cumulative frequencies of the CBS item obtained in other waves. The estimated mean on the 0-10 continuum for each of these waves is equal to the mean of the corresponding best fitting beta distribution resulting from the application of the Continuum Approach. To exemplify this, the cumulative frequencies of each wave of the CBS item in the period 1997 to 2009 and the parameters of the best fitting beta distributions are given in Tab. C.1 of appendix C.

An obvious question of interest is how the boundaries found using the Reference Distribution Method relate to the boundaries obtained using the Scale Interval Method, where the boundaries are based on assessments made by judges. This relationship is depicted in Fig. 19 for the CBS and the EB items, to give an impression of what the difference between the two methods means for the positions of the boundaries on the reconfigured scales.

**Figure 19 Boundaries as based on the assessments by judges or based on a reference distribution**




From Fig. 19 it can be seen that according to the Scale Interval Method the interval for the response option 'Not very satisfied' in the CBS item, does not overlap with the interval for this option as assessed for the similar option in the EB-item. The latter interval is fully covered by the interval for the response option 'Fairly satisfied' in the CBS item according to the Scale Interval Method. When the boundaries are derived from a reference distribution as done in the Reference Distribution Method, they show a dramatic change compared to those obtained using the Scale Interval Method. The boundaries based on the Reference Distribution Method for the CBS scale are more in harmony with those for the EB scale compared to the results obtained using the Scale Interval Method. Using the Reference Distribution Method, the interval for the response option 'Very satisfied' of the EB scale almost coincides with the combination of the intervals for the response options 'Very satisfied' and 'Extraordinarily satisfied'. On the other side of both scales a similar correlation can be noticed for the interval for the response option 'Not very satisfied' of the CBS scale with the combined intervals for the response options 'Not at all satisfied' and 'Not very satisfied' of the EB scale.

### 7.3 Scale transformation using the Reference Distribution Method

In the Reference Distribution Method the reference distribution used is the beta distribution that fits best to the frequency distribution in a certain

year, the reference year, of a happiness item from a deliberately selected survey. Suppose we want to transform the results of another survey for a specific item with a verbal response scale to the continuum from 0 to 10 using the Reference Distribution Method. To do so, given that the results of this other survey are also measured in the reference year, the positions on the continuum from 0 to 10 of the boundaries between the response options of the specific item can be derived from the reference distribution in the way we illustrated in Fig. 18. Once these boundaries have been derived, they are kept fixed in the Reference Distribution Method for the transformation of the survey results for the specific item measured in other years. In other words, to transform survey results for other years, the boundaries remain equal to those derived from the reference distribution for the reference year.

The transformation for each of the other years in which the survey has been fielded consists of estimating the parameters of the best fitting beta distribution based on the boundaries derived from the reference distribution and on the frequency distribution of the response on the primary verbal scale in the year in progress. The estimated survey mean is subsequently the outcome of the division of 10 * $\hat{\alpha}$ by $\hat{\alpha} + \hat{\beta}$, see formula (Eq. 3) in Sec. 6.2, with $\hat{\alpha}$ and $\hat{\beta}$ the estimated parameters of this best fitting beta distribution. The survey results of a whole time series can be transformed in this way.

In a certain year however, the mode of surveying may be changed. If so, it is plausible that this will influence the position of the boundaries between response options. An example of the effect a mode change can have is the Life Situation Survey of SCP in the Netherlands, which in 2004 was changed from face-to-face interviews responding to a questioner to a paper-and-pencil survey using a questionnaire. This mode change caused a dramatic fall in the percentage of people who rated themselves as either 'Happy' or 'Very happy', a drop of 6 percentage points from 2002 to 2004 in a time series that had been rather stable since 1997. In such a situation, the position of the boundaries has to be reconsidered and presumably determined anew. To derive new boundaries that comply with the new survey mode, the original reference distribution should not be used. Instead the best fitting beta distribution given the boundaries derived from the original reference distribution and the frequency distribution of the survey results in the year prior or equal to that in which the mode was changed should be selected as a new reference distribution. Whether the new reference distribution should be based on the survey results for the

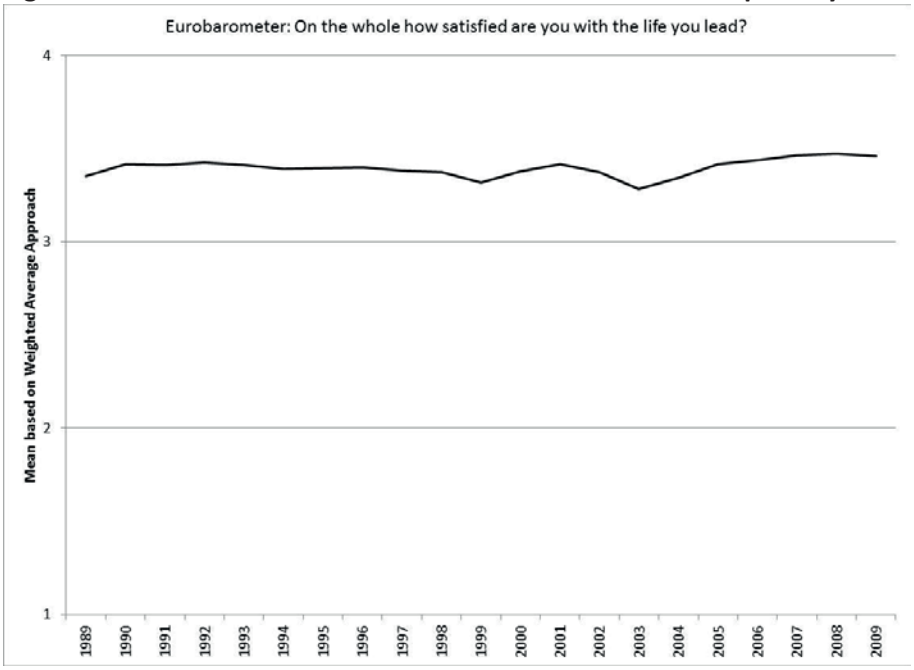
year the mode was changed or for the year prior to that, depends on whether there has been a double measurement: in the ideal situation a survey will be fielded in both modes in the year of change to get insight into the effect of the change. In this case the new reference distribution can be based on the survey results for the same year the mode was changed. If unfortunately no double measurement is available, but the survey results show minor changes from year to year, as a proxy the best fitting beta distribution estimated for the year prior to the year the questionnaire mode was changed can be used.

In the same way, two different surveys to measure happiness that partially overlap in the years they have been fielded can be transformed and combined if a reference distribution is available for one of them. This reference distribution does not necessarily have to be based on a different (third) survey, but can also be derived from one of the two surveys of concern. In this case a reference year has to be selected from the time period in which both surveys have been fielded. Next one of the two surveys should be selected to provide the reference distribution. If the item of interest in this survey has a numerical scale, a reference distribution can be estimated straightforwardly just as it is done for the example from the ESS. If, however, this item has a verbal scale, the boundaries between the response options must be specified first and the Scale Interval Recorder can be used for this purpose. The reference distribution can be estimated using these specified boundaries and the frequency distribution for the item in the reference year. Given the reference distribution, the time series of both surveys can then be transformed in the way we described earlier.

### 7.4 Application of the Reference Distribution Method

We will now illustrate how the Reference Distribution Method is used by applying it to the items taken from the CBS and the EB surveys for results obtained in the years from 1993 to 2009. This application consists of a trend analysis in terms of the comparability of the trends in responses to different questions about happiness in one country. In most of the years of this period, the EB was fielded in the spring and autumn. To demonstrate the Reference Distribution Method, we have selected the results for just one measurement per year. If available, we selected the results obtained in spring otherwise we incorporated the results for autumn. The means of the EB item in the period 1989-2009 when the common Weighted Average Approach was applied are given in Fig. 20.

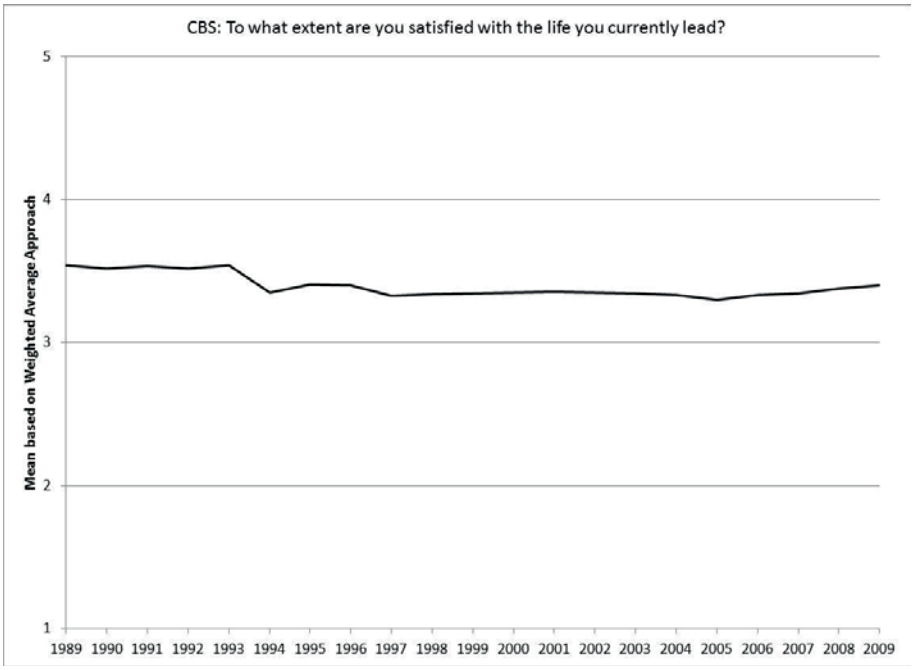
**Figure 20 Means of the EB item in 1989-2009 based on ranks primary scale**




In most of the years until 1996 the mean value of the EB item was nearly 3.40. In the following years dips were seen in the years 1999 and 2003 and from 2004 the line has climbed to around 3.46 in 2007 and this has been maintained until 2009.

In the period 1989-2009, there were two changes in the CBS survey that affected the responses. The first change was made in 1994 and consisted primarily of a comprehensive revision of the questionnaire forms and a reduction of the survey items in several domains. A major change of the survey design of the CBS survey took place in 1997. Amongst others, the mode of questioning was changed from paper-and-pencil surveying to face-to-face interviews and instead of drawing samples based on addresses, from then on the sample was drawn based on individual citizens. This change affected the survey results. We present the means of the CBS item for the period 1989-2009, when the common Weighted Average Approach was used, in Fig. 21.

**Figure 21 Means of the CBS item in 1989-2009 based on ranks primary scale**




The changes in the design in 1994 and 1997 of the CBS survey are clearly visible in the mean value presented in Fig. 8. In each of the three periods that can be distinguished for the CBS item, the mean values show a rather stable pattern.

We used the reference distribution based on the ESS results of 2008 to derive the boundaries between the response options of the items from both the EB and CBS surveys. Using these tuned boundaries we estimated the parameters of the best fitting beta distributions for the CBS results over the years 1997 to 2009 and for the EB results over the years 1994 to 2009. Fortunately in 1997 the CBS survey was fielded in both the old and the new design, therefore a best fitting beta distribution was available based on the survey results for 1997 according to the new design and on the boundaries derived from the ESS reference distribution. This best fitting beta distribution for 1997 and the survey results over 1997 according to the old design, we used to derive the boundaries between the response options for the survey results obtained in the years 1994-1996. In 1993 there was no double measurement. Therefore we used the beta distribution estimated

for 1994 as a reference to transform the survey results obtained in the period 1989-1993.

The time-invariant boundaries as assessed by the judges in the Scale Interval Method, the boundaries derived from the reference distribution based on the ESS results for 2008 and the adjusted boundaries for the changes in design for the CBS survey in 1997 and 1994 are given in Tab. 11.

**Table 11 Upper boundaries of response options CBS scale and EB scale**

| ***Item code Survey*** | ***Response options*** | ***Upper boundaries*** | | | |
|---|---|---|---|---|---|
| | | *Judges* | *Ref ESS 2008* | *Ref CBS 1997* | *Ref CBS 1994* |
| O-SLL-c-sq-v-5-d CBS | - Extraordinarily satisfied | 10.0 | 10.0 | 10.0 | 10.0 |
| | - Very satisfied | 8.8 | 9.0 | 8.8 | 8.6 |
| | - Satisfied | 7.2 | 7.8 | 7.5 | 7.2 |
| | - Fairly satisfied | 5.3 | 5.7 | 5.8 | 5.5 |
| | - Not very satisfied | 3.6 | 4.8 | 4.9 | 4.5 |
| O-SLL-u-sq-v-4-b EB | - Very satisfied | 10.0 | 10.0 | | |
| | - Fairly satisfied | 7.9 | 7.5 | | |
| | - Not very satisfied | 5.3 | 4.7 | | |
| | - Not at all satisfied | 3.0 | 3.6 | | |

In addition to what we exemplified for the difference in the position of the boundaries as presented in Fig. 19 when comparing the Scale Interval Method and the Reference Distribution Method, we can remark that before the design change of the CBS survey in 1997 the boundaries of the response options in the higher part of the scale were positioned a little lower and those in the lower part of the scale slightly higher. All the boundaries for the period 1989-1993 tuned to the reference distribution for 1994 are positioned somewhat lower on the continuum compared to the boundaries for the period 1994-1996.

In Fig. 22 the results based on boundaries obtained by application of the Scale Interval Method are shown and in Fig. 23 the results based on boundaries according to the Reference Distribution Method: for reasons of comparison, besides the results for the CBS and the EB items, we have also included in the graphs of Fig. 22 and 23 the results for the ESS item of the survey waves for 2002, 2004, 2006 and 2008.

**Figure 22 Results by application of the Scale Interval Method**

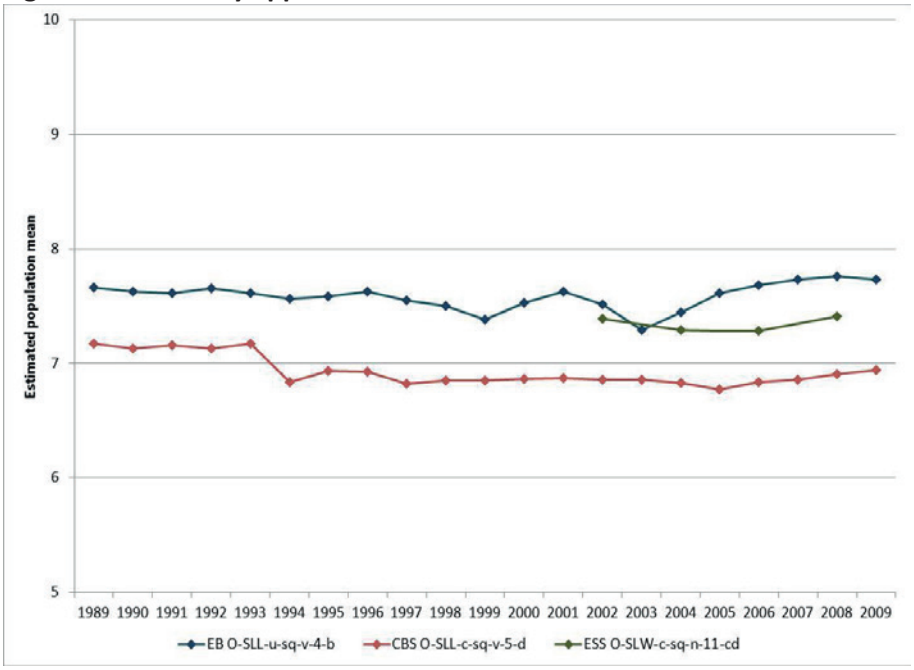
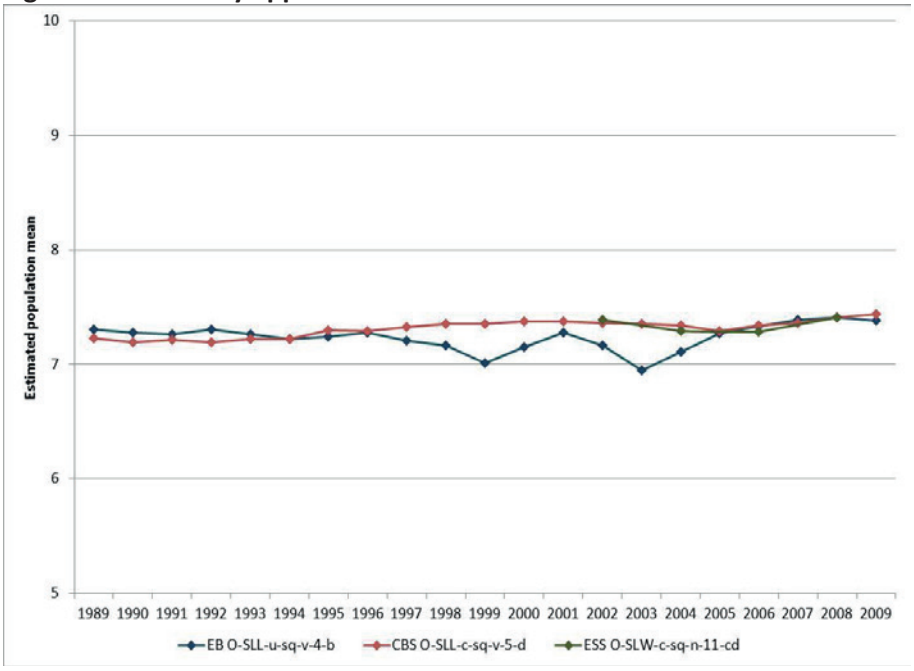



**Figure 23 Results by application of the Reference Distribution Method**



As can be seen, when applying the Scale Interval Method, the estimated population means for the EB item are too high compared to those for the ESS item, whereas for the CBS item they are too low. The means for the CBS item when using the Scale Interval Method furthermore show a large discontinuity in the transition from 1993 to 1994 and a small discontinuity in the transition from 1996 to 1997, which is due to changes in the survey design. After application of the Reference Distribution Method, the estimated means for the EB item are somewhat lower compared to the application of the Scale Interval Method, whereas the Reference Distribution Method causes an upward shift for the CBS results. Due to the adjustment of the boundaries for 1993 and 1997, the discontinuities from 1993 to 1994 and from 1996 to 1997 have also disappeared. The fluctuations in each survey over the years turn out to be similar for the results when applying the Reference Distribution Method and the results obtained by the Scale Interval Method. Application of the Reference Distribution Method brought the results for all three the surveys to a comparable level.

## 7.5 Discussion

In this chapter we introduced the Reference Distribution Method, which in our opinion, provides a valid way to estimate population means based on ratings on verbal and discrete numerical scales at truly comparable levels on a continuum from 0 to 10.

### *7.5.1 Strengths of scale transformation using a reference distribution*

The Reference Distribution Method is a variation of the Scale Interval Method and tunes survey results to the level of a reference distribution in a reference year. We have shown that this Reference Distribution Method is an effective tool for transforming survey results obtained with different items on the same topic to a comparable scale. In addition, the Reference Distribution Method allows corrections to be made for discontinuities due to changes in the design of a survey. As such the Reference Distribution Method can be used to extend time series as it permits combining results from different surveys that have been fielded in, partly, overlapping periods in time. We have elaborated this in Ch. 10.

### *7.5.2 Limitations*

The Reference Distribution Method can be used to correct much of the differences seen in different sets of findings on happiness that are due to dissimilarity in the measures used; yet it cannot solve all the comparability

problems.

One limitation is that the method requires a reference distribution, typically a survey in which the same subject is assessed using a 0-10 numerical scale in the same country in the same year. If not, as a second best option for transforming distributions on numerical scales the Scale Interval Method should be used, preceded, in the case of a verbal response scale, by a Scale Interval Study.

If a survey has been fielded only once and there is a reference distribution available, then the mean based on the upper boundaries derived from a reference distribution, is by definition, equal to the mean of this reference distribution. This saddles the transformed scores with the errors of the reference distribution, which causes them to become systematic rather than random.

The boundaries between response options that have been derived from a reference distribution are kept fixed as long as the survey design has not undergone a significant change. An obvious question that can be raised is whether it is a reasonable assumption that the boundaries will be more or less stable over time. The answer is affirmative which we will discuss in Ch. 8.

The primary verbal scales of the two items we used in this chapter to illustrate how the Reference Distribution Method works both had more than three response options. The Reference Distribution Method is invalid if a verbal scale has only two response options. There is always a perfectly fitting beta distribution, though with zero degrees of freedom, for a primary scale with only three response options. Some 15% of the survey studies on happiness in nations is based on 2- and 3-point response scales (Veenhoven, 2012) and thus cannot be used for comparison with the other 85% of the research findings using the Reference Distribution Method.

Another limitation is that the Reference Distribution Method applies only to the diversity in rating scales, that is, to the last three aspects of the differences in survey questions presented in Tab. 1. Survey questions on happiness also differ in the wording of the leading question, such as in the key word used, for example ‘happiness’ or ‘satisfaction with life’. Furthermore, the questions differ also in the time frame that is addressed, some referring to ‘current’ happiness, while others ask the respondent to appraise ‘the last year’. In addition, the Reference Distribution Method has been developed to be applied to single item questions. Yet, there are also multiple question inventories, such as Diener’s (1985) five item ‘satisfaction with life scale’. Although each of these items can be tuned in principle, the

chance of finding good reference items is lower than that for the case of single items.

## 7.6 Conclusion

Survey studies on the same topic often use different questions. One of the differences is in the response scales, which commonly differ in the number of options in verbal and numerical scales used and in the words used to label the response options or anchor points. As a result much of the available research findings cannot be compared. Several methods have been proposed for transforming observed scores on these different scales into common scores, typically on a 0-10 numerical scale. All of these methods have limitations and the transformed scores they produce appear to differ substantially from distributions obtained directly using 0-10 numerical scales. The Reference Distribution Method proposed in this thesis performs better.

# 8 Stability of the boundaries between response options

## 8.1 Research question

When the Continuum Approach is applied to the time series of a survey which has remained unchanged over time, the transition points are kept fixed. The idea behind this is that, although people may change the perception of the intensity of, for example, their own happiness intensity over time, they are assumed not to change the degree of appreciation they attribute to the terms used to label response options. This is an important assumption for research syntheses that require that everything remains unchanged, except for the change of interest. It means that if the Continuum Approach is applied to measurements at distinct points in time, differences in estimates of the mean and standard deviation can be solely attributed to changes in the frequency distributions on the primary scale. Thus the research question addressed in this section is: Is it reasonable to keep the transition points between response options fixed when we apply the Continuum Approach?

For this research question, we had four survey items available which were fielded in 2008. These were the CBS and EB items on life satisfaction and the CBS and Dutch Household Survey (DHS) items on happiness which we have described in Sec. 1.3 and have presented in Tab. 2 and 3 of that section. We have used a reference distribution derived from the ESS item on life satisfaction to derive reference boundaries for the CBS item on life satisfaction to demonstrate the Reference Distribution Method in Sec. 7.2. This same reference distribution can be used to derive reference boundaries for the EB item. For the two items on happiness, we used the reference distribution derived from the ESS data for 2008 on happiness, which we have also described in Sec. 1.3. The verbal scale items, their frequency distribution in 2008 and the reference boundaries obtained by application of the Reference Distribution Method, are summarized in Tab. 12 to 15.

**Table 12 CBS life satisfaction 2008**

| | To what extent are you satisfied with the life you currently lead? | | | | |
|---|---|---|---|---|---|
| | Not very satisfied | Fairly satisfied | Satisfied | Very satisfied | Extraordi-narily satisfied |
| Frequency | 3.4% | 7.6% | 45.1% | 35.5% | 8.4% |
| Reference boundary | 4.78 | 5.73 | 7.77 | 9.04 | 10.00 |

**Table 13 EB life satisfaction 2008**

| | On the whole how satisfied are you with the life you lead? | | | |
|---|---|---|---|---|
| | Very unsatisfied | Not very satisfied | Fairly satisfied | Very satisfied |
| Frequency | 0.6% | 3.1% | 44.8% | 51.5% |
| Reference boundary | 3.69 | 4.82 | 7.51 | 10.00 |

**Table 14 CBS happiness 2008**

| | To what extent do you consider yourself a happy person? | | | | |
|---|---|---|---|---|---|
| | Unhappy | Not very happy | Neither happy nor unhappy | Happy | Very happy |
| Frequency | 3.4% | 7.6% | 45.1% | 35.5% | 8.4% |
| Reference boundary | 4.26 | 5.13 | 6.05 | 8.46 | 10.00 |

**Table 15 DHS happiness 2008**

| | Taking all together, to what extent do you think of yourself as a happy person? | | | | |
|---|---|---|---|---|---|
| | Very unhappy | Unhappy | Neither happy nor unhappy | Happy | Very happy |
| Frequency | 0.3% | 1.6% | 16.7% | 61.6% | 19.7% |
| Reference boundary | 3.96 | 4.88 | 6.53 | 8.55 | 10.00 |

## 8.2 Approach for testing the stability of boundaries

If the Reference Distribution Method is applied, the best fitting beta distribution in the reference year by definition coincides with the cumulative frequency distribution of the verbal response scale at the position of the reference boundaries. It is unlikely that this coincidence will

also occur exactly for the beta distribution that fits best to these reference boundaries and the cumulative frequencies of a verbal scale item measured at different moments in time.

We recall from Sec. 7.2 that using the Reference Distribution Method forces the cumulative frequency of a verbal scale item in the reference year into the curve of a corresponding reference distribution and leaves us with a set of reference boundaries. As a result the estimated mean on basis of the verbal scale in the reference year is equal to that of the mean reference distribution.

The main reason for determining the reference boundaries is that they are necessary for the transformation of time series of means based on measurements using verbal scale items into time series of mutually comparable means. To achieve this, the Continuum Approach is applied to estimate the best fitting beta distribution for each frequency distribution of the time series of a given item and the reference boundaries for this item derived from the reference distribution in the reference year. In this way we get a series of beta distributions for each item in which every beta distribution is based on the same reference boundaries but each one on its own frequency distribution. The reference boundaries are thus kept fixed over time, whereas the frequency distributions vary within each time series. In doing so, we implicitly assume that the boundaries between the response options are stable over time and that the differences in the estimated means can be attributed solely to changes in the frequency distributions on the same verbal scale. When we use the term 'stability of the boundaries between response options over time' we mean that if we apply the Continuum Approach to estimate a beta distribution which fits best to the cumulative frequencies positioned at the fixed reference boundaries for a survey item at different moments in time:

- the beta distribution that fits best to the frequency distribution of each wave may only slightly deviate from the observed cumulative frequencies at the positions of the reference boundaries
- if there is a deviation, its size should not be related to the length of the period between the time of measurement and the reference year

The horizontal and vertical deviation can be formulated formally as follows.

- ***The deviation in horizontal direction:*** for each response option *i* this is the difference between reference boundary *i* and the position on the continuum where the cumulative frequency of the response option is equal to the value of the best fitting cumulative beta distribution. We will go into this in Sec. 8.4.

- ***The deviation in vertical direction:*** for each response option *i* this is the difference between the cumulative frequency of the response option and the value of the best fitting cumulative beta distribution at the position of reference boundary *i*. We will go into this in Sec. 8.5.

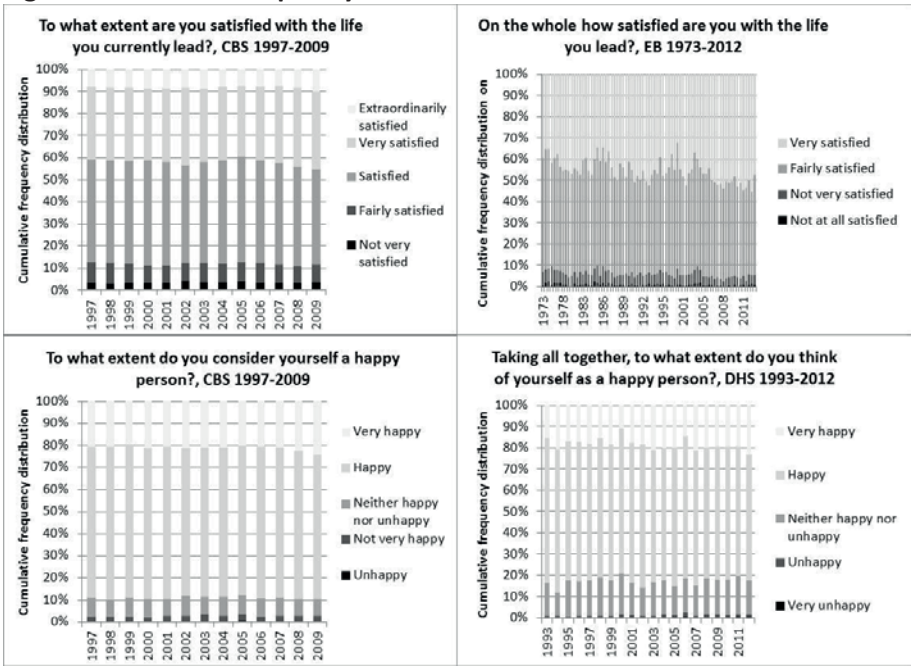
If for a given measurement both the horizontal deviation and the vertical deviation of the estimated beta distribution to the cumulative frequencies of the primary scale when positioned on the reference boundaries are small, it means that the estimated beta distribution fits well to the measurement on the primary scale. What 'small' in this context means, is a subjective judgement.

### 8.3 Available time series

For each of the survey items that we mentioned in Sec. 8.1 we had both the frequency distribution of 2008 available, and an entire time series. For the two items from CBS we had one frequency distribution for each year in the period 1997-2009 when the mode of surveying was unchanged. Frequency distributions for the DHS item were available for the period 1993-2012. The EB item was fielded in The Netherlands almost every year for one to four times between 1973 and 2012 (Schmitt et al, 2008; European Commission, 2012a, 2012b and 2013). An overview of the frequency distributions for the various surveys is given in Fig. 24, in which the stack diagrams are projections of the cumulative frequency distributions on the vertical scale.

Of all the four items presented in Fig. 24, the frequency distribution of the responses to the Eurobarometer item had the most fluctuating pattern over time, by which we mean that the share of respondents that select a certain response option largely fluctuated over the time period of the survey. The options 'Fairly satisfied' and 'Very satisfied' were dominant in the ratings for each year but regularly changed places over the years with respect to representing the highest frequency. The two CBS items showed the least fluctuations over time. The cumulative distributions of all items were skewed to the left, which means that they had a relatively long tail on the left where there were relatively few observations. Furthermore it is worth noting that for the two items on happiness with the most negative formulated options, 'Unhappy' for the CBS item and 'Very unhappy' for the DHS item, were nearly never chosen by the respondents.

**Fig. 24 Cumulative frequency distributions**



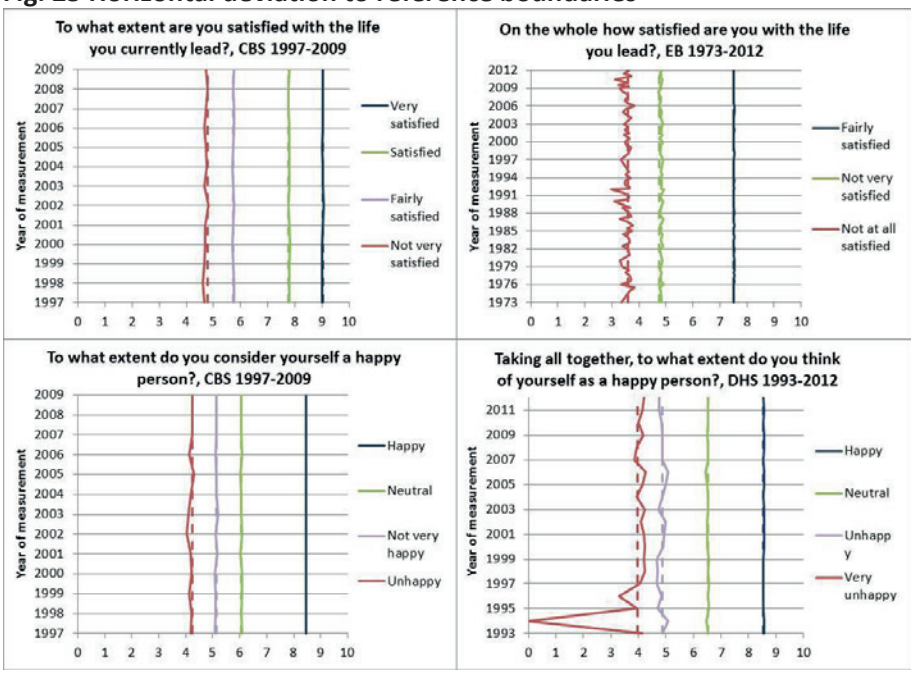

We applied the Continuum Approach to estimate a best fitting beta distribution to each frequency distribution and the corresponding reference boundaries for each item: for both the CBS item on life satisfaction and the CBS item on happiness we thus estimated 13 beta distributions, for the Eurobarometer item which was fielded several times in a year, we estimated 75 beta distributions and for the DHS item 20. An overview of the cumulative frequencies on the primary scales and of the parameters of the best fitting beta distributions is given in appendix C. We determined the horizontal deviation and the vertical deviation to the corresponding cumulative frequency distribution on the primary verbal scale for each of these beta distributions. The results are described in Sec. 8.4 and Sec. 8.5.

## 8.4 The deviation in horizontal direction

The deviation from the reference boundaries in the horizontal direction is an obvious choice of deviation from an intuitive point of view, since it gives insight into the distance between the reference boundary and the point on the continuum where the value of the best fitting cumulative beta distribution equals the cumulative frequency for a given response option. The fluctuations of the horizontal deviation over time for each response

option are presented in Fig. 25 where the reference boundaries of each item are represented by straight dashed lines. Since the value for the reference boundary of the most positively labelled option of each item is, by definition, equal to 10, this trivial boundary is ignored in the analysis of the stability of the boundaries.

**Fig. 25 Horizontal deviation to reference boundaries**




The deviation from the reference boundaries in the horizontal direction is the largest for cumulative frequencies in the lower tail of the distribution. This has to be attributed to the small slope of the cumulative beta distribution in the tail and does not necessarily imply a large deviation in the vertical direction as we will show in Sec. 8.5. In line with the low fluctuations of the CBS items over time, the horizontal deviations from the reference boundaries for these items are most stable. For the Eurobarometer item, the point on the continuum where the value of the beta distribution is equal to the cumulative frequency for the option 'Not very satisfied' is the only one that is positioned to the right of the reference boundary for nearly all of the observations over the years. For the other response options of this item these points curl around the reference boundaries. Most eye-catching for the DHS item, is the horizontal deviation

to the reference boundary of the response option ‘Very unhappy’ in 1994. This extremity is due to the fact that the response to this option was zero in 1994. Considering the low frequency at which this option was ticked over the course of time, it should be questioned whether it would not be better to combine this option with the option ‘Unhappy’ in the analysis. This would reduce the effective scale of the DHS item from five to four points, but given the low response to both options, it is not likely that this would affect the sample mean much.

The frequency distributions of the horizontal deviations from the reference boundaries over the years as presented in Fig. 25, characterized by their average value, the standard deviation and the standard error of this average value are summarized in Tab. 16 to 19. As a reference we have also included the values of the reference boundaries which we described in Sec. 8.1. Example: the horizontal deviation of the reference boundary to the cumulative frequency of the response option ‘Not at all satisfied' of the Eurobarometer item (Table 17) varies with a standard deviation of 0.16 about an average horizontal deviation of 0.06; this average is estimated with a standard error of 0.02. The horizontal deviation is positive if the value on the 0-10 continuum where the beta distribution is equal to the cumulative frequency is higher than the value of the reference boundary of the corresponding response option.

On average the horizontal deviation from the reference boundaries is small for all the response options of each item. The standard deviation is the largest, with a value of 0.93, for the option ‘Very Unhappy’ of the DHS item, which is mainly to be attributed to the zero response to this option in 1994. If the horizontal deviation in 1994 for this option is not taken into account, the standard deviation would reduce to 0.22. Leaving out the results for 1994, would reduce the standard error of the average horizontal deviation for the DHS item to 0.05. The relative high standard deviation for the horizontal deviation to the response option ‘Not at all satisfied’ of the Eurobarometer can be explained by the fact that the cumulative frequency for this option is situated in the lower tail of the life satisfaction distribution. In this part of the scale the number of respondents to this option is very small: in the Eurobarometer survey usually < 3% and often < 1%. Since the observed relative frequency acts as a weight of the contribution of the corresponding response option to the estimated parameters of the population distribution, it is acceptable to ignore its effect on the final conclusion.

**Table 16 Distribution horizontal deviation CBS life satisfaction item 1997-2009 (13 waves)**

| Indicators | Not very satisfied | Fairly satisfied | Satisfied | Very satisfied |
|---|---|---|---|---|
| *Reference boundary* | *4.78* | *5.73* | *7.77* | *9.04* |
| Average horizontal deviation | 0.076 | -0.002 | -0.018 | 0.016 |
| Standard deviation | 0.06 | 0.02 | 0.02 | 0.01 |
| Standard error | 0.02 | 0.01 | < 0.01 | < 0.01 |

**Table 17 Distribution horizontal deviation EB item 1973-2012 (75 waves)**

| Indicators | Not at all satisfied | Not very satisfied | Fairly satisfied |
|---|---|---|---|
| *Reference boundary* | *3.60* | *4.76* | *7.51* |
| Average horizontal deviation | 0.060 | -0.066 | -0.003 |
| Standard deviation | 0.16 | 0.04 | 0.01 |
| Standard error | 0.02 | < 0.01 | < 0.01 |

**Table 18 Distribution horizontal deviation CBS happiness item 1997-2009 (13 waves)**

| Indicators | Unhappy | Not very happy | Neither happy nor unhappy | Happy |
|---|---|---|---|---|
| *Reference boundary* | *4.26* | *5.13* | *6.05* | *8.46* |
| Average horizontal deviation | 0.070 | 0.007 | -0.007 | 0.001 |
| Standard deviation | 0.07 | 0.04 | 0.02 | < 0.01 |
| Standard error | 0.02 | 0.01 | < 0.01 | < 0.01 |

**Table 19 Distribution horizontal deviation DHS happiness item 1993-2012 (20 waves)**

| Indicators | Very unhappy | Unhappy | Neither happy nor unhappy | Happy |
|---|---|---|---|---|
| *Reference boundary* | *3.96* | *4.88* | *6.53* | *8.55* |
| Average horizontal deviation | 0.091 | 0.031 | 0.011 | -0.004 |
| Standard deviation | 0.93 | 0.13 | 0.03 | 0.01 |
| Standard error | 0.21 | 0.03 | 0.01 | < 0.01 |

## 8.5 The deviation in vertical direction

The deviation of the beta distribution from the primary frequency distribution at the reference boundaries in vertical direction gives insight into the extent to which the best fitting beta distribution under- or overestimates the cumulative frequency of a response option at the position of the corresponding reference boundary. We summarized the vertical deviation of the best fitting beta distributions from the cumulative frequencies of each response option of all items by their average, and the standard deviation and the standard error from this average over the years expressed in percentage points in Tab. 20 to 23. Unlike for the horizontal deviation from the reference boundaries, there is no reference value to compare the vertical deviation of the beta distribution from the primary frequency distribution at the reference boundaries. By a positive vertical deviation we refer to an overestimation of the primary frequency distribution by the beta distribution and by a negative vertical deviation to an underestimation of the primary frequency distribution.

The average vertical deviation of the reference distribution from the cumulative frequencies is small for all response options of each item, with a very small standard error. It is worth noting here, that when mutually comparing the average vertical deviations for two response options, one should keep in mind that these deviations are related to the cumulative frequencies. For example, the average vertical deviation for the response option 'Very satisfied' of the CBS item on life satisfaction (0.31 %pts) is not much smaller than that for the response option 'Not very satisfied' (0.38 %pts), but relatively speaking, the difference between the two average deviations is much larger: the 0.38 percentage points for the option 'Not very satisfied' corresponded to an on average cumulative frequency of less than 4%, whereas the 0.31 percentage points for the option 'Very satisfied' belonged to an on average cumulative frequency of over 90%. Keeping this in mind, the vertical deviations corresponding to the response options on the left side of the scale were relatively larger than those corresponding to the response options on the right side.

**Table 20 Distribution vertical deviation CBS life satisfaction item 1997-2009 (13 waves)**

| Indicators | Not very satisfied | Fairly satisfied | Satisfied | Very satisfied |
|---|---|---|---|---|
| Average vertical deviation | 0.38 %pts | -0.03 %pts | -0.54 %pts | 0.31 %pts |
| Standard deviation | 0.28 %pts | 0.26 %pts | 0.48 %pts | 0.25 %pts |
| Standard error | 0.08 %pts | 0.07 %pts | 0.13 %pts | 0.07 %pts |

**Table 21 Distribution vertical deviation EB item 1973-2012 (75 waves)**

| Indicators | Not at all satisfied | Not very satisfied | Fairly satisfied |
|---|---|---|---|
| Average vertical deviation | 0.22 %pts | -0.04 %pts | -0.04 %pts |
| Standard deviation | 0.26 %pts | 0.32 %pts | 0.22 %pts |
| Standard error | 0.03 %pts | 0.04 %pts | 0.03 %pts |

**Table 22 Distribution vertical deviation CBS happiness item 1997-2009 (13 waves)**

| Indicators | Unhappy | Not very happy | Neither happy nor unhappy | Happy |
|---|---|---|---|---|
| Average vertical deviation | 0.08 %pts | 0.03 %pts | -0.10 %pts | 0.04 %pts |
| Standard deviation | 0.09 %pts | 0.18 %pts | 0.21 %pts | 0.10 %pts |
| Standard error | 0.02 %pts | 0.05 %pts | 0.06 %pts | 0.03 %pts |

**Table 23 Distribution vertical deviation DHS happiness 1993-2012 (20 waves)**

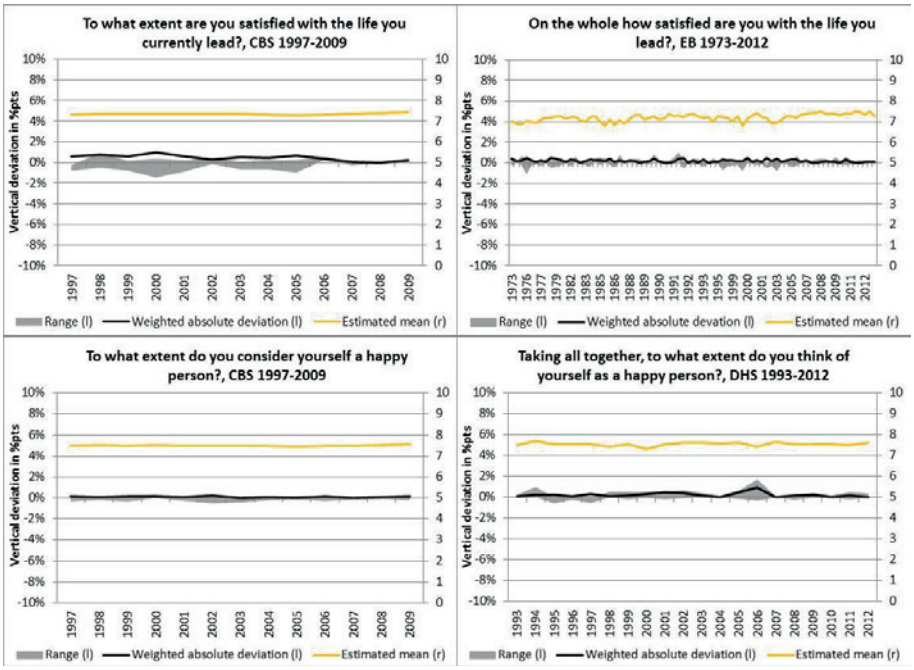
| Indicators | Very unhappy | Unhappy | Neither happy nor unhappy | Happy |
|---|---|---|---|---|
| Average vertical deviation | -0.09 %pts | 0.07 %pts | 0.21 %pts | -0.13 %pts |
| Standard deviation | 0.11 %pts | 0.34 %pts | 0.52 %pts | 0.24 %pts |
| Standard error | 0.03 %pts | 0.08 %pts | 0.12 %pts | 0.05 %pts |

To finalize our analysis, we calculated two indicators for the goodness-of-fit of the beta distributions. The first of these indicators was the range of the vertical deviation which is defined as the absolute difference between the minimum vertical deviation and the maximum vertical deviation of the

reference distribution from the cumulative frequencies, excluding the trivial deviation for the reference boundary at position 10, which, by definition, is equal to zero. The second indicator was the weighted absolute deviation, which is equal to the weighted average of the absolute vertical deviation for each reference boundary except the one at position 10, with the relative frequencies of the corresponding response options as weights. The use of these frequencies as weights was an arbitrary choice. Other possibilities would be the squares or roots of these frequencies and also the cumulative frequencies. The idea behind the weighted absolute deviation is that a large deviation that corresponds to a low frequency has a lower impact on the value of the estimated mean than a large deviation that corresponds to a high frequency. The range is an indicator that gives insight into whether the upper boundaries of all response options are overestimated or underestimated by the best fitting beta distribution or whether there is a mixture of over- and underestimations. The weighted absolute deviation provides guidance as to the extent to which the deviations affect the estimated population means. The range and the weighted absolute deviation of the vertical deviation are plotted on the left axis in Fig. 26. The underlying idea of the research question in this section was to make responses to different survey items which aim at measuring the same item, here happiness, comparable. Therefore we also included the estimated means for all the transformed time series of measurements in Fig. 26, which are plotted on the right axis.

On average the range over time was small for all four survey items. The average range over the years was largest for the CBS item on life satisfaction with a value of 0.72. For this item the standard deviation of the average range was also largest with 0.50 percentage points. The, in absolute sense, largest range occurred for the DHS item on happiness in 2006 when it was equal to 1.7. Clearly, the range for the reference measurement in 2008 was equal to zero for all items. The largest weighted absolute deviation could be observed for the CBS item on life satisfaction in 2000 when it was equal to one percentage point. The average weighted absolute deviation for this item was equal to 0.46 which was at least twice as large as the average for each of the other items. For each item neither the range nor the weighted absolute deviation showed a relationship between the size and the distance in time between the moment of measurement and the reference year.

**Fig. 26 Range and weighted absolute deviation best fitting beta distribution**




## 8.6 Discussion

The results we found were similar for all the survey items we considered in this section; for each response option for all four items we considered, we found that both the average horizontal deviation and the average vertical of the estimated beta distributions to the cumulative distributions of the primary scales were small. From this we conclude that the beta distributions fit well to the primary distributions. In addition, the standard error of each of the deviations was very small. The latter means that the average deviations in both horizontal and vertical direction are stable over time. We conclude that these results confirm the implicit assumption we formulated in Sec. 8.2 that the boundaries between response options are stable over time and that the differences in transformed means can solely be attributed to changes in the frequency distributions on the primary scale.

We need to remark that we used the range of the vertical deviation and the weighted absolute deviation between the cumulative frequency of the primary scale and the value of the best fitting beta distribution at the

position of the reference boundary as goodness-of-fit indicators. Future research is necessary to validate these indicators or to develop better ones.

When the Reference Distribution Method is applied, the estimated population mean of the reference distribution serves as a reference value for a comparison of the means estimated using the Continuum Approach to other survey items and other years of measurement. This reference value should not be considered to be the 'true' value of the perception of happiness on the continuum from 0 to 10. If another reference distribution is used, the reference value may be different. In other words, only population means that are estimated by using the Continuum Approach to survey items for which the reference boundaries are derived from the same reference distribution, can be compared.

## 8.7 Conclusion

The question we addressed in this chapter is whether it is reasonable to assume that the positions of the reference boundaries between response options on the continuum from 0 to 10 are stable over time if the Continuum Approach is applied. We conclude that the answer is affirmative, at least in the case of The Netherlands, which justifies that it is valid to use these boundaries as fixed values over a longer period of time.

# 9 Robustness of the conversion of verbal response scales across demographic categories

## 9.1 The consequence of a change from using a verbal scale to using a numerical scale

In Sec. 1.3 we already remarked that the present trend in measuring subjective well-being, is to use 10- and 11-point numerical scales with only the anchor points defined by verbal labels. A change from using a verbal to using a numerical scale to measure happiness and life satisfaction however, leads to an instantaneous discontinuity in the existing time series. This causes a severe problem for trend analyses, due to incomparability of the old and new measurements, meaning that, with the introduction of the numerical scale, current trends in well-being across demographic categories cannot be continued for long time series unless the discontinuity caused by the change of the scale type can be overcome in a suitable way.

In the previous sections we have shown that the Reference Distribution Method is a promising method with respect to the problem of instantaneous discontinuity sketched above. To date, in the development of the Reference Distribution Method the focus has been placed on the general population in nations and no distinction has been made between demographic groups within a nation. The research question we address in this chapter is: Can the reference boundaries derived for the general population be used for demographic categories to produce reliable extended time series to monitor differences in trends between these categories?

## 9.2 Availability of data for different demographic categories

An application of the Reference Distribution Method for trend analyses in different demographic categories requires time series of the responses to each of these categories on a verbal scale and for one wave of measuring the same topic on a numerical scale. A set of time series based on survey items taken from Statistics Netherlands (CBS) data and the Eurobarometer (EB) provided the data for the application described in this section.

### *9.2.1 Happiness and satisfaction with life: Statistics Netherlands*

CBS measured happiness[15] and satisfaction[16] with life with the items we described in Sec. 1.3 in the Permanent Onderzoek Leef Situatie[17] (POLS) in

[15] Happiness: http://statline.cbs.nl/StatWeb/publication/?DM=SLEN&PA=60027ENG&D1=43-47&D2=1-2,12,15,19-22,29-31&D3=a&LA=EN&VW=T

the period 1997-2008 and in the Social Cohesion Survey in 2009 and 2010. On average the number of respondents to the survey items on these topics added up to approximately 8,300 per year. CBS changed to a mix-mode of surveying in 2010. This caused a drop by approximately 5 percent in the percentage of happy and satisfied persons in the population according to the dichotomized scale which we have described in Sec. 5.4. For this reason CBS has not added these results for 2010 to the time series on happiness[16] and life satisfaction[17] and these results have also not been used for the research described in this thesis.

In 2012 Statistics Netherlands included the questions on happiness and life satisfaction in an experiment with a split-half design for the Social Cohesion Survey (Van Beuningen et al 2014). For this survey respondents were at random assigned to a group that had to rate the traditional 5-point verbal scales and a group that had to rate the same questions on a numerical 11-point response scales of which only the anchor points had verbal labels. The question on happiness with the 11-point numerical scale was phrased as: On a scale from 0 to 10 can you indicate to what extent you consider yourself to be a happy person? A score of 0 refers to being completely unhappy and a score of 10 refers to being completely happy. The question on life satisfaction with a numerical scale was worded as: On a scale from 0 to 10 can you indicate to what extent you are satisfied with the life you currently lead? A score of 0 refers to being completely dissatisfied and a score of 10 to being completely satisfied.

Data in the experiment were collected using a sequential mixed mode design. People were sent an invitation and two reminder letters asking them to fill out the questionnaire online. Those who did not respond to this invitation were interviewed by phone if a telephone number was available. When no telephone number was available people were interviewed face-to-face at their home.

The survey for the split-half experiment was distributed among respondents of 15 years or older, but for this thesis only respondents of 18 years or older were included in the analyses, which gave a total number of 7,641 respondents. Distinctions were made between gender, age, {18-24, 25-34, 35-44, 45-54, 55-64 and 65 or older}, and education, {low, middle and high}, to differentiate demographic categories of respondents. The frequency distributions for the two 5-point scales and the two 11-point

[16] Satisfaction with life: http://statline.cbs.nl/StatWeb/publication/?DM=SLEN&PA=60027ENG&D1=48-52&D2=1-2,12,15,19-22,29-31&D3=a&LA=EN&HDR=T&STB=G1,G2&VW=T

[17] Permanent Survey on Living Conditions

scales for each of these categories are included in Tab. D.1 to D.4 in appendix D.

### *9.2.2 Satisfaction with life: the Eurobarometer*

The EB is a series of public opinion surveys that have been conducted regularly in the member states of the European Union on behalf of the European Commission since 1973 as we have described in Sec. 1.3. Each EB survey consists of approximately 1,000 face-to-face interviews per participating country. We recall from Sec. 1.3 that in the autumn of 2011, prior to the standard survey, version 76.2 of the EB was launched, in which the item for life satisfaction had to be rated on a 10-point numerical scale, with the anchor points labelled 'Very dissatisfied' and 'Very satisfied' (European Commission, 2012b).

Since we are interested in time series for life satisfaction, we selected only countries which had participated in the European Union for more than 10 years. In the 15 European countries that joined the European Union before 2004, life satisfaction was measured for the entire period between 1997 and 2012 and also the item with the 10-point numerical scale was fielded in all these countries (Schmitt et al 2008; European Commission 2012a, 2012b and 2013). The number of respondents to the EB per country was too low to distinguish all the demographic categories we distinguished in the split-half experiment of CBS. Therefore, we divided the countries into two groups of countries:

- Northern Europe: Denmark, Sweden, Finland, Germany, Great Britain, The Netherlands, Belgium, France, Ireland, Luxemburg, Austria
- Southern Europe: Portugal, Spain, Italy and Greece

Then, for these two groups, we distinguished similar demographic categories to those used for the items of Statistics Netherlands. Only the three categories indicating the education level of a respondent are not fully comparable between the two datasets. We labelled these categories as 'low', 'middle' and 'high'. In the EB the education level classification is: in education until the age of 15, until the age between 16 and 19 years old and until an age of 20 years or over. The frequency distributions for the 5-point scale and the 11-point scale for each of these categories and the European regions can be found in Tab. E.1 to E.4 of appendix E.

It is interesting to compare the frequencies with which the upper anchor points have been ticked in The Netherlands on the numerical scale of Statistics Netherlands scale and in Northern Europe on the numerical Eurobarometer scale. This frequency is more than three times higher in Northern Europe than in The Netherlands. This difference is rather due to

the labelling of each of the anchor points than to a real difference in satisfaction: the wording ‘Completely satisfied’ used by Statistics Netherlands is likely to be interpreted differently from the wording ‘Very satisfied’ used in the Eurobarometer. Moreover the Northern Europeans ticked the anchor point more frequently than the preceding option labelled ‘9’. This phenomenon according to which the frequency with which respondents tick a ‘10’ stands out and is sometimes higher than the frequency by which a ‘9’ is ticked by respondents in the same sample is not unique as has been described by Brulé and Veenhoven (2014), who named this the ‘10-excess phenomenon’.

### 9.3 Reference boundaries in different demographic categories

People will interpret verbal response options differently for all kinds of reasons such as personality, cultural context, demographic characteristics or the context of the scale. This makes it unlikely that the reference boundaries for the response options of a given scale for the general population are equal to those for subcategories with specific characteristics.

To apply the Reference Distribution Method in an appropriate way, we had to adjust the verbal scale for happiness from Statistics Netherlands, because only a few of the respondents had ticked the option ‘Unhappy’ in the split-half experiment. For the application of the reference distribution method such response options have to be combined with the preceding or succeeding option to obtain proper reference boundaries. Except for the age category from 55-64 years, in all demographic categories less than one percent of the respondents selected this option as can be seen from Tab. D.3 of appendix D. As a consequence, the application of the Reference Distribution Method will return a reference boundary equal to zero for this option for respondents aged from 35-44 years. Since the frequency with which the option ‘Not very happy’ was chosen was also small in all demographic categories, we combined the options ‘Unhappy’ and ‘Not very happy’ for the application of the Reference Distribution Method.

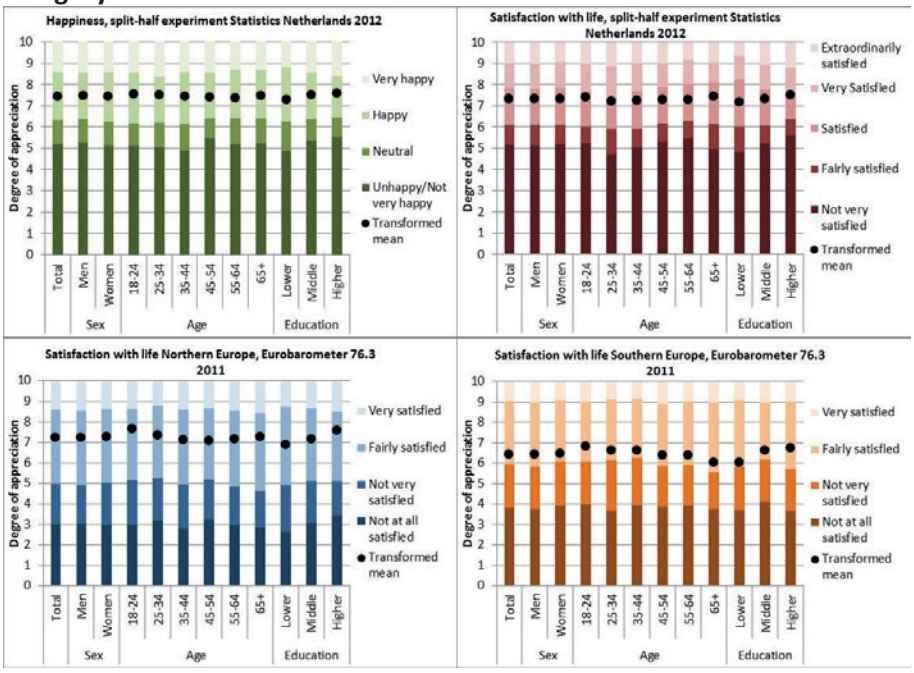
We applied the Reference Distribution Method to obtain reference boundaries for the verbal scales for life satisfaction of CBS and the EB and the adjusted verbal scale for happiness of CBS. For each demographic category and each topic, either happiness or life satisfaction, we first determined the best fitting beta distribution to the cumulative frequencies measured on the corresponding numerical scale from the split-half experiment of CBS or EB version 76.2 in the way we illustrated in Fig. 16 of Sec. 6.4. Next, we used these best fitting beta distributions to derive the

reference boundaries between the response options of the verbal scales for each demographic category in the way we demonstrated in Fig. 18 of Sec. 7.2 based on the related cumulative frequency distribution measured in the split-half experiment of Statistics Netherlands or in version 76.3 of the EB.

The results are displayed in Fig. 27 in which the response scales are visualized as vertical bars in which each interval represents the degree of appreciation expressed by a response option and in which the reference boundaries separate two adjacent response options. The estimated mean for a demographic category d in Fig. 27 is, in accordance with formula (3) in appendix C equal to $10*\alpha_d/(\alpha_d + \beta_d)$, with $\alpha_d$ and $\beta_d$ the shape parameters of the beta distribution that fits best to the reference boundaries and the cumulative frequency distribution for the given demographic category.

From Fig. 27 it is obvious that differences in reference boundaries between demographic categories and the general population are most prominent for age categories and for categories based on the education level of the respondents. Apparently, gender does not make much of a difference, not in the split-half experiment or in the EB.

**Figure 27 Reference boundaries and estimated means per demographic category**



The levels of the estimated means for EB version 76.3 cannot be compared with those for life satisfaction in the split-half experiment, because the reference distribution in the split-half experiment was based on an 11-point numerical scale, whereas in the EB a 10-point numerical scale was used[18]. For this reason, the estimated means for the split-half experiment are visibly higher than those for the Eurobarometer. In addition, the anchor points of both numerical scales for life satisfaction were labelled differently.

The patterns of estimated means for the EB are similar for Northern and Southern Europe, except that there is a substantial difference in the level: life satisfaction in Southern Europe is, in all demographic categories, lower than in Northern Europe. This comes with noticeably differences in reference boundaries between the two clusters of European countries.

As we stated above, the differences in boundaries can be attributed to a number of reasons. For the differences between age categories, it is imaginable that for example the response time of younger people is shorter than that of older people, whereas persons belonging to the latter category have a longer life experience than those who are younger. Both aspects may, among others, have an effect on the response patterns. The reasons for differences in boundaries however, fall outside the scope of this thesis and have been addressed by others, see for example Hazelrigg and Hardy (2000), Storm et al (1996), and the National Research Council (2013). We recall that the research question addressed in this chapter is whether the reference boundaries derived for the general population can be used for demographic categories to produce reliable extended time series to monitor differences in trends among these categories.

## 9.4 Differences in estimated means

A first step to answer the research question is to look at the differences in estimated means for demographic categories between the situation in which the reference boundaries for the general population are used and the situation in which category-specific boundaries are used. In Tab. 24 and 25 the estimated means as presented in Fig. 27 are given in the columns headed by 'Reference boundaries category'. The estimated means based on the best fitting beta distributions to the reference boundaries for the general population and the cumulative frequencies measured in each demographic category are given in the columns headed 'Reference boundaries population'.

[18] The boundaries between the response options of a 10-point numerical scale are positioned at different values than those between the response options of an 11-point numerical scale.

**Table 24 Differences in estimated means depending on reference boundaries used, Statistics Netherlands 2012**

| Split-half experiment Statistics Netherlands 2012 | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Happiness | | | Satisfaction with life | | |
| | | Estimated means | | | Estimated means | | |
| Category | Sub-category | Reference boundaries category | Reference boundaries population | Difference | Reference boundaries category | Reference boundaries population | Difference |
| Total | Total | 7.45 | 7.45 | | 7.32 | 7.32 | |
| Sex | Men | 7.46 | 7.44 | 0.02 | 7.32 | 7.34 | -0.02 |
| | Women | 7.46 | 7.47 | -0.01 | 7.32 | 7.30 | 0.02 |
| Age | 18-24 | 7.57 | 7.63 | -0.06 | 7.40 | 7.47 | -0.07 |
| | 25-34 | 7.53 | 7.71 | -0.18 | 7.22 | 7.51 | -0.29 |
| | 35-44 | 7.44 | 7.52 | -0.08 | 7.27 | 7.37 | -0.10 |
| | 45-54 | 7.40 | 7.36 | 0.04 | 7.30 | 7.23 | 0.07 |
| | 55-64 | 7.38 | 7.26 | 0.12 | 7.30 | 7.09 | 0.21 |
| | 65+ | 7.48 | 7.36 | 0.12 | 7.42 | 7.30 | 0.12 |
| Edu-cation | Low | 7.28 | 7.23 | 0.05 | 7.17 | 7.03 | 0.14 |
| | Middle | 7.51 | 7.49 | 0.02 | 7.31 | 7.35 | -0.04 |
| | High | 7.60 | 7.68 | -0.08 | 7.52 | 7.56 | -0.04 |

**Table 25 Differences in estimated means depending on reference boundaries used, Eurobarometer 2011**

| Eurobarometer 2011 version 76.3 | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Satisfaction with life Northern Europe | | | Satisfaction with life Southern Europe | | |
| | | Estimated means | | | Estimated means | | |
| Category | Sub-category | Reference boundaries category | Reference boundaries population | Difference | Reference boundaries category | Reference boundaries population | Difference |
| Total | Total | 7.26 | 7.26 | | 6.46 | 6.46 | |
| Sex | Men | 7.24 | 7.27 | -0.03 | 6.43 | 6.54 | -0.11 |
| | Women | 7.28 | 7.24 | 0.04 | 6.49 | 6.39 | 0.10 |
| Age | 18-24 | 7.66 | 7.60 | 0.06 | 6.84 | 6.78 | 0.06 |
| | 25-34 | 7.34 | 7.14 | 0.20 | 6.63 | 6.53 | 0.10 |
| | 35-44 | 7.13 | 7.17 | -0.04 | 6.64 | 6.41 | 0.23 |
| | 45-54 | 7.09 | 6.97 | 0.12 | 6.41 | 6.49 | -0.08 |
| | 55-64 | 7.18 | 7.26 | -0.08 | 6.42 | 6.44 | -0.02 |
| | 65+ | 7.28 | 7.47 | -0.19 | 6.05 | 6.29 | -0.24 |
| Edu-cation | Low | 6.91 | 6.90 | 0.01 | 6.07 | 6.15 | -0.08 |
| | Middle | 7.16 | 7.06 | 0.10 | 6.66 | 6.54 | 0.12 |
| | High | 7.59 | 7.61 | -0.02 | 6.78 | 6.92 | -0.14 |

What is clear from Tab. 24 and 25 is that the differences between the estimated means obtained using the category-specific reference boundaries or by using the reference boundaries for the general population are small in general. Nevertheless, at least for some categories, there are differences. The difference is largest for satisfaction with life in the age category from 25-34 years of the split-half experiment. Using the reference boundaries for this category to estimate a mean for the best fitting beta distribution, would lead to a value which is almost 0.3 points lower than if the reference boundaries for the general population are used.

It would be premature to conclude on basis of Tab. 24 and 25 that category-specific reference boundaries should be used to estimate a mean for the demographic category. It very much depends on the purpose of the transition to the 0 to 10 continuum. If one is interested in the absolute difference in happiness or life satisfaction between demographic categories on a continuum from 0 to 10, than it seems evident to use category-specific reference boundaries. This is not obvious though, if one is interested, as we are given our research question, in whether or not trends in happiness and life satisfaction evolve in the same way in different demographic categories. In this case the absolute difference in happiness or life satisfaction is of less importance as long as the development of the trend in each category shows a reliable pattern after transition to the 0 to 10 continuum. From this perspective, to be able to answer the research question requires insight into the stability over time of the differences between the estimated means on the continuum for each demographic category obtained using the category-specific boundaries or the boundaries for the general population.

## 9.5 Trends in estimated means in different demographic categories

The evolution of happiness and life satisfaction over time may differ across demographic categories. These differences in evolution have to be preserved after estimation of the corresponding means on a continuum from 0 to 10, independent of whether reference boundaries for the population of a whole are used or category-specific boundaries. If the latter is the case, then it suffices to apply the reference boundaries for the general population to obtain estimated means on a 0 to 10 continuum for monitoring differences in trends in extended time series.

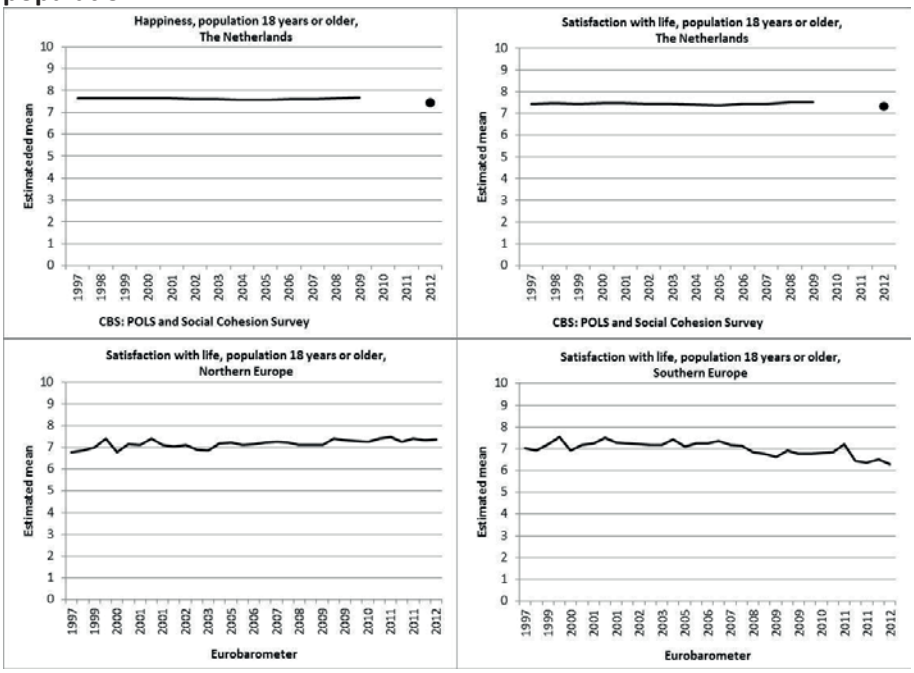
### *9.5.1 Trends in estimated means over time for the general population*

We applied the Continuum Approach to the response of the general population to the items on happiness and life satisfaction for each wave of the surveys of CBS in the period 1997-2009 and to the item on life

satisfaction for each wave of the EB in the period 1997-2012. We made use of the reference boundaries for the general population for each wave, as shown in Fig. 27. In these periods there were no changes in the mode of surveying for both surveys, see also Sec. 10.3. Because of that, the reference boundaries can be kept fixed over time and the differences in estimated means can solely be attributed to changes in the frequency distributions on the primary scale as we stated in Sec. 8.6.

The application of the Continuum Approach resulted in a best fitting beta distribution for each item and each wave in the corresponding period. We used the estimated parameters of each of these beta distributions to estimate a population mean, the results of which are depicted in Fig. 28.

**Figure 28 Estimated means happiness and life satisfaction, general population**




It is clear from Fig. 28 that there has been little variation in the levels of happiness and life satisfaction over the years in The Netherlands. The average level of life satisfaction in The Netherlands is somewhat lower than that for happiness. In 2012 the levels for both items of CBS dropped. Since no appropriate data were available for 2010 and 2011, it cannot be said

whether this is due to a different measurement techniques, or that it is an effect of the economic crisis or that it has some other cause.

Compared to the flat lines for happiness and life satisfaction based on the surveys of CBS, the time series of estimated Eurobarometer means show a much more fluctuating pattern. Moreover, the time series for Northern Europe shows a slightly upward trend, whereas in Southern Europe the trend is heading downwards.

### *9.5.2 Differences in estimated means over time depending on the boundaries used*

Just as we did for the general population, we applied the Continuum Approach to the responses for each demographic category for each wave of each item. For each wave we applied the Continuum Approach twice, one time making use of the reference boundaries for the general population and one time making use of the category-specific reference boundaries which are shown in Fig. 27.

As a result we found two best fitting beta distributions for each item and each wave, for which we, just as before, could use the parameters to estimate means on a 0-10 continuum for the corresponding demographic category. As was to be expected and in line with what we presented in Tab. 24 and 25, the estimated means based on the category-specific reference boundaries are not equal to the estimated means based on the reference boundaries for the general population. The differences between the estimated means based on the category-specific boundaries and those based on the boundaries for the entire population however, turn out to be very stable over time. We have summarized this in Tab. 26 and 27, which contains two columns for each item. The first of these columns contains the average difference over the number of waves in the period of observation between the estimated means. The second column contains the standard deviation of these differences from the average difference.

The averages in Tab. 26 and 27 are similar to those presented in Tab. 24 and 25. The standard deviations are very small for each demographic category and each item indicating that the differences are rather stable over time. The, in the absolute sense, largest average difference is once more found in the age category 25-34 years for the CBS item on life satisfaction. A possible explanation for finding the largest difference for this particular category may be that at this stage in their lives individuals in this category must deal with numerous changes, many of which may affect their response to surveys. We have depicted the differences for this demographic category in Fig. 29.

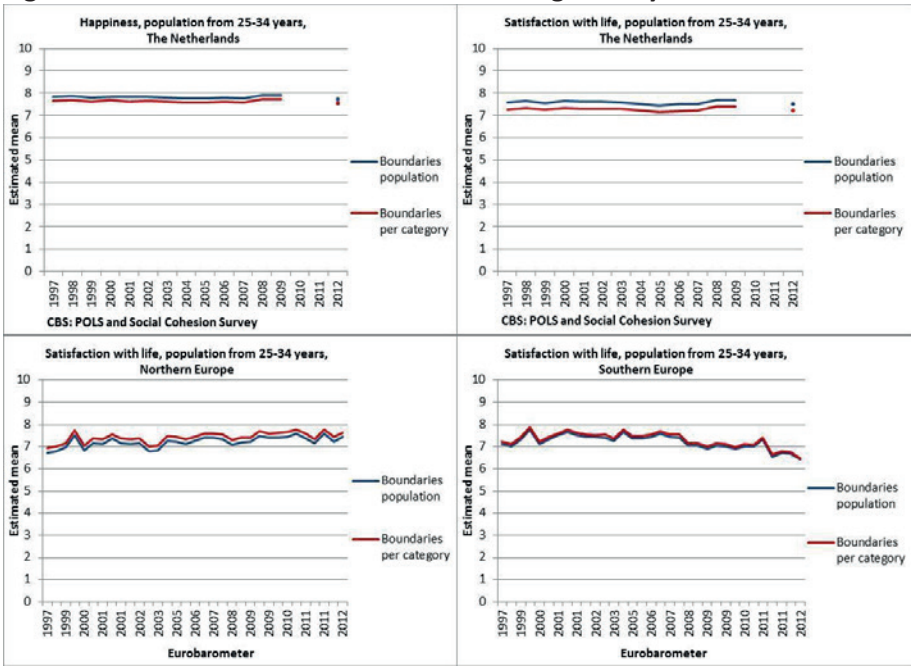
**Table 26 Average and standard deviation difference in estimated mean, Statistics Netherlands**

| CBS: POLS and Social Cohesion Survey, 1997-2009, 13 waves | | | | | | |
|---|---|---|---|---|---|---|
| | | Happiness | | | Satisfaction with life | |
| | Sub- | Difference in estimated mean | | | Difference in estimated mean | |
| Category | category | Average | Standard deviation | | Average | Standard deviation |
| Sex | Men | 0.02 | 0.01 | | -0.02 | 0.01 |
| | Women | -0.02 | 0.01 | | 0.02 | 0.01 |
| Age | 18-24 | -0.07 | 0.01 | | -0.08 | 0.01 |
| | 25-34 | -0.19 | 0.01 | | -0.29 | 0.01 |
| | 35-44 | -0.07 | 0.01 | | -0.11 | 0.01 |
| | 45-54 | 0.02 | 0.01 | | 0.06 | 0.01 |
| | 55-64 | 0.11 | < 0.01 | | 0.20 | 0.01 |
| | 65+ | 0.11 | 0.01 | | 0.12 | 0.01 |
| Education | Low | 0.12 | 0.01 | | 0.23 | 0.01 |
| | Middle | 0.01 | 0.01 | | -0.07 | 0.01 |
| | High | -0.08 | 0.01 | | -0.08 | 0.02 |

**Table 27 Average and standard deviation difference in estimated mean, Eurobarometer**

| Eurobarometer 1997-2012, 33 waves | | | | | | |
|---|---|---|---|---|---|---|
| | | Satisfaction with life Northern Europe | | | Satisfaction with life Southern Europe | |
| | | Difference in estimated mean | | | Difference in estimated mean | |
| Category | Sub-category | Average | Standard deviation | | Average | Standard deviation |
| Sex | Men | -0.04 | < 0.01 | | -0.10 | < 0.01 |
| | Women | 0.03 | < 0.01 | | 0.09 | < 0.01 |
| Age | 18-24 | 0.06 | 0.01 | | 0.04 | 0.01 |
| | 25-34 | 0.21 | 0.01 | | 0.10 | 0.01 |
| | 35-44 | -0.04 | < 0.01 | | 0.22 | 0.01 |
| | 45-54 | 0.12 | 0.01 | | -0.09 | 0.01 |
| | 55-64 | -0.08 | < 0.01 | | -0.02 | < 0.01 |
| | 65+ | -0.20 | 0.01 | | -0.21 | 0.01 |
| Education | Low | 0.01 | 0.01 | | -0.06 | 0.01 |
| | Middle | 0.10 | < 0.01 | | 0.08 | 0.02 |
| | High | -0.01 | 0.01 | | -0.12 | 0.01 |

**Fig. 29 Time series of estimated means among 25-34 years old**




From Fig. 29 it can be seen that for the two CBS items, the estimated means based on the boundaries for the general population are higher than the estimated means based on the category- specific boundaries, and that this is the opposite for the EB item in both Northern Europe and Southern Europe. The patterns of the time series are, however, consistent from which we conclude that if the main interest of a piece of research, as here, is to determine how trends develop then which reference boundaries are used for the conversion becomes less important.

The results shown in Tab. 26 and 27, supported by the visualization in Fig. 29, lead us to conclude that the reference boundaries define the levels of the estimated population means, but do not influence the evolution of the trends. This conclusion makes sense for the flat patterns of the time series for happiness and life satisfaction of CBS, and for the fluctuation patterns for life satisfaction of the EB the trend for which go in a different direction in Northern Europe and Southern Europe.

## 9.6 Discussion

Using three different populations, The Netherlands, Northern Europe and Southern Europe, and items from two surveys, we showed that, from the

perspective of monitoring trends in happiness and life satisfaction among demographic categories, it is not necessary to use category-specific reference boundaries to obtain estimated means on a 0 to 10 continuum if the Continuum Approach is applied. Although we covered some variants for which this conclusion is valid, it would be worthwhile applying the Reference Distribution Method presented in this thesis to other topics than happiness and life satisfaction, for example to social cohesion or self-reported health. This would allow us to validate the conclusions presented here more generally.

In Sec. 9.3 we remarked that the levels of the estimated means for the EB cannot be compared with those for life satisfaction from the split-half experiment of CBS, due to the different number of response options on the numerical scales used in the surveys. Since we concluded that it is not necessary to use category-specific boundaries to make the transition to the estimated means on a 0 to 10 continuum, the numerical scale for life satisfaction from the split-half experiment could be used to derive reference boundaries between the response options of the verbal scale of the EB item based on the response of the general population measured in 2012. These boundaries in turn could be used to apply the Continuum Approach to the Eurobarometer time series of the different social categories in Northern Europe and Southern Europe. This would enable comparison of trends in life satisfaction among demographic categories in The Netherlands with equivalent categories in Northern Europe and Southern Europe.

The numerical scale from the EB was only used in 2011, and as there was no wave of the survey of CBS in this time, this numerical scale from the EB cannot be used to apply the Reference Distribution Method to derive reference boundaries for the verbal scale of the life satisfaction item of CBS. There are however, other surveys, such as the European Social Survey, that contain an item on life satisfaction with a numerical scale and that have been fielded over a number of years. The response to such a survey item in a year when both the EB and the survey of CBS were fielded could be used substituted to estimate a reference distribution. This reference distribution could then be used accordingly to derive reference boundaries for the general population for the verbal scales of the EB item and the CBS item.

In this chapter we distinguished between gender, age, and education to differentiate demographic categories of respondents. Another distinction that might have been of interest is that of ethnicity. The conversion of verbal terms into numerical scales is likely to vary between

people from a different cultural and linguistic background. This is underpinned by our findings in Ch. 4. The data used for this thesis, however, fell short to make the distinction to ethnical groups, at least for The Netherlands. We will leave this therefore as a potential direction for future research.

For the trend analyses in this chapter we have only focused on mean happiness within demographic categories. We are aware, however, that in practice scholars are often not interested in just the mean, but want to know what the interrelations of the means in subgroups of these demographic groups are, for example from the perspective of inequality or marginalization. Although this falls outside the scope of this thesis, we have spent some words on the technique to determine means for subgroups based on a given beta distribution in appendix F.

**9.7 Conclusion**

In this chapter we addressed the question of whether the transition points derived for the general population can be used for demographic categories to produce reliable extended time series to monitor differences in trends among these categories. We conclude that this is possible and that it is not necessary to derive transition points for each demographic category separately.

# 10 Pooling time series based on slightly different questions about the same topic

## 10.1 Research question

We started in Ch. 1 with the problem of incomparability of time series from different surveys. We illustrated what the effect is for time series on survey questions about happiness and life satisfaction if the means for both topics are based on just the ranks of response options of the primary scale in Fig. 1 and 2 of Sec. 1.4. In the intermediate sections we have discussed a number of conventional and more recent approaches to tackle this problem of incomparability of which the Reference Distribution Method comes closest to overcoming the problem.

The main question we address in this section is: Is the Reference Distribution suitable to use for pooling time series based on slightly different survey questions about the same topic? We investigated the suitability of the method using the Reference Distribution Method to pool the time series for happiness and life satisfaction in The Netherlands described in Ch. 1. As a by-product of this investigation we answered the question: Have the Dutch become more happy and satisfied with their lives over the past 4 decennia?

## 10.2 Distortion of trends due to biases in measurement

A number of sources of measurement bias can be determined which may affect a respondent's response to survey items and muddle the trend analyses by causing discontinuities. Some of these biases can be reduced using the Reference Distribution Method. Therefore a thorough inspection of a time series of data is necessary to determine possible biases that may occur and which have to be addressed before any application of the method to obtain a transformed time series.

In Ch. 5 we introduced the issue of scale effects arising from differences in the numbers of response options, the labelling of the response options and whether a scale is numerical or verbal, however, something we did not mention above, is that the visual presentation of a scale can also influence results. This set of scale effects provided the main motivation and angle of approach for the development of the Reference Distribution Method and the preceding scale transformation methods commonly used in social research. All the methods used to date are aimed at overcoming the incomparability of survey results arousing from the scale effects. Further, if these methods are going to be used to transform time

series to perform analyses of trends over a long time span, than a number of other sources of response bias must be addressed. Here we give an overview of these response biases in relation to the Reference Distribution Method. This overview is based on the description of these issues to be found in the book of the National Research Council (NRC) (2013, Ch. 4). In this book the sources of response bias we will deal with are, besides that of scale effects, mode of surveying, ordering of questions, response shift and contextual influences.

### *10.2.1 Survey mode effects*

Surveys can be conducted in a number of modes, such as face-to-face, by telephone, paper-and-pencil surveying, via the internet or a mixture of the modes. Which mode is chosen, depends amongst others on the goal of the survey, the costs of surveying or the size of the sample, and, something which is of importance in relation to time series, the time when, historically, the survey was conducted. Telephone interviews, for example, only became possible at the time it was reasonable that almost all of the respondents would have a telephone at home, however, the rise of mobile phone usage has come at the cost of the adequate registration of telephone numbers, this causing this mode of questioning to become statistically invalid. The internet has opened up new opportunities for interviewing and its use is becoming ever more popular for survey research, not in the least for the relatively low costs associated with this mode of surveying.

The mode of surveying used has a large impact on a respondent's response. We already mentioned the mode change from computer assisted personal interviews to a paper-and-pencil questionnaire of the Life Situation Survey of SCP in Sec. 7.3, which caused a dramatic fall in the percentage of people who rated themselves as either 'Happy' or 'Very happy'[19]. Another example of the mode effect is described by Dolan and Kavetsos (2012), who found large differences in survey mode which swamped all the other bias effects that they had taken into account in their analysis.

[19] This change may be attributed to those responding to a paper-and- pencil survey having more time to think about their answer than someone doing a face-to-face interview, where the pace is set by the interviewer. This leads to a more cognitive than intuitive response which comes with lower reported happiness (Studer and Winkelmann,2014). Another explanation for the change from 2002 to 2004 is the commonly recognized interviewer bias, caused by the effect an interviewer can have on the response in a face-to-face interview (Katz, 1942; Davis et al, 2010). The interviewer may, unintentionally, influence respondents to give for example socially desirable answers that may be more positive than when that respondent answers questions in an uninfluenced situation like a paper-and-pencil survey.

We recall from Sec. 7.3 that as long as the mode of surveying is unchanged, the application of the Continuum Approach results for each wave in an estimated beta distribution that tightly fits to the reference boundaries derived from a reference distribution and the cumulative distribution of the primary scale. If the mode of surveying does not change, the reference boundaries can thus be kept fixed over time and the differences in estimated means can solely be attributed to changes in the frequency distributions on the primary scale.

### *10.2.2 Ordering of questions*

The NRC (2013) mentions a number of studies in which the effect question ordering has on the response to a survey is discussed. They refer among others to a paper of Deaton (2012), who found a large effect of question ordering on the responses to a question on subjective life satisfaction. How a question was answered, was affected by, for example, whether or not respondents had to answer a question about a subject such as politics or unemployment before they assessed their own satisfaction with life. More in general it is acknowledged in the literature that preceding questions may affect how respondents interpret the meaning of survey items and that this effect may be significantly large (Saris and Gallhofer, 2007; OECD, 2013).

The bias in trend analyses caused by the effect the question ordering has on the survey response cannot be solved using the Reference Distribution Method if the order of questions is changed in different waves of one and the same survey, unless the size of the effect is comparable to that of a mode change and can be addressed as such. The Reference Distribution Method however, can deal with a difference in question ordering if this difference is present *between* surveys, but not in succeeding waves of each survey separately. This is because in the Reference Distribution Method the sample means of different surveys are forced to a level that is equivalent to that of the reference distribution, this causing the potential effect of a difference in question ordering between surveys to vanish.

### *10.2.3 Response shift*

The term ‘response shift’ originates from a study by Howard et al (1979) conducted to study to which extent participants changed their level of self-perceived dogmatism during a communication skills training. The effect of the training made the participants change their internal standards for their measurement of dogmatism. Researchers introduced the term ‘response shift’ to denote this kind of change. Since this time a growing number of

studies on response shift have been conducted, most in the field of health care (e.g. Schwartz et al, 2013).

Within the field of subjective well-being the NRC (2013) summarizes response shift as a term that is used to characterize changes in reporting over time. They differentiate the observed differences over time in self-reports of well-being into two main streams. One reflects true changes in the quality-of-life assessments made by respondents as a result of adaptation to changed circumstances. The second reflects measurement errors associated with a respondent's internal scale recalibrations. As an example of this type of internal scale recalibration given by the NRC is the case described of a person with chronic pain who rates this pain on average at 7 on a scale from 1 to 10. After having experienced the more intense pain of a kidney stone, the same person lowered their rating of the unchanged chronic pain at 5. Thus, this person made an internal scale recalibration of how they rate chronic pain.

The NRC concludes that adaptation cannot be characterized as a process that occurs uniformly, because every person is unique and adapts differently to changing situations. In analyzing time series however, it is of interest to monitor if there are general upwards or downwards trends in target groups or subgroups, and in subjective well-being trends which reflect true changes in quality-of-life, but, as described by the NRC, it is not yet possible to decompose these trends into what are 'true' changes in quality-of-life and effects of scale recalibration. Use of the Reference Distribution Method does not contribute to solving this entanglement. It is very well possible that over the course of time a slow change in scale interpretation has evolved which means that on average the present population will appreciate the response options in a response scale differently than that of earlier times, which can also be considered as a type of scale recalibration. For an application of the Reference Distribution Method, in which we try to bring sample means from different surveys to a comparable level, we assume that if at present response options are appreciated differently than in the past, this will be the case for all the response scales under consideration, verbal and numerical.

### *10.2.4 Contextual influences*

Contextual influences refer to the research instruments used and the external circumstances at the time of a respondent's assessment for a self-report. Contextual influences of the research instrument can be brought back to the mode of surveying used and the ordering of questions as discussed above. There are manifold influential external circumstances that

one can think of, such as the weather conditions at the time of the assessment, recent life events or breaking news of the day, but also the demographic characteristics of a respondent, such as age, level of education or cultural background and whether a questionnaire is computer mediated or not.

When analyzing trends one has to be aware of these contextual influences. This holds especially when studying trends over a long period of time as a consequence of which the sample of respondents in the earlier waves of the time series represent a different population than the sample of respondents in the more recent waves. In an ageing society for example, an increasing trend in life satisfaction might be due to a positivity effect of an age-related trend of the older person having a positive view towards life (NRC, 2013). In another example, the composition of the population may change over the years due to migration movements, introducing an effect of cultural influences into the outcome of happiness measurements (Senik, 2013). The above examples serve to illustrate how contextual influences in general affect trend analyses, independent of the response scales used. The Reference Distribution Method can be used appropriately to take a limited number of the possible contextual influences into account. If, for example, a survey is always conducted in the autumn or spring, than there may be a seasonal effect in the response to each wave, also, when an exceptional influential event takes places during the period of surveying the results may be biased. These types of contextual influences are, to a certain extent, corrected when the Reference Distribution Method is used, because the sample means of the surveys they belong to are shifted to a level that is equivalent to that of the reference distribution.

### 10.3 Inspecting the available time series

We have argued that when using the Reference Distribution Method it is possible to correct for discontinuities due to mode changes and, to a limited extent, for contextual influences. This requires a thorough inspection of the available time series, and decisions have to be made as to which waves require new reference boundaries to be derived and which estimated beta distributions can serve as a reference for these decisions. In addition we have to decide whether some waves have to be excluded from a transformation and for what valid reason.

### *10.3.1 Searching for sources for response bias to correct*

***The CBS items on happiness and life satisfaction***

The surveys which include the CBS items on happiness and life satisfaction have been changed several times since their introduction. These changes have affected the responses and led to serious discontinuities in the time series. In the first years the survey containing the items was only fielded over a period of a few weeks, changing to a continuous survey from 1989 on. The effect of this is undoubtedly visible, especially for the life satisfaction item in Fig. 2 of Sec. 1.3. A comprehensive revision of the questionnaire forms and a reduction in the number of survey items in several domains in 1994 led once more to a response discontinuity. Another major change in the CBS survey took place in 1997 when among other things the survey mode was changed from paper-and-pencil surveying into face-to-face interviews. This change was introduced using a split-half measurement with half of the respondents being required to fill in a paper-and-pencil questionnaire and the other half being interviewed. As we mentioned in Sec. 9.2.1 CBS changed to a mix-mode of surveying in 2010 which caused a drop by approximately 5 percent in the percentage of happy and satisfied persons. For that reason we left the results for 2010 out of the conversion process. In the most recent change to the CBS survey, in 2012, 10-point numerical scales were introduced, breaking with the previous tradition of using verbal scales (Van Beuningen et al, 2014). Thus we left the 2012 wave of CBS out of the conversion process, since as yet it provides only a singular point in the data flow.

***The SCP items on happiness and life satisfaction***

Much of which has been said for the CBS items also holds for the SCP items. In addition, as we mentioned in Sec. 7.3 and Sec. 10.2.1, SCP changed the mode of its Life Situation Survey, which included the item on happiness, from a computer assisted personal interview to a paper-and-pencil questionnaire in 2004. With the exception of the Life Situation Survey, SCP has included the life satisfaction item also over a long period in its survey of Cultural Changes in The Netherlands. The response to this survey is yet un-weighted and we therefore left it out of the conversion process. Note: in 2004, for rating life satisfaction SCP changed from using its traditional 5-point verbal scale to a 10-point numerical scale.

***The EB items on happiness and life satisfaction***

Happiness was measured in the first years of the EB until 1986 and not in later years. In the waves of the EB from before 2002 the questions on life satisfaction and before 1987 also on happiness, were preceded by a number of opinion questions on different topics. In almost all waves of the EB from 2002 on, the question on life satisfaction is the first subjective question asked. We have seen no noticeable change due to this difference in question ordering between older and newer waves. The main point of concern with respect to the EB is the fact that there is a spring and a summer wave in each year and in some years one or more extra waves. This may be the reason why the individual waves of the time series of the EB show a rather irregular patter, which may be due to seasonal effects. For this reason, but also because we preferred to have one measurement a year, we averaged the response to all waves per year for the EB.

***The DHS item on happiness***

In the time series of the DHS item on happiness there are noticeable discontinuities in the waves of 2000 and 2006. We have no information about the background of these discontinuities and therefore decided to leave them as they are.

***The WVS items on happiness and life satisfaction***

The WVS items have been fielded with large gaps in the major series. Despite this, we had no reason to assume that there is a need to correct for discontinuities, except for the item on happiness for the wave of 2008. Compared to the wave of 2006, the frequency of respondents who rated themselves as very happy increased from 42 percent to 56 percent. We do not see such a large increase in any of the other surveys carried out in the same period and therefore considered this result of the WVS for 2008 to be implausible, and left it out of the conversion process.

### *10.3.2 Preparation for the conversion of time series*

The ESS is the only recent survey available in which both happiness and life satisfaction are measured on an 11-point numerical scale. The ESS is therefore an obvious choice to select when estimating initial reference distributions. The ESS has been fielded every two years since 2002. We estimated a best fitting beta distribution to the frequency distributions of each wave of the ESS, keeping the numerical response options equidistant. In this way we got a conversion of the time series from the ESS into a series of continuous distributions which fit to the 0 to 10 continuum.

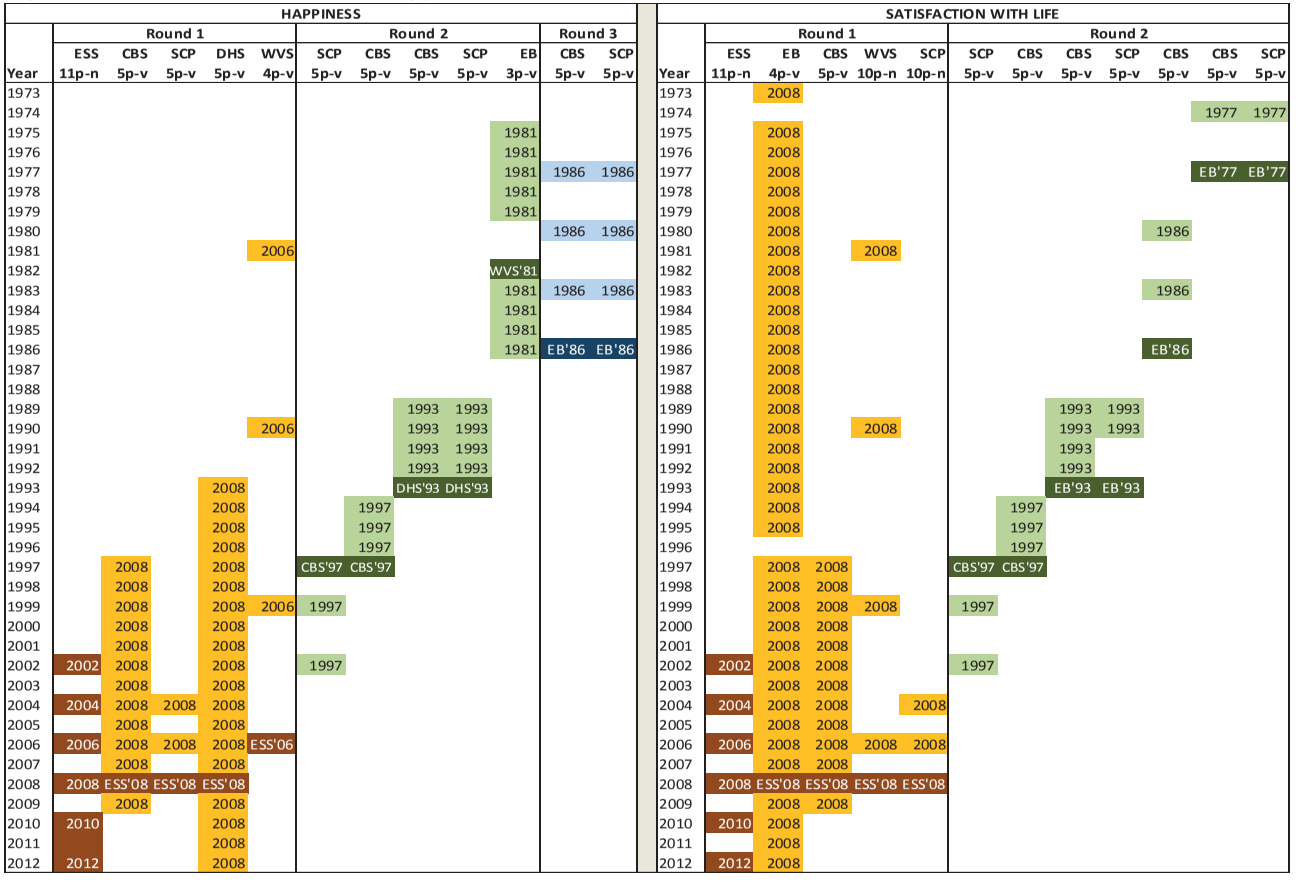
Because most of the time series from other surveys we have at our disposal, contain a wave for 2008, this became the best choice to serve as the initial reference year. Starting with 2008, we began the conversion process for the other surveys to convert the responses from waves of 2008 and before, but we also converted the responses from more recent waves. The exception to this was the happiness item of the WVS, for which we derived reference boundaries from the beta distribution estimated for the ESS-results over 2006, due to having excluded the WVS over 2008 from the conversion process for the reasons discussed above.

We have made conversion schemes for the time series for happiness and life satisfaction to prepare and support the conversion process. These schemes are shown in Fig. 30 in which the waves for the ESS items are indicated by the corresponding year on the left side of each conversion scheme. It can be seen in Fig. 30 that several rounds were needed for the conversion process. This is a consequence of the discontinuities in the time series of some surveys and of the fact that not all surveys were in use in the entire period spanned by all the time series together. The dark coloured boxes with text in white in Fig. 30 represent a year and survey item for which a beta distribution is estimated which is used to derive reference boundaries from for other survey items. The reference boundaries derived for each of those other surveys on their turn are used to apply the Continuum Approach for estimating a best fitting beta distribution for each of the other waves in the corresponding column. This is indicated, for example, for the waves of the CBS life satisfaction item in the period 1997 - 2009 by the text '2008' in Fig. 30.

From Fig. 30 it can be seen that we needed to derive reference boundaries for a number of items and a number of years. The reference boundaries for the 1982 wave of the EB item were derived from the estimated beta distribution for the 1981 wave of the WVS, which means that there was a one year difference between the waves from both surveys. This was a choice we made, because otherwise we would not have been able to convert the EB time series and the results from the older waves of the CBS and SCP time series.

We have presented the reference boundaries we derived in each round of the conversion process for happiness in Tab. 28 and similarly, the reference boundaries for life satisfaction are contained in Tab. 29. We have included the parameters $\alpha$ and $\beta$ and the estimated population mean for each reference distribution used.

## Figure 30 Conversion schemes happiness and life satisfaction



**HAPPINESS**

| Year | Round 1 | | | | | Round 2 | | | | | Round 3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ESS 11p-n | CBS 5p-v | SCP 5p-v | DHS 5p-v | WVS 4p-v | SCP 5p-v | CBS 5p-v | CBS 5p-v | SCP 5p-v | EB 3p-v | CBS 5p-v | SCP 5p-v |
| 1973 | | | | | | | | | | | | |
| 1974 | | | | | | | | | | | | |
| 1975 | | | | | | | | | | 1981 | | |
| 1976 | | | | | | | | | | 1981 | | |
| 1977 | | | | | | | | | | 1981 | 1986 | 1986 |
| 1978 | | | | | | | | | | 1981 | | |
| 1979 | | | | | | | | | | 1981 | | |
| 1980 | | | | | | | | | | | 1986 | 1986 |
| 1981 | | | | | 2006 | | | | | | | |
| 1982 | | | | | | | | | | WVS'81 | | |
| 1983 | | | | | | | | | | 1981 | 1986 | 1986 |
| 1984 | | | | | | | | | | 1981 | | |
| 1985 | | | | | | | | | | 1981 | | |
| 1986 | | | | | | | | | | 1981 | EB'86 | EB'86 |
| 1987 | | | | | | | | | | | | |
| 1988 | | | | | | | | | | | | |
| 1989 | | | | | | | | 1993 | 1993 | | | |
| 1990 | | | | | 2006 | | | 1993 | 1993 | | | |
| 1991 | | | | | | | | 1993 | 1993 | | | |
| 1992 | | | | | | | | 1993 | 1993 | | | |
| 1993 | | | | 2008 | | | | DHS'93 | DHS'93 | | | |
| 1994 | | | | 2008 | | | 1997 | | | | | |
| 1995 | | | | 2008 | | | 1997 | | | | | |
| 1996 | | | | 2008 | | | 1997 | | | | | |
| 1997 | | 2008 | | 2008 | | CBS'97 | CBS'97 | | | | | |
| 1998 | | 2008 | | 2008 | | | | | | | | |
| 1999 | | 2008 | | 2008 | 2006 | 1997 | | | | | | |
| 2000 | | 2008 | | 2008 | | | | | | | | |
| 2001 | | 2008 | | 2008 | | | | | | | | |
| 2002 | 2002 | 2008 | | 2008 | | 1997 | | | | | | |
| 2003 | | 2008 | | 2008 | | | | | | | | |
| 2004 | 2004 | 2008 | 2008 | 2008 | | | | | | | | |
| 2005 | | 2008 | | 2008 | | | | | | | | |
| 2006 | 2006 | 2008 | 2008 | 2008 | ESS'06 | | | | | | | |
| 2007 | | 2008 | | 2008 | | | | | | | | |
| 2008 | 2008 | ESS'08 | ESS'08 | ESS'08 | | | | | | | | |
| 2009 | | 2008 | | 2008 | | | | | | | | |
| 2010 | 2010 | | | 2008 | | | | | | | | |
| 2011 | | | | 2008 | | | | | | | | |
| 2012 | 2012 | | | 2008 | | | | | | | | |

**SATISFACTION WITH LIFE**

| Year | Round 1 | | | | | Round 2 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ESS 11p-n | EB 4p-v | CBS 5p-v | WVS 10p-n | SCP 10p-n | SCP 5p-v | CBS 5p-v | CBS 5p-v | SCP 5p-v | CBS 5p-v | CBS 5p-v | SCP 5p-v |
| 1973 | | 2008 | | | | | | | | | | |
| 1974 | | | | | | | | | | | 1977 | 1977 |
| 1975 | | 2008 | | | | | | | | | | |
| 1976 | | 2008 | | | | | | | | | | |
| 1977 | | 2008 | | | | | | | | | EB'77 | EB'77 |
| 1978 | | 2008 | | | | | | | | | | |
| 1979 | | 2008 | | | | | | | | | | |
| 1980 | | 2008 | | | | | | | | 1986 | | |
| 1981 | | 2008 | | 2008 | | | | | | | | |
| 1982 | | 2008 | | | | | | | | | | |
| 1983 | | 2008 | | | | | | | | 1986 | | |
| 1984 | | 2008 | | | | | | | | | | |
| 1985 | | 2008 | | | | | | | | | | |
| 1986 | | 2008 | | | | | | | | EB'86 | | |
| 1987 | | 2008 | | | | | | | | | | |
| 1988 | | 2008 | | | | | | | | | | |
| 1989 | | 2008 | | | | | | 1993 | 1993 | | | |
| 1990 | | 2008 | | 2008 | | | | 1993 | 1993 | | | |
| 1991 | | 2008 | | | | | | 1993 | | | | |
| 1992 | | 2008 | | | | | | 1993 | | | | |
| 1993 | | 2008 | | | | | | EB'93 | EB'93 | | | |
| 1994 | | 2008 | | | | | 1997 | | | | | |
| 1995 | | 2008 | | | | | 1997 | | | | | |
| 1996 | | | | | | | 1997 | | | | | |
| 1997 | | 2008 | 2008 | | | CBS'97 | CBS'97 | | | | | |
| 1998 | | 2008 | 2008 | | | | | | | | | |
| 1999 | | 2008 | 2008 | 2008 | | 1997 | | | | | | |
| 2000 | | 2008 | 2008 | | | | | | | | | |
| 2001 | | 2008 | 2008 | | | | | | | | | |
| 2002 | 2002 | 2008 | 2008 | | | 1997 | | | | | | |
| 2003 | | 2008 | 2008 | | | | | | | | | |
| 2004 | 2004 | 2008 | 2008 | | 2008 | | | | | | | |
| 2005 | | 2008 | 2008 | | | | | | | | | |
| 2006 | 2006 | 2008 | 2008 | 2008 | 2008 | | | | | | | |
| 2007 | | 2008 | 2008 | | | | | | | | | |
| 2008 | 2008 | ESS'08 | ESS'08 | ESS'08 | ESS'08 | | | | | | | |
| 2009 | | 2008 | 2008 | | | | | | | | | |
| 2010 | 2010 | 2008 | | | | | | | | | | |
| 2011 | | 2008 | | | | | | | | | | |
| 2012 | 2012 | 2008 | | | | | | | | | | |

**Table 28 Reference distributions and reference boundaries for happiness items**

| | Round 1 | | | | Round 2 | | | | | Round 3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Reference distribution** | **ESS 2008** | | | **ESS 2006** | **CBS 1997** | | **DHS 1993** | | **WVS 1981** | **EB 1986** | |
| **Estimated alpha** | **10.20** | | | **8.72** | **10.51** | | **12.88** | | **11.72** | **10.09** | |
| **Estimated beta** | **3.33** | | | **2.99** | **3.53** | | **4.25** | | **4.22** | **3.73** | |
| **Estimated mean** | **7.54** | | | **7.45** | **7.49** | | **7.52** | | **7.35** | **7.30** | |
| **Response option** | ***CBS 5p-v*** | ***SCP 5p-v*** | ***DHS 5p-v*** | ***WVS 4p-v*** | ***CBS 5p-v*** | ***SCP 5p-v*** | ***CBS 5p-v*** | ***SCP 5p-v*** | ***EB 3p-v*** | ***CBS 5p-v*** | ***SCP 5p-v*** |
| - Very happy | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 |
| - Happy | 8.5 | 8.7 | 8.6 | | 8.4 | 8.5 | 8.3 | 8.3 | | 8.3 | 8.3 |
| - Quite happy | | | | 7.8 | | | | | | | |
| - Pretty happy | | | | | | | | | 7.6 | | |
| - Neither happy nor unhappy | 6.1 | 6.4 | 6.5 | | 6.2 | 6.1 | 6.3 | 6.3 | | 6.1 | 6.1 |
| - Not too happy | | | | | | | | | 5.7 | | |
| - Not very happy | 5.1 | 5.1 | | 5.3 | 5.2 | 5.1 | 5.3 | 5.3 | | 4.9 | 4.9 |
| - Unhappy | 4.3 | 4.2 | 4.9 | | 3.9 | 4.2 | 4.4 | 4.4 | | 3.9 | 3.8 |
| - Very unhappy | | | 4.0 | | | | | | | | |
| - Not at all happy | | | | 3.8 | | | | | | | |

**Table 29 Reference distributions and reference boundaries for life satisfaction items**

| | Round 1 | | | | Round 2 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Reference distribution** | **ESS 2008** | | | | **CBS 1997** | | **EB 1993** | | **EB 1986** | **EB 1977** | |
| **Estimated alpha** | **7.92** | | | | **7.67** | | **6.35** | | **6.78** | **6.12** | |
| **Estimated beta** | **2.76** | | | | **2.80** | | **2.51** | | **3.05** | **2.67** | |
| **Estimated mean** | **7.41** | | | | **7.33** | | **7.17** | | **6.89** | **6.96** | |
| **Response option** | ***EB 4p-v*** | ***CBS 5p-v*** | ***WVS 10p-n*** | ***SCP 10p-n*** | ***SCP 5p-v*** | ***CBS 5p-v*** | ***CBS 5p-v*** | ***SCP 5p-v*** | ***CBS 5p-v*** | ***CBS 5p-v*** | ***SCP 5p-v*** |
| - Extraordinarily satisfied | | 10.0 | | | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 |
| - Very satisfied | 10.0 | 9.0 | | | 9.0 | 8.8 | 8.6 | 8.7 | 8.7 | 8.9 | 8.9 |
| - Satisfied | 7.5 | 7.8 | | | 7.8 | 7.5 | 7.2 | 7.2 | 7.4 | 7.7 | 7.7 |
| - Fairly satisfied | 4.7 | 5.7 | | | 5.8 | 5.9 | 5.4 | 5.4 | 5.5 | 5.7 | 5.7 |
| - Not very satisfied | | 4.8 | | | 4.7 | 4.9 | 4.3 | 4.4 | 4.4 | 4.6 | 4.5 |
| - Not at all satisfied | 3.5 | | | | | | | | | | |
| - 10[20] | | | 10.0 | 10.0 | | | | | | | |
| 9 | | | 9.0 | 9.0 | | | | | | | |
| 8 | | | 8.2 | 8.4 | | | | | | | |
| 7 | | | 6.6 | 6.9 | | | | | | | |
| 6 | | | 5.4 | 5.6 | | | | | | | |
| 5 | | | 4.8 | 4.8 | | | | | | | |
| 4 | | | 4.3 | 4.3 | | | | | | | |
| 3 | | | 3.9 | 3.7 | | | | | | | |
| 2 | | | 3.3 | 3.5 | | | | | | | |
| 1[21] | | | 3.0 | 3.2 | | | | | | | |

[20] In the WVS labelled with 'Satisfied' and in the SCP survey labelled with 'Completely satisfied'.
[21] In the WVS labelled with 'Dissatisfied' and in the SCP survey labelled with 'Completely dissatisfied'.

It can be seen from Tab. 28, that reference boundaries for different waves of one time series, do not differ much. For response options with the same label but in different scales, the differences are much larger, which is most notably for the response options labelled with 'Unhappy'. This corroborates the idea that how a response option with a certain label is appreciated, depends on the context of the scale, see Ch. 5.

Two of the items included for life satisfaction have a 10-point numerical scale. The reference boundaries for the numerically labelled response options of these scales are clearly not equidistant as can be seen in Tab. 29. Despite the differences in the labels used for the anchor points of both scales, the reference boundaries do not differ much.

## 10.4 Combining converted survey results into long time series

Having prepared the conversion process, the time series of the different survey items for happiness and life satisfaction can be converted such that the resulting estimated population means are on a comparable level. Once this is achieved, the converted times series can be pooled into long consistent times series for happiness and life satisfaction in The Netherlands, spanning a period of almost 40 years.

### *10.4.1 Conversion of population means for time series of individual survey items*

Given estimates of the parameters $\alpha$ and $\beta$ of the beta distribution of happiness and life satisfaction, we can calculate an estimate of the population mean as $(10*\alpha/(\alpha + \beta))$ for each survey in a time series. By doing this for all estimated beta distributions, we obtain a time series of converted population means for each individual survey item. We recall from Sec. 7.3 that if the mode of surveying does not change, the reference boundaries can be kept fixed over time when applying the Continuum Approach and the differences in estimated means can solely be attributed to changes in the frequency distributions on the primary scale. The use of reference distributions brings the means for different survey items to a comparable level. The time series of the converted means are presented in Fig. 31 for happiness and in Fig. 32 for life satisfaction.

**Figure 31 Converted time series for happiness in The Netherlands**

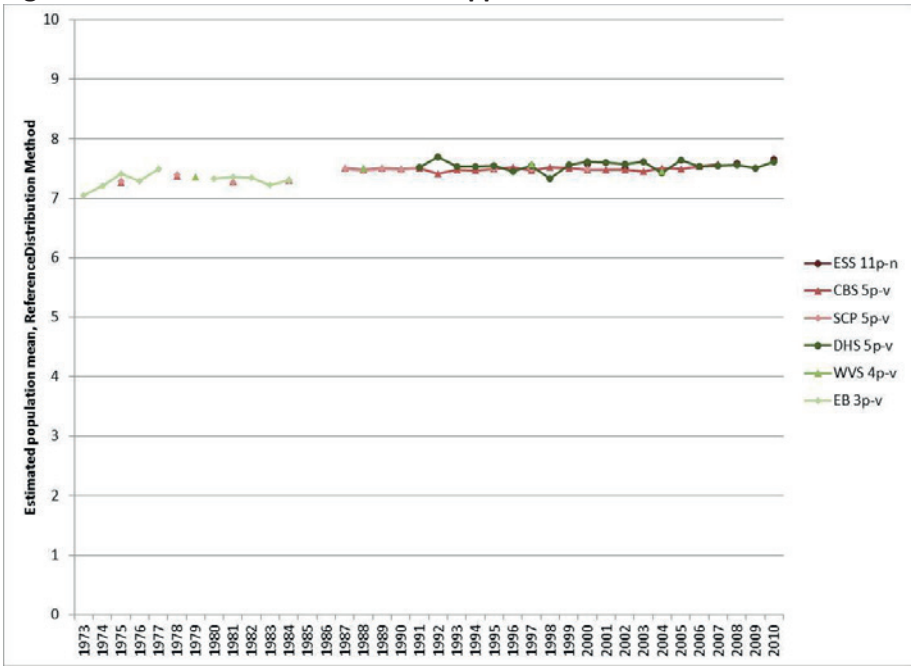
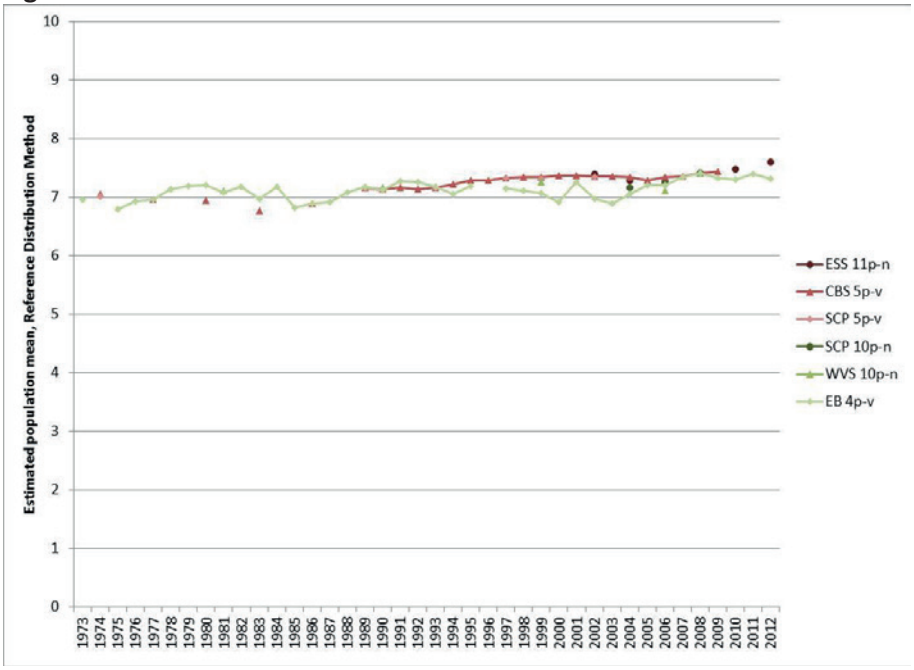



**Figure 32 Converted time series for life satisfaction in The Netherlands**



Comparing the converted time series given in Fig. 31 and 32 with the unconverted time series in Fig. 1 and 2 of Sec. 1.3, it can be seen that the conversion process, as intended, brought the estimated population means to a comparable level, also, the pattern over the year for each individual time series remains largely unchanged after the conversion, except for the waves for which we corrected for discontinuities and for the magnitude of the fluctuations. The magnitude of the fluctuations is stretched somewhat, due to the conversion of the data from a low number of response options to a continuum spanning a larger numerical range.

### *10.4.2 Pooling of converted time series*

The last step to obtain one long time series for happiness and life satisfaction is to pool the converted times series of the individual survey items. We have chosen to pool the time series in the most straightforward manner one can think of, which corresponds to taking the average of the estimated population means for each year. This gives one un-weighted average population mean per year, joining together all individual converted time series into the required time series. We have depicted these pooled time series in Fig. 33 and 34 to which we have also added the trend line to give an indication of the trends in happiness and life satisfaction in The Netherlands over the past four decennia.

It can be seen from Fig. 33 and 34 that pooling the converted time series has flattened the fluctuations of the individual time series. We conclude that the Dutch population has become slightly happier and more satisfied with life since the early seventies of the twentieth century. According to the trend lines, average happiness has increased by a little more than 0.3 points from 7.27 in 1975 to 7.60 in 2012, whereas the increase in average life satisfaction in the same period has amounted to a little more than 0.4 points going from 6.95 in 1975 to 7.37 in 2012.

**Figure 33 Pooled converted time series for happiness in The Netherlands**

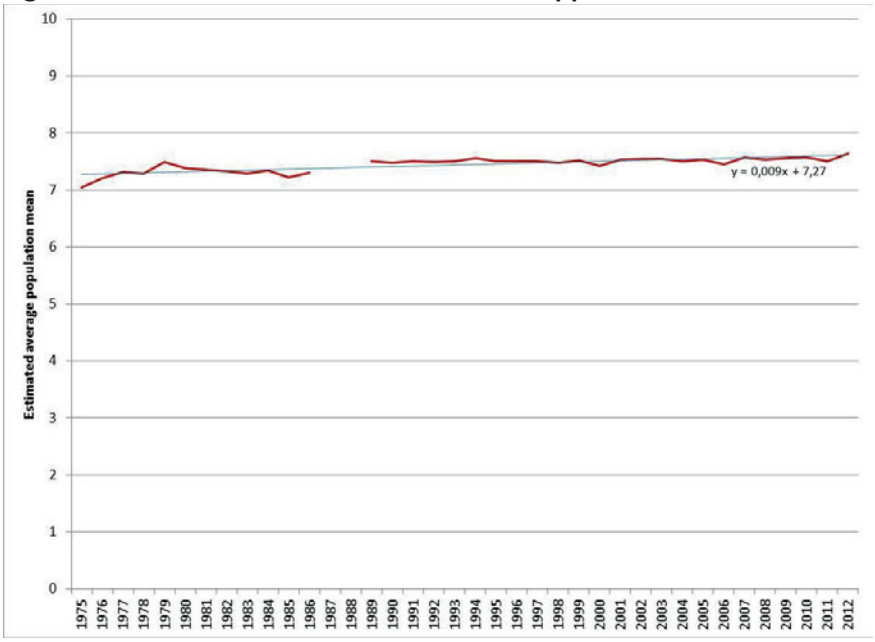
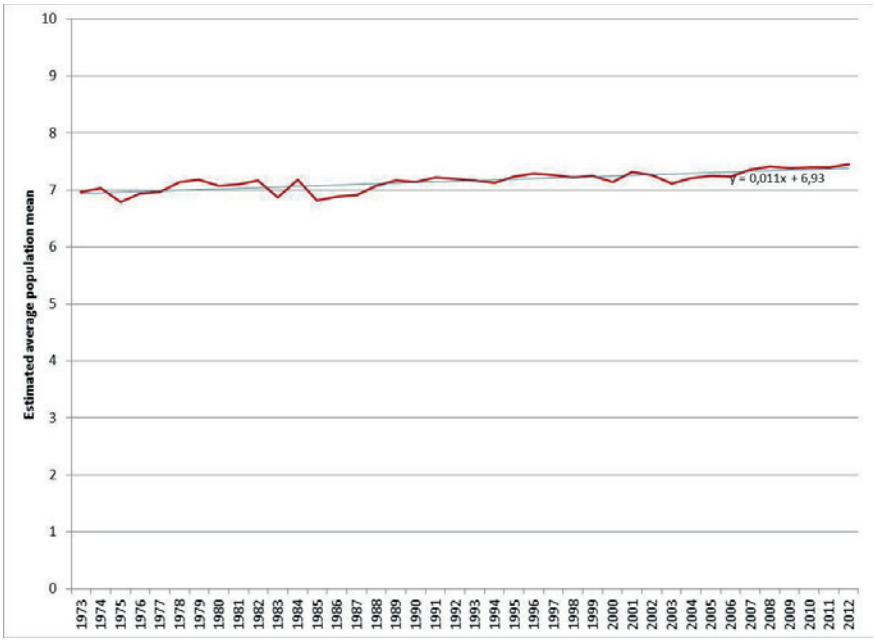



**Figure 34 Pooled converted time series for life satisfaction in The Netherlands**



## 10.5 Discussion

In this chapter we present an example application of the Reference Distribution Method for pooling time series based on slightly different survey questions on the same topic. We use the method to pool time series for the topics happiness and satisfaction with life. In this section we will discuss some methodological considerations.

As we stated in Sec. 7.1 the Reference Distribution Method is an attempt to deal with the fact that, for a given year and a given population, one would expect the estimated distribution means for similar questions about happiness asked in different representative surveys to be approximately the same irrespective of the primary response scales used. This equality of distribution means implies that the estimated cumulative distribution functions are exactly the same for both primary scales, but leading to different reference boundaries for these primary scales. This requires that the reference distribution must be based on a model which does not depend on the scale of measurement. Beta distributions meet this requirement. In appendix G we have shown that although the application of the Reference Distribution Method would be easier if the fully continuous model, the beta distribution, would be replaced by a semi-continuous model, this would not result in comparable estimates of the population means.

The standard version of the Eurobarometer has a spring wave and an autumn wave and sometimes one or more extra waves in a year. For the pooling of time series as described in this thesis, we calculated the un-weighted average of the frequency distributions for each year of the EB, to obtain one sample mean per year. Another option would have been to consider the spring waves and the autumn waves each as constituting different time series and to convert them separately. In this case, the possible seasonal effect would not have been averaged out before the application of the Reference Distribution Method, but after the conversion when pooling them together with other survey time series into one long time series. It would then be possible to investigate whether or not the seasonal effect is more or less stable over time.

There are sizeable differences in the number of respondents per survey to the survey items we looked at. In pooling the converted time series, we did not take these differences into consideration and calculated an un-weighted average for each year. It would be interesting to investigate if the trends found would differ very much if, when averaging over the estimated means in a year, these means were weighted by the number of respondents to the corresponding survey.

We only used one wave of the ESS to estimate an initial reference distribution for the conversion of the time series of one survey. The initial reference distribution defines the level of reference for the conversion results of all waves from each survey. To obtain a more stable pooled time series that is less prone to systematic errors, it would be useful to consider repeating the whole conversion process using each wave of the ESS once to derive an initial reference distribution from. Instead of pooling time series which are converted on the basis of just one initial reference distribution, conversion results for initial reference distributions based on other waves could be included in the pooling.

We kept the boundaries between response options that were derived from a reference distribution fixed for all subsequent waves for which the survey design had not undergone a significant change. It is a reasonable assumption that the boundaries will be more or less stable over time and may be kept fixed, as we have described in Ch. 8.

The Reference Distribution Method can be used to correct for discontinuities due to mode changes and, to a limited extent, to account for contextual influences; yet, as we already stated in Sec. 7.5.2, it cannot solve all the comparability problems. For example an effect on the response due to a re-ordering of questions cannot be corrected for by the Reference Distribution Method if the effect is small or the re-ordering occurs frequently and affects the response of a number of waves. Furthermore, if there is a response shift, this will be difficult to correct for since in time series this shift may occur over a long period of time and we assume that if it occurs, this will be the case for all response scales under consideration, both verbal and numerical.

## 10.6 Conclusion

We conclude that the Reference Distribution Method is a useful method to pool time series from different surveys to make the analyses of trends over a long time span possible. Having applied this method to pool time series of responses to different survey items on happiness and life satisfaction in the Netherlands, we conclude that in the past 40 years the Dutch have become slightly happier and more satisfied with their lives.

# 11 Directions for further research

In the preceding chapters we have presented some innovative methods to 1) get a better understanding of how response options are interpreted in the context of the scale and the topic of concern and 2) to make responses to different survey questions on the same topic comparable in a feasible way. Although we believe that we have made a step forward, there is still a number of questions left to be answered. We will discuss some possible refinements and applications of the new methods that come to mind in this chapter.

## 11.1 Refinements

### *11.1.1 Differential interpretation of items by subgroups*

In Ch. 4 we have described that in The Netherlands there are at least two subgroups of the general population that assign an equal degree of appreciation to equivalent response options for happiness and life satisfaction. Between these subgroups there were no noteworthy differences between the assessments of the scales. It is possible that other subgroups would appreciate the same response scales differently, but that each subgroup would still appreciate equivalent response options for happiness and life satisfaction in the same way. In addition to this, we found that in Spain and Chile at least the subgroup of students does not appreciate equivalent response options for happiness and life satisfaction in the same way. If this difference in appreciation does not hold for the general population: logically there must be at least one other subgroup in each of the two countries that appreciates response options for happiness differently given equivalent options for life satisfaction. In this case this difference in appreciation for this latter subgroup must be opposite to the appreciation of the students, if it is assumed that a difference in interpretation cannot be found in the general population.

These findings combined with the findings of Ch. 5 that 'very happy' is not always equally happy offer a direction for further research into the differential possible interpretation of response items by subgroups of a general population based on, for example, level of education, type of occupation or age and to mutually compare the outcomes for similar subgroups from different countries. The results of such research give insight in the extent to which people within subgroups agree on the appreciation of response options and whether this differs between subgroups. It could

also shed light on whether a difference or indifference in the interpretation of response options in equivalent scales for different topics is present population wide or just in some subgroups.

The research for subgroups of the general population should not be restricted to just the response options. The leading questions may also be an influential factor. We have assumed that the leading questions of the items used for the time series which we have pooled in Ch. 10 aimed at measuring the same or at least were interpreted as such by the respondents. Whether this is true could be checked by conducting a Scale Interval Study in which the leading question from each of these items would be combined with the response options of the other items. Instead of studying the interpretation of identical response options in the context of the scale, such a study would amount to studying identical response scales in the context of the leading question.

### *11.1.2 Improvement of HSIS-ratings*

In the HSIS-studies conducted so far, no attention has been paid to the subjective well-being (SWB) of the judges employed for these studies. Thus we do not know whether the assessment of the boundaries between response options is affected by this SWB. It is desirable to set up a HSIS-study to investigate this, as has been recommended by Kalmijn (2010, p. 179) previously. In such a study it would also be worthwhile to distinguish between the experienced well-being and the evaluative well-being of the judges. The first of these two kinds of SWB refers to the momentary assessment of affect and the second to the appreciation of life as a whole. A person's experienced well-being and evaluative well-being may opposite one-another (National Research Council, 2013, pp. 15-16) and thus each of these kinds of SWB may have a different effect on the assessment of the boundaries between response options.

In Sec. 5.3 we have discussed the occurrence of zero-width intervals assigned by judges to anchor points and referred to a more elaborated discussion about this issue by Kalmijn (2010, p. 147 sqq). A suggestion for further research with respect to this is to compare the occurrence of such zero-width intervals for the items between groups of native speakers from different countries and to investigate whether there is a cultural pattern in the results. In connection with this, it would be interesting to determine whether there is a relation between the occurrence of zero-width intervals and the '10-excess phenomenon' described in Brulé and Veenhoven (2014) according to which the frequency with which respondents tick a 10 stands

out and is sometimes higher than the frequency with which a 9 is ticked by respondents in the same sample.

The precision of the estimates of the boundaries between response options using the Scale Interval Recorder could be improved, if in future HSIS-studies the number of judges was considerably increased compared to the average number used for studies to date.

The stability over time of the boundaries between response options obtained by using the Scale Interval Recorder could be checked by repeating a HSIS-study after an interval of several years using the same or similar subgroups of judges.

### *11.1.3 Improvement of estimates obtained with the Reference Distribution Method*

To generalize the conclusion of Ch. 8 about the stability of the reference boundaries between response options, further research into the stability of the boundaries is required for other countries such as Greece, where the average level of life satisfaction has undergone a large change since the start of the economic crisis of 2008, when, according to the Eurobarometer, the percentage of respondents who were at least fairly satisfied with their life dropped from around 65 percent to around 50 percent.

In Sec. 10.5.1 we mentioned two options for further research on the validation of the Reference Distribution Method. These options were 1) to weigh by the number of respondents to each survey when averaging over the estimated means in a year to pool time series, 2) to repeat the process of pooling time series with the other waves of the European Social Survey (ESS) that are available for this, in order to obtain a more stable pooled time series that is less prone to systematic errors.

The main emphasis of this thesis has been put on the introduction of new techniques to deal with the comparability problem caused by differences in survey items used the measure the same latent variable. We have restricted the inspection of time series to distortions caused by the influential factors we have discussed in Sec. 10.2. Although we pooled time series from different surveys, we have not paid attention yet to the quality of the data sources used in terms as for instance the non-response, the selectivity of the response, and whether or not correction methods such as re-weighting of the results were applied. To improve the reliability of the pooled time series, it is necessary to also investigate the quality of the data sources and to decide which data sources meet the required quality and which data sources do not.

The items used for pooling time series as presented in Ch. 10 all come from surveys conducted in The Netherlands: to further improve and validate the method, it is necessary to apply the method to time series of survey results from other countries which at the same time will allow us to study the differences between countries in the interpretation of scales and how respondents in practice cope with response options.

In addition to the preceding points, it would be interesting to repeat the process of pooling time series by deriving initial reference distributions from survey results measured using other numerical scales than those taken from the ESS to investigate whether that would lead to similar outcomes. Candidate response scales for such an exercise are, for example, the 10-point scales of the items on happiness and life satisfaction from the European Quality of Life Survey.

From the perspective of international comparisons, it is sensible to investigate how much difference it makes whether a 10-point or an 11-point numerical scale is used to derive a reference distribution from. In both cases, the primary numerical scale has to be transformed to fit into the 0 to 10 continuum. For a 10-point scale this corresponds to shifting down the ranks of all the response options by 0.5 (Kalmijn, 2013). The differences in sample means between countries when based on the Weighted Average Approach, would therefore not change due to the transformation of the 10-point scale. If an 11-point numerical scale is used, however, the value of the response options with a rank lower than 5 would be higher on the transformed scale than the rank on the primary scale, whereas the value of the other response options ranked higher than 5 would be lower than the rank of the primary scale. The transformation of the 11-point numerical scale would thus affect the differences between the samples means according to the Weighted Average Approach. It has to be investigated whether this difference in effect of using a 10-point or 11-point scale as a basis is of importance for the conclusions that can be drawn from the results of applying the Reference Distribution Method.

The 10-excess phenomenon which we referred to in Sec. 11.1.2 does not occur as often in each country Brulé and Veenhoven (2014). This begs for research to find out what the impact of the 10-excess phenomenon is on the derivation of a reference distribution and what solutions can be thought of to assure that this impact does not come at the cost of the validity of an international comparison.

### *11.1.4 Comparison of the estimates obtained with the HSIS and the Reference Distribution Method*

In Sec. 6.5 we have shown that the combination of the assessment of boundaries between response options using the Happiness Scale Interval Recorder and the Continuum Approach cannot be used to solve the comparability problem. We explained in Secs. 5.5 and 6.5 that this may be due to the fact that a verbal response scale does not necessarily offer response options that meet the perception of respondents well, which may force them to choose between two less than optimal alternatives. As a consequence, the boundaries between response options obtained using the Scale Interval Method differ dramatically from the boundaries derived from a reference distribution as done in the Reference Distribution Method, which we illustrated in Fig. 19 of Sec. 7.2. From a scientific point of view, it would be of interest to study how the boundaries found using the Scale Interval Method differ from the boundaries obtained using the Reference Distribution Method for a number of countries from all over the world to reveal whether groups of countries can be detected for which these differences resemble and whether this can be attributed to cultural differences. The results of such a study would contribute to increasing insight into the extent to which language and culture affect the interpretation of response options in surveys and, as a corollary, the response to survey questions.

## 11.2 Applications

### *11.2.1 Application in research synthesis of happiness*

The innovations we have presented in this thesis were all developed in the context of the World Database of Happiness (WDH). Apart from the refinements we have suggested in Sec. 11.1, these methods appear to be applicable for synthetic analysis. We will discuss some of the options for synthetic analysis in this section.

In the WDH distributional findings on happiness based on survey questions that validly tap an individual's 'overall appreciation of life as a whole' are presented. These distributional findings include the mean and the standard deviation for which the comparison within and across countries is facilitated by a transformation to a 0 to 10 scale. A better alternative for such a transformation is to apply the Reference Distribution Method to estimate the mean and the standard deviation in the general population. This would allow a better comparison of survey results on happiness for which the estimates of the mean and the standard deviation

are based on a reference distribution which is rooted in one and the same survey and the frequency distributions in a particular year. The item on happiness from the European Social Survey and the corresponding frequency distributions in 2010 may, for example, serve as a source to derive reference distributions from for all participating countries.

The pooling of time series as demonstrated in Sec. 10 allows presenting longer time series for happiness in the WDH and to study trends over a longer time span than is possible at this moment. In Sec. 8.6 we have stated that the estimated population mean of a reference distribution can serve as a reference value for comparison purposes, but should not be considered to be the 'true' value of the perception of happiness on the continuum from 0 to 10. Other reference distributions may constitute a different reference value. If we assume however, that a reference value is mainly determinant for bringing survey results to a comparable level, but does not affect the pattern of the trend of a time series, than the trends of different pooled time series which are based on different reference distributions can be compared. Similar to what we remarked in Sec. 9.4 for the trends in demographic groups, the absolute difference between the pooled time series is of less importance for trend analysis as long as the development of the trend in each series shows a reliable pattern after transition to the 0 to 10 continuum.

A thorough analysis of the results from all the HSIS-studies that have been conducted would be helpful to reveal groups of countries where response scales are interpreted similarly and to identify for which response scale this is the case. A comparison of the level of happiness between countries within those groups will be more reliable since no differential interpretation of the response scales has to be taken into account.

### *11.2.2 Application in new research on happiness*

From the comparison of the assessments by Dutch-speaking judges and the two groups of Spanish-speaking judges, described in Ch. 4, it is clear that results for one country cannot be generalized to other languages and cultures. An extended study using the Scale Interval Recorder would contribute to investigating further the hypothesis of Saris and Andreenkova (2001) which we mentioned in Sec. 4.1, that the relationship between happiness and satisfaction with life varies with the cultural and linguistic environment in which it is studied.

We recall from Ch. 4 that although people in different countries may speak the same language, due to cultural influences, they do not necessarily interpret response scales in the same way. The Scale Interval

Recorder offers an opportunity to study the differences in interpretation of response options by native speakers of one language, but living in different countries or forming parts of different subcultures. Examples of languages and countries would be Chinese and China, Malaysia and other countries where Chinese is a primary language for a cultural group, French and France, Canada, Belgium, Morocco, Haiti and Switzerland where French is a primary language for a cultural group, or English and Great Britain, the United States, Canada, New Zealand and South Africa where English is a primary language for a cultural group.

A variant on the study described above, would be to compare the interpretation of response options by groups of native speakers who live in the same country but speak a different language. The aim of this type of research comes close to the research reported in a paper by Senik (2013), who compares the happiness of French natives with the happiness of immigrants living inside and outside France and to that of the survey study by Van der Houwen and Moonen (2014), who compared the happiness of Dutch natives with that of immigrants.

### *11.2.3 Application to other topics than happiness*

In Sec. 4.5.3 we introduced the term 'interpreters' bias' for the phenomenon that an adverb used in the label of a response options is not necessarily always translated by same the adverb in another language, such as when the translation in Spanish of the adverb 'Fairly' in the label 'Fairly satisfied' in English would be 'Más bien' for one scale and 'Bastante' for another scale. To improve the comparison of happiness across nations it is of importance to gain more uniformity in the translation of adverbs. Therefore it is necessary to get more insight in the size of the interpreters' bias in terms of the frequency of occurrence in different countries. To reduce the diversity in adverbs used in a language for the translation of a single adverb in another language, the Happiness Scale Interval Recorder can be used as an instrument to get insight into the extent to which different adverbs are interpreted differently in the context of the scale and the keyword of concern. We recommend not to restrict this to the keywords happiness and satisfaction, but to include also response options being labelled with other keywords. In European socio-economic and opinion surveys which are mentioned on the website[22] of the European Data Center for Work and Welfare (EDAC), we found, for example, response

[22] http://www.edac.eu/

options labelled with the adverb ‘fairly’ in combination with a number of keywords such as ‘easily’, ‘difficult’, ‘likely’, ‘safe’, ‘good’ and ‘bad’.

In this thesis we have considered the comparability problem only from the perspective of happiness. The comparability problem, however, is ubiquitous in social science research for a range of topics, such as how satisfied people are with their housing, working conditions or their present job, but also how safe they feel when walking alone in their neighbourhood at night, how they assess their own health, how interested they are in politics or how well they are able to make ends meet given their household income. Examples of the survey items and the diversity of response scales used, can be found in the surveys mentioned at the previously introduced website of EDAC and also in surveys of National Statistical Offices such as the New Zealand General Social Survey we mentioned in Sec. 1.5.1. Besides social science research there are also other fields of research which have to cope with the comparability problem. Two fields of research in which the comparability problem is as prominent as in happiness research are those of consumer satisfaction (Bartikowski, Kamei, and Chandon, 2010) and of customer satisfaction (MORI 2002; Daner and Haddrell, 2011; Market Directions, 2014). What makes customer satisfaction especially interesting from the perspective of the interpretation of response items, is that it is of concern to parties in both the private and public sector. With respect to the comparability problem the field of health research is also worth mentioning here. In this field the perceptions people have of their physical functioning, of the pain they experience or of their mental well-being are measured (Schwartz and Sprangers, 1999; McDowell, 2006; Davis et al, 2010). Just like in happiness research and the other fields of research we referred to, the differences in response scales used over time makes comparison of survey results in health research difficult.

Based on the preceding we recommend making an inventory of other topics within the social sciences and other fields of research for which the perceptions of people have been measured using survey research and discrete scales over a long period of time and for which the comparability problem is a serious issue. Based on this inventory and the results of, and directions for further research in this thesis, a plan can be made for research to address the comparability problem found in these fields. This would also give us an opportunity to find out whether or not using the family of beta distributions is appropriate for other topics than happiness and life satisfaction when converting time series and, if so, or not, determine the methods that are most appropriate to use in these situations.

# APPENDICES

## Appendix A Survey questions on happiness from the HSIS-studies used in this thesis

**Table A.1 Items from the Dutch HSIS-study**

| Wording in Dutch | Wording in English | Question code |
|---|---|---|
| Vindt u zichzelf gelukkig?<br>- Ja<br>- Nee | Are you happy with your life?<br>- Yes<br>- No | O-HL-u-sq-v-2-a |
| Voelt u zichzelf..<br>- Gelukkig<br>- Tamelijk gelukkig<br>- Ongelukkig | Do you feel…….?<br>- Happy<br>- Fairly happy<br>- Unhappy | M-FH-u-sq-v-3-d |
| Voelt u zich<br>- Gelukkig<br>- Niet zo gelukkig<br>- Ongelukkig | Do you feel…….?<br>- Happy<br>- Not very happy<br>- Unhappy | M-FH-u-sq-v-3-a |
| Nu een vraag over de manier waarop u op het ogenblik vooruit komt: bent u daarover<br>- Zeer tevreden<br>- Tevreden<br>- Ontevreden | How satisfied are you with the way you are getting on now?<br>- Very satisfied<br>- Satisfied<br>- Not satisfied | O-SLS-c-sq-v-3-ab |
| Hoe tevreden bent u over het algemeen met het leven dat u leidt?<br>- Zeer tevreden<br>- Tamelijk gelukkig<br>- niet zo tevreden<br>- helemaal niet tevreden | On the whole how satisfied are you with the life you lead?<br>- Very satisfied<br>- Fairly satisfied<br>- Not very satisfied<br>- Not at all satisfied | O-SLL-u-sq-v-4-b |

**Table A.1 Continued**

| Wording in Dutch | Wording in English | Question code |
|---|---|---|
| In welke mate bent u tevreden met het leven dat u op dit moment leidt?<br>- Buitengewoon tevreden<br>- Zeer tevreden<br>- Tevreden<br>- Tamelijk tevreden<br>- Niet zo tevreden | To what extent are you satisfied with the life you currently lead?<br>- Extraordinarily satisfied<br>- Very satisfied<br>- Satisfied<br>- Fairly satisfied<br>- Not very satisfied | O-SLL-c-sq-v-5-d |
| In welke mate vindt u zichzelf een gelukkig mens?<br>- Erg gelukkig<br>- Gelukkig<br>- Niet gelukkig, niet ongelukkig<br>- Niet zo gelukkig<br>- Ongelukkig | To what extent do you consider yourself a happy person....?<br>- Very happy<br>- Happy<br>- Neither happy nor unhappy<br>- Not very happy<br>- Unhappy | O-HP-u-sq-v-5-a |
| In welke mate vindt u zichzelf een gelukkig mens?<br>- Erg gelukkig<br>- Gelukkig<br>- Niet gelukkig, niet ongelukkig<br>- Ongelukkig<br>- Erg ongelukkig | To what extent do you consider yourself a happy person....?<br>- Very happy<br>- Happy<br>- Neither happy nor unhappy<br>- Unhappy<br>- Very unhappy | O-HP-u-sq-v-5-d |

**Table A.1 Continued**

| Wording in Dutch | Wording in English | Question code |
|---|---|---|
| Hoe tevreden bent u met uw leven in het algemeen op dit moment?<br>- Zeer tevreden<br>- Tamelijk tevreden<br>- Tevreden noch ontevreden<br>- Tamelijk ontevreden<br>- Zeer ontevreden | How satisfied are you with your life in general these days?<br>- Very satisfied<br>- Fairly satisfied<br>- Satisfied nor dissatisfied<br>- Fairly dissatisfied<br>- Very dissatisfied | O-SLW-c-sq-v-5-fb |
| Hoe tevreden bent u met uw leven in het algemeen?<br>- Volledig tevreden<br>- Tevreden<br>- Niet tevreden, niet ontevreden<br>- Ontevreden<br>- Volledig ontevreden | How satisfied are you with your life in general?<br><br>- Completely satisfied<br>- Satisfied<br>- Unsatisfied nor satisfied<br>- Unsatisfied<br>- Completely unsatisfied | O-SLu-g-sq-v-5-c |
| Hoe gelukkig of ongelukkig bent u met uw leven in het algemeen?<br><br>- Volkomen gelukkig<br>- Zeer gelukkig<br>- Tamelijk gelukkig<br>- Noch gelukkig, noch ongelukkig<br>- Tamelijk ongelukkig<br>- Zeer ongelukkig<br>- Volkomen ongelukkig | If you were to consider your life in general, how happy or unhappy would you say you are, on the whole?<br>- Completely happy<br>- Very happy<br>- Fairly happy<br>- Neither happy nor unhappy<br>- Fairly unhappy<br>- Very unhappy<br>- Completely unhappy | O-HL-g-sq-v-7-a |

**Table A.2 Pairs of items on happiness and life satisfaction in the Dutch HSIS-study**

| Item wording in Dutch | Item wording in English | # judges | Item code |
|---|---|---|---|
| D7_2 Hoe gelukkig bent u, alles bijeengenomen?<br><br>- Heel gelukkig<br>- Tamelijk gelukkig<br>- Niet zo gelukkig<br>- Helemaal niet gelukkig | D7_2 Taking all things together, would you say you are.....?<br>- Very happy<br>- Fairly happy<br>- Not very happy<br>- Not at all happy | 345 | ***O-HL-u-sq-v-4-a*** |
| D7_4 Hoe tevreden bent u over het algemeen met het leven dat u leidt?<br>- Zeer tevreden<br>- Tamelijk tevreden<br>- Niet zo tevreden<br>- Helemaal niet tevreden | D7_4 On the whole how satisfied are you with the life you lead?<br>- Very satisfied<br>- Fairly satisfied<br>- Not very satisfied<br>- Not at all satisfied | 344 | ***O-SLL-u-sq-v-4-b*** |
| | | | |
| D7_6 In welke mate vindt u zichzelf een gelukkig mens?<br>- Buitengewoon gelukkig<br>- Zeer gelukkig<br>- Gelukkig<br>- Tamelijk gelukkig<br>- Niet zo gelukkig | D7_6 To what extent do you consider yourself a happy person?<br>- Extraordinarily happy<br>- Very happy<br>- Happy<br>- Fairly happy<br>- Not very happy | 341 | ***O-HP-u-sq-v-5-h*** |
| D6_6. In welke mate bent u tevreden met het leven dat u op dit moment leidt?<br>- Buitengewoon tevreden<br>- Zeer tevreden<br>- Tevreden<br>- Tamelijk tevreden<br>- Niet zo tevreden | D6_6. To what extent are you satisfied with the life you currently lead?<br>- Extraordinarily satisfied<br>- Very satisfied<br>- Satisfied<br>- Fairly satisfied<br>- Not very satisfied | 375 | ***O-SLL-c-sq-v-5-d*** |

**Tabel A.2 Continued**

| Item wording in Dutch | Item wording in English | # judges | Item code |
|---|---|---|---|
| D6_5 In welke mate vindt u zichzelf een gelukkig mens?<br>- Erg gelukkig<br>- Gelukkig<br>- Niet gelukkig, niet ongelukkig<br>- Ongelukkig<br>- Erg ongelukkig | D6_5 To what extent do you consider yourself a happy person….?<br>- Very happy<br>- Happy<br>- Neither happy nor unhappy<br>- Unhappy<br>- Very unhappy | 373 | ***O-HP-u-sq-v-5-d*** |
| D6_8 In welke mate bent u tevreden met het leven dat u op dit moment leidt?<br>- Erg tevreden<br>- Tevreden<br>- Niet tevreden, niet ontevreden<br>- Ontevreden<br>- Erg ontevreden | D6_8 To what extent are you satisfied with the life you currently lead?<br>- Very satisfied<br>- Satisfied<br>- Neither satisfied, nor dissatisfied<br>- Dissatisfied<br>- Very dissatisfied | 364 | ***O-SLL-c-sq-v-5-f*** |

**Table A.3 Pairs of items on happiness and life satisfaction in the Spanish and Chilean HSIS-studies**

| Item wording in Spanish | Item wording in English | # judges | Item code |
|---|---|---|---|
| En general, tal y como están las cosas hoy en día, ¿cómo diría usted que es hoy en día?<br>- Muy feliz<br>- Bastante feliz<br>- No muy feliz | Taking all things together, would you say you are.....?<br>- Very happy<br>- Fairly happy<br>- Not very happy | Spain 401<br>Chile 24 | ***O-HL-c-sq-v-3-ab*** |
| Qué tan satisfecho/a está usted de como le está yendo actualemente?[23]/ ¿Como satisfecho/a está usted con cómo le está yendo actualmente?[24]<br>- Muy satisfecho/a<br>- Bastante satisfecho/a<br>- No muy satisfecho/a | In general, would you say that you are satisfied with your life?<br>- Very satisfied<br>- Fairly satisfied<br>- Not very satisfied | Spain 408<br>Chile 39 | ***O-SLS-c-sq-v-3-ab*** |
| | | | |
| En términos generales, ¿diría Ud. que es...?<br>- Muy feliz<br>- Bastante feliz<br>- No muy feliz<br>- Nada feliz | Taking all things together, would you say you are.....?<br>- Very happy<br>- Fairly happy<br>- Not very happy<br>- Not at all happy | Spain 410<br>Chile 23 | ***O-HL-u-sq-v-4-a*** |
| En términos generales, ¿diría Ud. que está satisfecho/a con su vida? Diría Ud. Que está..<br>- Muy satisfecho/a<br>- Bastante satisfecho/a<br>- No muy satisfecho/a<br>- Para nada satisfecho/a | In general, would you say that you are satisfied with your life?<br>- Very satisfied<br>- Fairly satisfied<br>- Not very satisfied<br>- Not at all satisfied | Spain 370<br>Chile 67 | ***O-SLu-g-sq-v-4-c*** |

[23] Leading question in Chilean version
[24] Leading question in Spanish version

## Appendix B Differences in assessment upper boundaries by employees, students and total

**Table B.1 Mean and standard error of assessment upper boundaries by employees, students and total**

| Question code | Wording in English | Mean (employees) | Standard error (employees) | Mean (students) | Standard error (students) | Mean (total) | Standard error (total) |
|---|---|---|---|---|---|---|---|
| O-HL-u-sq-v-2-a | Are you happy with your life? | | | | | | |
| | - Yes | 10.0 | | 10.0 | | 10.0 | |
| | - No | 5.8 | 0.1 | 5.8 | 0.1 | 5.8 | 0.1 |
| M-FH-u-sq-v-3-d | Do you feel.....? | | | | | | |
| | - Happy | 10.0 | | 10.0 | | 10.0 | |
| | - Fairly happy | 7.4 | 0.0 | 7.1 | 0.1 | 7.2 | 0.1 |
| | - Unhappy | 4.7 | 0.1 | 4.0 | 0.1 | 4.4 | 0.1 |
| M-FH-u-sq-v-3-a | Do you feel.....? | | | | | | |
| | - Happy | 10.0 | | 10.0 | | | |
| | - Not very happy | 6.6 | 0.1 | 6.4 | 0.1 | 6.5 | 0.1 |
| | - Unhappy | 3.8 | 0.1 | 3.1 | 0.1 | 3.4 | 0.1 |
| O-SLS-c-sq-v-3-ab | How satisfied are you with the way you are getting on now? | | | | | | |
| | - Very satisfied | 10.0 | | 10.0 | | 10.0 | |
| | - Satisfied | 8.0 | 0.1 | 7.9 | 0.1 | 7.9 | 0.0 |
| | - Not satisfied | 4.8 | 0.1 | 4.3 | 0.1 | 4.5 | 0.1 |
| O-SLL-u-sq-v-4-b | On the whole how satisfied are you with the life you lead? | | | | | | |
| | - Very satisfied | 10.0 | | 10.0 | | 10.0 | |
| | - Fairly satisfied | 7.9 | 0.1 | 7.8 | 0.1 | 7.9 | 0.0 |
| | - Not very satisfied | 5.5 | 0.1 | 5.1 | 0.1 | 5.3 | 0.1 |
| | - Not at all satisfied | 3.1 | 0.1 | 2.9 | 0.1 | 3.0 | 0.1 |

**Table B.1 Continued**

| Question code | Wording in English | Mean (employees) | Standard error (employees) | Mean (students) | Standard error (students) | Mean (total) | Standard error (total) |
|---|---|---|---|---|---|---|---|
| O-SLL-c-sq-v-5-d | To what extent are you satisfied with the life you currently lead? | | | | | | |
| | - Extraordinarily satisfied | 10.0 | | 10.0 | | 10.0 | |
| | - Very satisfied | 8.8 | 0.1 | 8.9 | 0.0 | 8.8 | 0.0 |
| | - Satisfied | 7.2 | 0.1 | 7.1 | 0.1 | 7.2 | 0.1 |
| | - Fairly satisfied | 5.5 | 0.1 | 5.2 | 0.1 | 5.3 | 0.1 |
| | - Not very satisfied | 3.8 | 0.1 | 3.5 | 0.1 | 3.6 | 0.1 |
| O-HP-u-sq-v-5-a | To what extent do you consider yourself a happy person....? | | | | | | |
| | - Very happy | 10.0 | | 10.0 | | 10.0 | |
| | - Happy | 8.5 | 0.0 | 8.5 | 0.0 | 8.5 | 0.0 |
| | - Neither happy nor unhappy | 6.3 | 0.1 | 5.9 | 0.1 | 6.1 | 0.1 |
| | - Not very happy | 4.5 | 0.1 | 4.1 | 0.1 | 4.3 | 0.1 |
| | - Unhappy | 2.5 | 0.1 | 2.1 | 0.1 | 2.3 | 0.1 |
| O-HP-u-sq-v-5-d | To what extent do you consider yourself a happy person....? | | | | | | |
| | - Very happy | 10.0 | | 10.0 | | 10.0 | |
| | - Happy | 8.5 | 0.0 | 8.3 | 0.1 | 8.4 | 0.0 |
| | - Neither happy nor unhappy | 6.1 | 0.1 | 5.8 | 0.1 | 5.9 | 0.1 |
| | - Unhappy | 4.3 | 0.1 | 4.0 | 0.1 | 4.1 | 0.1 |
| | - Very unhappy | 2.0 | 0.1 | 1.6 | 0.1 | 1.8 | 0.1 |

**Table B.1 Continued**

| Question code | Wording in English | Mean (employees) | Standard error (employees) | Mean (students) | Standard error (students) | Mean (total) | Standard error (total) |
|---|---|---|---|---|---|---|---|
| O-SLW-c-sq-v-5-fb | How satisfied are you with your life in general these days? | | | | | | |
| | - Very satisfied | 10.0 | | 10.0 | | 10.0 | |
| | - Fairly satisfied | 8.1 | 0.1 | 8.1 | 0.1 | 8.1 | 0.0 |
| | - Satisfied nor dissatisfied | 5.9 | 0.1 | 5.7 | 0.1 | 5.8 | 0.0 |
| | - Fairly dissatisfied | 4.5 | 0.1 | 4.2 | 0.1 | 4.3 | 0.1 |
| | - Very dissatisfied | 2.4 | 0.1 | 2.3 | 0.1 | 2.3 | 0.1 |
| O-SLu-g-sq-v-5-c | How satisfied are you with your life in general? | | | | | | |
| | - Completely satisfied | 10.0 | | 10.0 | | 10.0 | |
| | - Satisfied | 8.7 | 0.1 | 8.5 | 0.1 | 8.6 | 0.0 |
| | - Unsatisfied nor satisfied | 6.2 | 0.1 | 5.9 | 0.1 | 6.0 | 0.1 |
| | - Unsatisfied | 4.3 | 0.1 | 4.0 | 0.1 | 4.2 | 0.1 |
| | - Completely unsatisfied | 1.8 | 0.1 | 1.5 | 0.1 | 1.6 | 0.1 |
| O-HL-g-sq-v-7-a | If you were to consider your life in general, how happy or unhappy would you say you are, on the whole? | | | | | | |
| | - Completely happy | 10.0 | | 10.0 | | 10.0 | |
| | - Very happy | 9.2 | 0.1 | 9.2 | 0.1 | 9.2 | 0.0 |
| | - Fairly happy | 7.6 | 0.1 | 7.5 | 0.1 | 7.5 | 0.0 |
| | - Neither happy nor unhappy | 5.8 | 0.0 | 5.6 | 0.1 | 5.7 | 0.0 |
| | - Fairly unhappy | 4.6 | 0.1 | 4.3 | 0.1 | 4.4 | 0.0 |
| | - Very unhappy | 3.0 | 0.1 | 2.8 | 0.1 | 2.9 | 0.0 |
| | - Completely unhappy | 1.1 | 0.1 | 1.1 | 0.1 | 1.1 | 0.0 |

## Appendix C Cumulative frequencies and parameters beta distributions

The tables given below contain the frequency distributions of the verbal scales depicted in Fig. 24 in Sec. 8.3. The parameters of the reference distribution and the frequency distribution for 2008 that has been used to derive the reference boundaries for the response options of the scale are given in bold in each table. The values of these fixed reference boundaries are given in Tab. 12 to 15 in Sec. 8.2.

**Table C.1 Cumulative frequencies life satisfaction and parameters beta distributions, CBS**

| | To what extent are you satisfied with the life you currently lead? | | | | | Parameters best fitting beta distribution | |
|---|---|---|---|---|---|---|---|
| Year | Not very satisfied | Fairly satisfied | Satisfied | Very satisfied | Extraordinarily satisfied | α | β |
| 1997 | 4.1% | 12.4% | 58.7% | 92.5% | 100.0% | 7.69 | 2.81 |
| 1998 | 3.8% | 11.9% | 58.1% | 92.4% | 100.0% | 7.83 | 2.82 |
| 1999 | 4.0% | 12.1% | 57.9% | 92.1% | 100.0% | 7.69 | 2.77 |
| 2000 | 3.8% | 11.7% | 57.4% | 92.0% | 100.0% | 7.79 | 2.78 |
| 2001 | 3.7% | 11.6% | 57.2% | 92.0% | 100.0% | 7.80 | 2.78 |
| 2002 | 4.1% | 12.3% | 57.2% | 91.5% | 100.0% | 7.41 | 2.66 |
| 2003 | 4.2% | 12.4% | 57.6% | 91.7% | 100.0% | 7.46 | 2.69 |
| 2004 | 4.0% | 12.3% | 58.2% | 92.3% | 100.0% | 7.66 | 2.78 |
| 2005 | 4.4% | 13.2% | 59.7% | 92.8% | 100.0% | 7.55 | 2.81 |
| 2006 | 3.9% | 12.1% | 58.4% | 92.5% | 100.0% | 7.80 | 2.83 |
| 2007 | 3.7% | 11.6% | 57.8% | 92.4% | 100.0% | 7.94 | 2.84 |
| **2008** | **3.5%** | **11.0%** | **56.1%** | **91.6%** | **100.0%** | **7.92** | **2.76** |
| 2009 | 3.9% | 11.6% | 54.7% | 90.0% | 100.0% | 7.24 | 2.50 |

**Table C.2 Cumulative frequencies life satisfaction and parameters beta distributions, Eurobarometer**

| Year | Version | On the whole, how satisfied are you with the life you lead? | | | | Parameters best fitting beta distribution | |
|---|---|---|---|---|---|---|---|
| | | Not at all satisfied | Not very satisfied | Fairly satisfied | Very satisfied | α | β |
| 1973 | ECS73 | 1.2% | 6.6% | 61.0% | 100.0% | 7.52 | 3.22 |
| 1975 | EB3 | 1.7% | 8.3% | 64.8% | 100.0% | 7.19 | 3.28 |
| 1975 | EB4 | 2.0% | 9.0% | 64.5% | 100.0% | 6.76 | 3.10 |
| 1976 | EB5 | 2.1% | 8.5% | 58.9% | 100.0% | 6.11 | 2.60 |
| 1976 | EB6 | 1.8% | 8.1% | 60.5% | 100.0% | 6.61 | 2.86 |
| 1977 | EB7 | 1.8% | 8.1% | 62.3% | 100.0% | 6.87 | 3.04 |
| 1977 | EB8 | 1.7% | 7.4% | 56.4% | 100.0% | 6.33 | 2.57 |
| 1978 | EB9 | 1.4% | 6.5% | 54.7% | 100.0% | 6.57 | 2.58 |
| 1978 | EB10 | 1.3% | 6.3% | 54.6% | 100.0% | 6.74 | 2.64 |
| 1979 | EB11 | 0.7% | 4.5% | 54.1% | 100.0% | 8.03 | 3.06 |
| 1980 | EB13 | 0.9% | 4.9% | 53.0% | 100.0% | 7.49 | 2.82 |
| 1981 | EB15 | 1.6% | 7.1% | 55.7% | 100.0% | 6.39 | 2.56 |
| 1982 | EB17 | 0.8% | 4.8% | 54.4% | 100.0% | 7.79 | 2.99 |
| 1982 | EB18 | 1.5% | 6.6% | 52.8% | 100.0% | 6.27 | 2.40 |
| 1983 | EB19 | 1.0% | 5.9% | 59.5% | 100.0% | 7.79 | 3.24 |
| 1983 | EB20 | 1.6% | 7.6% | 60.3% | 100.0% | 6.86 | 2.94 |
| 1984 | EB21 | 1.2% | 5.9% | 54.2% | 100.0% | 6.94 | 2.69 |
| 1984 | EB22 | 1.0% | 5.3% | 53.0% | 100.0% | 7.18 | 2.72 |
| 1985 | EB23 | 2.3% | 9.0% | 60.2% | 100.0% | 6.08 | 2.65 |
| 1985 | EB24 | 2.2% | 9.7% | 65.7% | 100.0% | 6.65 | 3.12 |
| 1986 | EB25 | 0.9% | 5.5% | 58.8% | 100.0% | 7.99 | 3.28 |
| 1986 | EB26 | 2.0% | 9.2% | 65.5% | 100.0% | 6.88 | 3.20 |
| 1987 | EB27 | 1.4% | 6.8% | 58.9% | 100.0% | 7.09 | 2.96 |
| 1987 | EB28 | 1.7% | 8.1% | 63.7% | 100.0% | 7.10 | 3.19 |

**Table C.2 (continued) Cumulative frequencies life satisfaction and parameters beta distributions, Eurobarometer**

| Year | Version | On the whole how satisfied are you with the life you lead? | | | | Parameters best fitting beta distribution | |
|---|---|---|---|---|---|---|---|
| | | Not at all satisfied | Not very satisfied | Fairly satisfied | Very satisfied | α | β |
| 1988 | EB29 | 1.6% | 7.0% | 55.8% | 100.0% | 6.48 | 2.60 |
| 1989 | EB31 | 0.8% | 4.7% | 51.5% | 100.0% | 7.41 | 2.73 |
| 1989 | EB31A | 1.0% | 5.2% | 50.0% | 100.0% | 6.79 | 2.46 |
| 1989 | EB32A | 1.0% | 5.7% | 58.2% | 100.0% | 7.70 | 3.14 |
| 1989 | EB32B | 1.1% | 5.8% | 56.0% | 100.0% | 7.30 | 2.90 |
| 1990 | EB33 | 1.2% | 5.7% | 52.1% | 100.0% | 6.74 | 2.53 |
| 1990 | EB340 | 0.8% | 5.2% | 58.7% | 100.0% | 8.16 | 3.33 |
| 1990 | EB341 | 1.6% | 7.1% | 55.2% | 100.0% | 6.32 | 2.52 |
| 1991 | EB350 | 0.9% | 4.7% | 49.4% | 100.0% | 7.11 | 2.54 |
| 1991 | EB36 | 1.0% | 5.3% | 52.6% | 100.0% | 7.07 | 2.66 |
| 1992 | EB370 | 1.4% | 6.0% | 50.7% | 100.0% | 6.28 | 2.32 |
| 1992 | EB371 | 1.1% | 5.5% | 53.9% | 100.0% | 7.17 | 2.76 |
| 1992 | EB380 | 1.3% | 5.8% | 49.9% | 100.0% | 6.33 | 2.30 |
| 1992 | EB381 | 1.7% | 6.5% | 48.1% | 100.0% | 5.61 | 2.00 |
| 1993 | EB390 | 1.0% | 5.3% | 52.5% | 100.0% | 7.06 | 2.66 |
| 1993 | EB40 | 1.3% | 6.1% | 54.5% | 100.0% | 6.84 | 2.67 |
| 1994 | EB410 | 1.5% | 6.5% | 53.1% | 100.0% | 6.33 | 2.44 |
| 1994 | EB42 | 1.5% | 7.3% | 60.9% | 100.0% | 7.10 | 3.06 |
| 1995 | EB431 | 1.7% | 6.9% | 52.3% | 100.0% | 6.02 | 2.29 |
| 1997 | EB471 | 1.5% | 6.5% | 53.7% | 100.0% | 6.43 | 2.49 |
| 1998 | EB49 | 1.0% | 5.6% | 56.0% | 100.0% | 7.43 | 2.94 |
| 1999 | EB520 | 0.8% | 5.2% | 62.3% | 100.0% | 8.83 | 3.76 |
| 1999 | EB521 | 0.6% | 4.0% | 54.7% | 100.0% | 8.68 | 3.31 |
| 2000 | EB530 | 1.4% | 7.8% | 68.0% | 100.0% | 8.09 | 3.81 |

**Table C.2 (continued) Cumulative frequencies life satisfaction and parameters beta distributions, Eurobarometer**

| Year | Version | On the whole how satisfied are you with the life you lead? | | | | Parameters best fitting beta distribution | |
|---|---|---|---|---|---|---|---|
| | | Not at all satisfied | Not very satisfied | Fairly satisfied | Very satisfied | α | β |
| 2000 | EB541 | 1.2% | 6.2% | 55.9% | 100.0% | 7.00 | 2.79 |
| 2001 | EB551 | 1.0% | 5.2% | 52.1% | 100.0% | 7.09 | 2.65 |
| 2001 | EB561 | 1.2% | 5.3% | 47.8% | 100.0% | 6.37 | 2.23 |
| 2001 | EB562 | 1.2% | 5.8% | 53.2% | 100.0% | 6.80 | 2.60 |
| 2002 | EB571 | 1.3% | 6.2% | 55.0% | 100.0% | 6.82 | 2.68 |
| 2002 | EB581 | 1.5% | 7.7% | 63.6% | 100.0% | 7.35 | 3.28 |
| 2003 | EB601 | 2.2% | 9.0% | 60.8% | 100.0% | 6.16 | 2.70 |
| 2004 | EB620 | 1.9% | 7.9% | 56.2% | 100.0% | 6.05 | 2.46 |
| 2005 | EB634 | 0.8% | 4.7% | 52.7% | 100.0% | 7.62 | 2.86 |
| 2005 | EB642 | 0.8% | 4.7% | 52.9% | 100.0% | 7.60 | 2.86 |
| 2006 | EB652 | 0.9% | 5.2% | 55.5% | 100.0% | 7.69 | 3.01 |
| 2006 | EB661 | 1.2% | 5.5% | 50.1% | 100.0% | 6.57 | 2.39 |
| 2007 | EB672 | 0.6% | 3.8% | 49.7% | 100.0% | 7.95 | 2.83 |
| 2007 | EB681 | 0.8% | 4.4% | 48.1% | 100.0% | 7.11 | 2.49 |
| **2008** | **EB692** | **0.7%** | **3.9%** | **48.5%** | **100.0%** | **7.92** | **2.76** |
| 2008 | EB701 | 0.4% | 2.8% | 46.2% | 100.0% | 8.64 | 2.89 |
| 2009 | EB711 | 0.6% | 3.9% | 50.0% | 100.0% | 7.93 | 2.84 |
| 2009 | EB712 | 0.8% | 4.5% | 49.0% | 100.0% | 7.17 | 2.54 |
| 2009 | EB724 | 0.9% | 4.7% | 49.5% | 100.0% | 7.06 | 2.53 |
| 2010 | EB734 | 1.0% | 5.3% | 52.0% | 100.0% | 7.03 | 2.63 |
| 2010 | EB742 | 0.8% | 4.4% | 47.5% | 100.0% | 7.02 | 2.43 |
| 2011 | EB753 | 0.8% | 4.5% | 49.1% | 100.0% | 7.18 | 2.55 |
| 2011 | EB754 | 1.2% | 5.2% | 45.8% | 100.0% | 6.08 | 2.06 |
| 2011 | EB763 | 0.5% | 3.2% | 46.7% | 100.0% | 8.15 | 2.76 |

**Table C.2 (continued) Cumulative frequencies life satisfaction and parameters beta distributions, Eurobarometer**

| Year | Version | On the whole how satisfied are you with the life you lead? | | | | Parameters best fitting beta distribution | |
|---|---|---|---|---|---|---|---|
| | | Not at all satisfied | Not very satisfied | Fairly satisfied | Very satisfied | α | β |
| 2012 | EB773 | 1.3% | 5.8% | 50.0% | 100.0% | 6.34 | 2.31 |
| 2012 | EB774 | 1.3% | 5.5% | 44.8% | 100.0% | 5.76 | 1.92 |
| 2012 | EB782 | 1.2% | 5.7% | 52.8% | 100.0% | 6.83 | 2.59 |

**Table C.3 Cumulative frequencies happiness and parameters beta distributions, CBS**

| Year | To what extent do you consider yourself a happy person? | | | | | Parameters best fitting beta distribution | |
|---|---|---|---|---|---|---|---|
| | Unhappy | Not very happy | Neither happy nor unhappy | Happy | Very happy | α | β |
| 1997 | 0.6% | 3.0% | 11.2% | 79.5% | 100.0% | 10.52 | 3.53 |
| 1998 | 0.5% | 2.5% | 10.2% | 79.6% | 100.0% | 11.19 | 3.70 |
| 1999 | 0.5% | 2.9% | 11.1% | 80.6% | 100.0% | 10.91 | 3.70 |
| 2000 | 0.5% | 2.7% | 10.5% | 79.0% | 100.0% | 10.82 | 3.57 |
| 2001 | 0.5% | 2.8% | 10.7% | 79.5% | 100.0% | 10.85 | 3.61 |
| 2002 | 0.7% | 3.3% | 11.7% | 79.0% | 100.0% | 10.09 | 3.39 |
| 2003 | 0.7% | 3.3% | 11.7% | 79.1% | 100.0% | 10.10 | 3.40 |
| 2004 | 0.7% | 3.2% | 11.5% | 79.3% | 100.0% | 10.27 | 3.45 |
| 2005 | 0.8% | 3.6% | 12.5% | 79.7% | 100.0% | 9.88 | 3.38 |
| 2006 | 0.5% | 2.8% | 10.7% | 79.7% | 100.0% | 10.91 | 3.64 |
| 2007 | 0.6% | 3.1% | 11.4% | 79.0% | 100.0% | 10.26 | 3.43 |
| **2008** | **0.6%** | **3.0%** | **10.8%** | **77.5%** | **100.0%** | **10.20** | **3.33** |
| 2009 | 0.7% | 3.0% | 10.7% | 75.7% | 100.0% | 9.77 | 3.13 |

**Table C.4 Cumulative frequencies happiness and parameters beta distributions, DHS**

| Year | Taking all together, to what extent do you think of yourself as a happy person? | | | | | Parameters best fitting beta distribution | |
|---|---|---|---|---|---|---|---|
| | Very unhappy | Unhappy | Neither happy nor unhappy | Happy | Very happy | α | β |
| 1993 | 0.1% | 1.1% | 16.8% | 84.4% | 100.0% | 12.97 | 4.28 |
| 1994 | 0.1% | 0.7% | 12.8% | 79.2% | 100.0% | 13.19 | 3.94 |
| 1995 | 0.1% | 1.2% | 17.0% | 83.2% | 100.0% | 12.28 | 4.02 |
| 1996 | 0.2% | 1.3% | 17.1% | 82.7% | 100.0% | 12.03 | 3.93 |
| 1997 | 0.2% | 1.4% | 17.3% | 82.1% | 100.0% | 11.63 | 3.79 |
| 1998 | 0.2% | 1.7% | 19.5% | 84.5% | 100.0% | 11.57 | 3.96 |
| 1999 | 0.2% | 1.4% | 17.3% | 81.7% | 100.0% | 11.47 | 3.72 |
| 2000 | 0.2% | 1.7% | 21.6% | 88.9% | 100.0% | 12.78 | 4.66 |
| 2001 | 0.2% | 1.4% | 17.0% | 81.7% | 100.0% | 11.66 | 3.77 |
| 2002 | 0.1% | 1.0% | 15.0% | 81.2% | 100.0% | 12.58 | 3.94 |
| 2003 | 0.2% | 1.6% | 16.9% | 78.8% | 100.0% | 10.60 | 3.34 |
| 2004 | 0.3% | 1.8% | 17.9% | 79.6% | 100.0% | 10.35 | 3.32 |
| 2005 | 0.2% | 1.2% | 15.8% | 79.8% | 100.0% | 11.58 | 3.62 |
| 2006 | 0.3% | 1.9% | 20.5% | 84.8% | 100.0% | 11.20 | 3.89 |
| 2007 | 0.2% | 1.2% | 15.4% | 78.7% | 100.0% | 11.35 | 3.49 |
| **2008** | **0.3%** | **2.0%** | **18.8%** | **80.1%** | **100.0%** | **10.20** | **3.33** |
| 2009 | 0.3% | 1.8% | 18.2% | 80.3% | 100.0% | 10.43 | 3.38 |
| 2010 | 0.3% | 1.7% | 17.9% | 80.5% | 100.0% | 10.69 | 3.46 |
| 2011 | 0.4% | 2.3% | 19.8% | 80.4% | 100.0% | 9.75 | 3.24 |
| 2012 | 0.4% | 2.0% | 17.7% | 76.8% | 100.0% | 9.51 | 2.97 |

# Appendix D Frequency distributions split-half experiment Statistics Netherlands

**Table D.1 Frequency distribution on verbal scale of question "To what extent do you consider yourself a happy person?"**

| Category | Sub-category | N | %DK/NA | Effective N | Unhappy | Not very happy | Neither happy nor unhappy | Happy | Very happy | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| Total | Total | 3,811 | 0.6 | 3,787 | 0.6 | 3.4 | 12.7 | 65.8 | 17.5 | 100.0 |
| Sex | Men | 1,869 | 1.7 | 1,837 | 0.7 | 3.6 | 13.4 | 64.3 | 17.9 | 100.0 |
| | Women | 1,974 | 2.8 | 1,918 | 0.5 | 3.3 | 11.9 | 67.2 | 17.1 | 100.0 |
| Age | 18-24 | 424 | 11.6 | 375 | 0.9 | 1.7 | 9.4 | 68.0 | 20.0 | 100.0 |
| | 25-34 | 566 | 8.0 | 521 | 0.2 | 1.8 | 10.1 | 64.1 | 23.8 | 100.0 |
| | 35-44 | 652 | 0.2 | 651 | 0.0 | 3.7 | 12.2 | 64.7 | 19.5 | 100.0 |
| | 45-54 | 759 | 1.2 | 750 | 1.0 | 4.7 | 13.0 | 65.2 | 16.0 | 100.0 |
| | 55-64 | 706 | 10.6 | 631 | 1.1 | 4.2 | 15.8 | 65.2 | 13.6 | 100.0 |
| | 65+ | 798 | 4.1 | 765 | 0.5 | 3.5 | 14.0 | 67.7 | 14.3 | 100.0 |
| Education | Low | 1,069 | 3.3 | 1,034 | 0.9 | 4.8 | 16.6 | 65.1 | 12.6 | 100.0 |
| | Middle | 1,619 | 0.2 | 1,616 | 0.5 | 3.0 | 11.7 | 67.7 | 17.1 | 100.0 |
| | High | 1,105 | 5.6 | 1,043 | 0.2 | 2.9 | 9.5 | 64.0 | 23.4 | 100.0 |

Source: Statistics Netherlands, Social Cohesion Survey 2012

**Table D.2 Frequency distribution on numerical scale of question "To what extent do you consider yourself a happy person?"**

| Category | Sub-category | N | %DK/NA | Effective N | Completely unhappy | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Completely happy | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total | Total | 3,701 | 0.2 | 3,695 | 0.3 | 0.2 | 0.2 | 0.6 | 0.8 | 3.3 | 5.7 | 23.3 | 43.5 | 16.1 | 6.2 | 100.0 |
| Sex | Men | 1,808 | 1.2 | 1,787 | 0.3 | 0.2 | 0.2 | 0.5 | 0.9 | 3.0 | 5.4 | 22.5 | 44.3 | 17.3 | 5.4 | 100.0 |
| | Women | 1,914 | 1.5 | 1,886 | 0.2 | 0.1 | 0.2 | 0.6 | 0.7 | 3.6 | 5.9 | 24.0 | 42.7 | 15.0 | 6.9 | 100.0 |
| Age | 18-24 | 405 | 11.9 | 357 | 0.0 | 0.2 | 0.4 | 0.2 | 0.9 | 1.0 | 6.9 | 24.6 | 41.3 | 16.1 | 8.3 | 100.0 |
| | 25-34 | 596 | 6.5 | 557 | 0.0 | 0.0 | 0.5 | 0.5 | 0.9 | 3.3 | 5.0 | 21.0 | 45.8 | 17.5 | 5.5 | 100.0 |
| | 35-44 | 664 | 1.8 | 652 | 0.7 | 0.3 | 0.0 | 1.1 | 0.3 | 3.3 | 5.5 | 22.0 | 42.1 | 18.8 | 5.9 | 100.0 |
| | 45-54 | 688 | 1.3 | 679 | 0.1 | 0.1 | 0.1 | 0.6 | 1.4 | 4.3 | 5.1 | 23.7 | 44.8 | 14.1 | 5.6 | 100.0 |
| | 55-64 | 670 | 11.6 | 592 | 0.5 | 0.3 | 0.0 | 0.2 | 0.6 | 5.1 | 5.6 | 23.3 | 43.4 | 15.6 | 5.3 | 100.0 |
| | 65+ | 777 | 2.3 | 759 | 0.2 | 0.1 | 0.3 | 0.5 | 0.8 | 2.2 | 6.3 | 24.8 | 42.8 | 14.9 | 7.0 | 100.0 |
| Education | Low | 1,064 | 5.2 | 1,009 | 0.9 | 0.3 | 0.1 | 0.9 | 0.9 | 4.1 | 6.4 | 27.5 | 39.2 | 12.5 | 7.3 | 100.0 |
| | Middle | 1,577 | 0.9 | 1,563 | 0.0 | 0.0 | 0.3 | 0.5 | 0.8 | 3.1 | 6.2 | 23.3 | 44.2 | 15.3 | 6.4 | 100.0 |
| | High | 1,062 | 5.5 | 1,004 | 0.1 | 0.0 | 0.1 | 0.4 | 0.5 | 2.6 | 4.3 | 19.0 | 47.2 | 21.4 | 4.3 | 100.0 |

Source: Statistics Netherlands, Social Cohesion Survey 2012

**Table D.3 Frequency distribution on verbal scale of question "To what extent are you satisfied with the life you currently lead?"**

| Category | Sub-category | N | %DK/NA | Effective N | Not very satisfied | Fairly satisfied | Satisfied | Very satisfied | Extraordinarily satisfied | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| Total | Total | 3,819 | 0.6 | 3,796 | 5.9 | 11.1 | 44.6 | 31.2 | 7.2 | 100.0 |
| Sex | Men | 1,873 | 1.8 | 1,840 | 5.9 | 11.5 | 43.5 | 30.9 | 8.3 | 100.0 |
| | Women | 1,979 | 2.8 | 1,923 | 6.0 | 10.7 | 45.7 | 31.6 | 6.1 | 100.0 |
| Age | 18-24 | 425 | 11.5 | 376 | 5.3 | 8.6 | 41.1 | 37.0 | 7.9 | 100.0 |
| | 25-34 | 567 | 7.9 | 522 | 3.3 | 11.9 | 37.3 | 39.5 | 7.9 | 100.0 |
| | 35-44 | 654 | 0.2 | 653 | 6.5 | 10.2 | 41.1 | 34.0 | 8.2 | 100.0 |
| | 45-54 | 759 | 1.1 | 751 | 7.1 | 11.4 | 45.9 | 28.7 | 6.8 | 100.0 |
| | 55-64 | 707 | 10.6 | 632 | 9.9 | 11.8 | 45.8 | 26.8 | 5.6 | 100.0 |
| | 65+ | 801 | 4.1 | 768 | 3.2 | 11.7 | 52.6 | 25.6 | 6.8 | 100.0 |
| Education | Low | 1,071 | 3.3 | 1,036 | 7.9 | 14.2 | 50.3 | 22.9 | 4.6 | 100.0 |
| | Middle | 1,620 | 0.1 | 1,618 | 5.1 | 9.6 | 46.0 | 32.8 | 6.4 | 100.0 |
| | High | 1,109 | 5.6 | 1,047 | 5.0 | 9.7 | 36.4 | 38.1 | 10.9 | 100.0 |

Source: Statistics Netherlands, Social Cohesion Survey 2012

**Table D.4 Frequency distribution on numerical scale of question "To what extent are you satisfied with the life you currently lead?"**

| Category | Sub-category | N | %DK/NA | Effective N | Completely dissatisfied | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Completely satisfied | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total | Total | 3,694 | 0.3 | 3,682 | 0.3 | 0.3 | 0.3 | 0.9 | 1.0 | 4.1 | 8.0 | 24.8 | 39.4 | 15.1 | 5.7 | 100.0 |
| Sex | Men | 1,804 | 1.1 | 1,785 | 0.4 | 0.4 | 0.4 | 0.9 | 1.3 | 4.0 | 6.9 | 23.8 | 40.4 | 16.8 | 4.9 | 100.0 |
| | Women | 1,909 | 1.6 | 1,878 | 0.3 | 0.2 | 0.3 | 0.9 | 0.7 | 4.2 | 9.1 | 25.8 | 38.4 | 13.5 | 6.5 | 100.0 |
| Age | 18-24 | 402 | 11.4 | 356 | 0.0 | 0.0 | 0.6 | 0.9 | 0.5 | 4.3 | 7.9 | 26.3 | 37.3 | 14.4 | 7.9 | 100.0 |
| | 25-34 | 592 | 6.3 | 555 | 0.0 | 0.5 | 0.2 | 1.2 | 2.1 | 3.4 | 9.9 | 26.4 | 37.5 | 14.7 | 4.1 | 100.0 |
| | 35-44 | 662 | 1.7 | 651 | 0.7 | 0.3 | 0.0 | 1.3 | 0.9 | 5.5 | 7.3 | 24.9 | 37.2 | 16.6 | 5.5 | 100.0 |
| | 45-54 | 686 | 1.5 | 676 | 0.1 | 0.3 | 0.7 | 0.5 | 1.5 | 3.9 | 8.2 | 27.3 | 36.9 | 15.0 | 5.6 | 100.0 |
| | 55-64 | 671 | 11.6 | 593 | 1.0 | 0.0 | 0.4 | 0.8 | 0.3 | 4.5 | 7.3 | 22.6 | 42.9 | 15.1 | 5.2 | 100.0 |
| | 65+ | 775 | 2.5 | 756 | 0.2 | 0.3 | 0.2 | 0.8 | 0.6 | 3.2 | 7.6 | 22.4 | 43.4 | 14.8 | 6.5 | 100.0 |
| Education | Low | 1,063 | 5.0 | 1,010 | 1.1 | 0.5 | 0.2 | 1.6 | 1.2 | 4.4 | 9.1 | 24.9 | 36.6 | 12.7 | 7.8 | 100.0 |
| | Middle | 1,574 | 1.1 | 1,557 | 0.0 | 0.1 | 0.5 | 0.9 | 0.8 | 4.8 | 8.4 | 25.3 | 40.8 | 13.4 | 5.0 | 100.0 |
| | High | 1,058 | 5.6 | 999 | 0.1 | 0.1 | 0.2 | 0.1 | 0.9 | 2.5 | 6.4 | 24.5 | 40.1 | 20.6 | 4.5 | 100.0 |

Source: Statistics Netherlands, Social Cohesion Survey 2012

# Appendix E Frequency distributions Eurobarometer 76.3 2011 and 76.2 2011

**Table E.1 Frequency distribution Northern Europe on verbal scale of question "On the whole how satisfied are you with the life you lead?"**

| Category | Sub-category | N | %DK/NA | Effective N | Not at all satisfied | Not very satisfied | Fairly satisfied | Very satisfied | Total |
|---|---|---|---|---|---|---|---|---|---|
| Total | Total | 9,940 | 0.3 | 9,914 | 2.9 | 11.2 | 56.0 | 29.9 | 100.0 |
| Sex | Men | 5,769 | 0.2 | 4,759 | 2.9 | 10.5 | 56.9 | 29.7 | 100.0 |
| | Women | 5,171 | 0.3 | 5,155 | 2.9 | 11.8 | 55.1 | 30.2 | 100.0 |
| Age | 18-24 | 1,123 | 0.2 | 1,121 | 1.7 | 9.4 | 51.3 | 37.6 | 100.0 |
| | 25-34 | 1,520 | 0.1 | 1,519 | 3.3 | 12.2 | 56.7 | 27.7 | 100.0 |
| | 35-44 | 1,715 | 0.1 | 1,713 | 2.6 | 12.6 | 57.1 | 27.7 | 100.0 |
| | 45-54 | 1,830 | 0.1 | 1,828 | 4.2 | 13.5 | 57.5 | 24.8 | 100.0 |
| | 55-64 | 1,514 | 0.1 | 1,512 | 3.6 | 11.0 | 54.2 | 31.2 | 100.0 |
| | 65+ | 2,237 | 0.8 | 2,219 | 2.0 | 8.4 | 56.9 | 32.7 | 100.0 |
| Education | Low | 1,771 | 0.6 | 1,761 | 3.8 | 15.2 | 57.5 | 23.5 | 100.0 |
| | Middle | 4,252 | 0.2 | 4,244 | 3.9 | 13.1 | 55.6 | 27.3 | 100.0 |
| | High | 3,165 | 0.1 | 3,162 | 1.6 | 7.4 | 55.9 | 35.1 | 100.0 |

Source: Eurobarometer 76.3 2011

**Table E.2 Frequency distribution Northern Europe on numerical scale of question "On the whole how satisfied are you with the life you lead?"**

| Category | Sub-category | N | %DK/NA | Effective N | Very dissatisfied | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Very satisfied | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total | Total | 9,989 | 0.2 | 9,972 | 0.8 | 1.2 | 2.0 | 1.7 | 7.4 | 6.4 | 14.8 | 27.5 | 18.0 | 20.2 | 100.0 |
| Sex | Men | 4,840 | 0.1 | 4,833 | 0.9 | 1.1 | 2.2 | 1.8 | 6.6 | 6.4 | 14.9 | 28.1 | 19.3 | 18.7 | 100.0 |
| | Women | 5,149 | 0.2 | 5,139 | 0.7 | 1.3 | 1.8 | 1.7 | 8.1 | 6.4 | 14.7 | 27.0 | 16.7 | 21.7 | 100.0 |
| Age | 18-24 | 1,135 | 0.4 | 1,131 | 0.7 | 1.1 | 1.4 | 0.8 | 3.9 | 4.5 | 14.8 | 25.0 | 20.1 | 27.8 | 100.0 |
| | 25-34 | 1,511 | 0.1 | 1,510 | 0.9 | 1.1 | 2.6 | 1.5 | 5.9 | 4.8 | 15.9 | 25.2 | 21.3 | 21.0 | 100.0 |
| | 35-44 | 1,725 | 0.1 | 1,724 | 1.3 | 1.3 | 2.1 | 1.3 | 7.8 | 5.4 | 16.1 | 29.7 | 17.7 | 17.2 | 100.0 |
| | 45-54 | 1,853 | 0.2 | 1,849 | 0.5 | 1.7 | 1.6 | 2.3 | 9.2 | 8.7 | 13.0 | 29.1 | 16.5 | 17.3 | 100.0 |
| | 55-64 | 1,519 | 0.3 | 1,514 | 1.3 | 1.6 | 2.2 | 2.0 | 6.9 | 6.5 | 14.3 | 26.5 | 19.0 | 19.7 | 100.0 |
| | 65+ | 2,248 | 0.1 | 2,245 | 0.4 | 0.5 | 2.0 | 1.9 | 8.7 | 7.3 | 14.7 | 28.0 | 15.4 | 21.0 | 100.0 |
| Education | Low | 1,869 | 0.2 | 1,866 | 1.1 | 1.9 | 2.6 | 3.1 | 10.6 | 8.6 | 14.6 | 25.8 | 13.0 | 18.8 | 100.0 |
| | Middle | 4,356 | 0.1 | 4,353 | 1.0 | 1.4 | 2.2 | 1.4 | 8.8 | 6.8 | 15.2 | 26.9 | 16.4 | 19.8 | 100.0 |
| | High | 3,088 | 0.4 | 3,077 | 0.5 | 0.5 | 1.5 | 1.2 | 4.0 | 4.6 | 13.9 | 30.1 | 23.0 | 20.9 | 100.0 |

Source: Eurobarometer 76.2 2011

**Table E.3 Frequency distribution Southern Europe on verbal scale of question "On the whole how satisfied are you with the life you lead?"**

| Category | Sub-category | N | %DK/NA | Effective N | Not at all satisfied | Not very satisfied | Fairly satisfied | Very satisfied | Total |
|---|---|---|---|---|---|---|---|---|---|
| Total | Total | 5,037 | 0.6 | 5,009 | 9.1 | 28.0 | 57.3 | 5.6 | 100.0 |
| Sex | Men | 2,426 | 0.7 | 2,408 | 8.4 | 26.7 | 58.8 | 6.0 | 100.0 |
| | Women | 2,611 | 0.4 | 2,601 | 9.7 | 29.2 | 55.9 | 5.2 | 100.0 |
| Age | 18-24 | 508 | 0.8 | 504 | 6.3 | 24.4 | 61.1 | 8.1 | 100.0 |
| | 25-34 | 826 | 0.6 | 821 | 7.1 | 30.3 | 56.0 | 6.6 | 100.0 |
| | 35-44 | 1,059 | 0.4 | 1,055 | 9.2 | 30.1 | 54.7 | 6.0 | 100.0 |
| | 45-54 | 814 | 0.0 | 814 | 8.7 | 27.8 | 57.7 | 5.8 | 100.0 |
| | 55-64 | 705 | 0.3 | 703 | 10.0 | 27.0 | 57.5 | 5.5 | 100.0 |
| | 65+ | 1,123 | 1.1 | 1,111 | 11.3 | 26.8 | 58.7 | 3.2 | 100.0 |
| Education | Low | 1,757 | 0.5 | 1,748 | 12.5 | 30.4 | 53.6 | 3.5 | 100.0 |
| | Middle | 1,807 | 0.3 | 1,802 | 7.7 | 28.5 | 57.5 | 6.2 | 100.0 |
| | High | 968 | 0.2 | 966 | 5.5 | 21.2 | 64.7 | 8.6 | 100.0 |

Source: Eurobarometer 76.3 2011

**Table E.4 Frequency distribution Southern Europe on numerical scale of question "On the whole how satisfied are you with the life you lead?"**

| Category | Sub-category | N | %DK/NA | Effective N | Very dissatisfied | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Very satisfied | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total | Total | 5,088 | 0.1 | 5,084 | 0.9 | 1.1 | 2.3 | 3.9 | 10.4 | 15.5 | 22.6 | 26.3 | 10.2 | 6.9 | 100.0 |
| Sex | Men | 2,464 | 0.1 | 2,462 | 0.7 | 1.5 | 2.6 | 4.1 | 10.4 | 15.4 | 22.0 | 27.3 | 9.7 | 6.5 | 100.0 |
| | Women | 2,624 | 0.1 | 2,622 | 1.1 | 0.8 | 2.1 | 3.8 | 10.4 | 15.6 | 23.1 | 25.2 | 10.7 | 7.2 | 100.0 |
| Age | 18-24 | 534 | 0.0 | 534 | 0.2 | 1.3 | 0.4 | 3.4 | 8.2 | 14.0 | 20.8 | 28.3 | 13.7 | 9.7 | 100.0 |
| | 25-34 | 809 | 0.0 | 809 | 1.2 | 0.7 | 2.1 | 4.1 | 8.2 | 13.8 | 20.9 | 26.0 | 15.3 | 7.7 | 100.0 |
| | 35-44 | 1,020 | 0.0 | 1,020 | 1.5 | 0.9 | 1.9 | 1.9 | 9.2 | 14.3 | 22.1 | 28.1 | 11.2 | 9.0 | 100.0 |
| | 45-54 | 882 | 0.0 | 882 | 0.3 | 1.6 | 2.9 | 3.5 | 11.2 | 13.5 | 24.8 | 28.1 | 8.3 | 5.7 | 100.0 |
| | 55-64 | 709 | 0.0 | 709 | 1.0 | 1.0 | 2.3 | 5.1 | 9.6 | 14.0 | 26.1 | 26.0 | 7.9 | 7.2 | 100.0 |
| | 65+ | 1,129 | 0.4 | 1,125 | 0.9 | 1.3 | 3.2 | 5.3 | 14.0 | 21.0 | 21.2 | 22.6 | 6.8 | 3.6 | 100.0 |
| Education | Low | 1,793 | 0.1 | 1,791 | 1.5 | 1.5 | 3.5 | 5.7 | 12.8 | 18.1 | 22.5 | 22.1 | 7.0 | 5.2 | 100.0 |
| | Middle | 1,721 | 0.0 | 1,721 | 0.3 | 0.6 | 1.2 | 3.5 | 9.3 | 15.6 | 24.5 | 27.1 | 9.9 | 8.0 | 100.0 |
| | High | 1,093 | 0.0 | 1,093 | 0.6 | 1.3 | 2.6 | 2.7 | 6.9 | 11.0 | 20.5 | 31.5 | 15.1 | 8.0 | 100.0 |

Source: Eurobarometer 76.2 2011

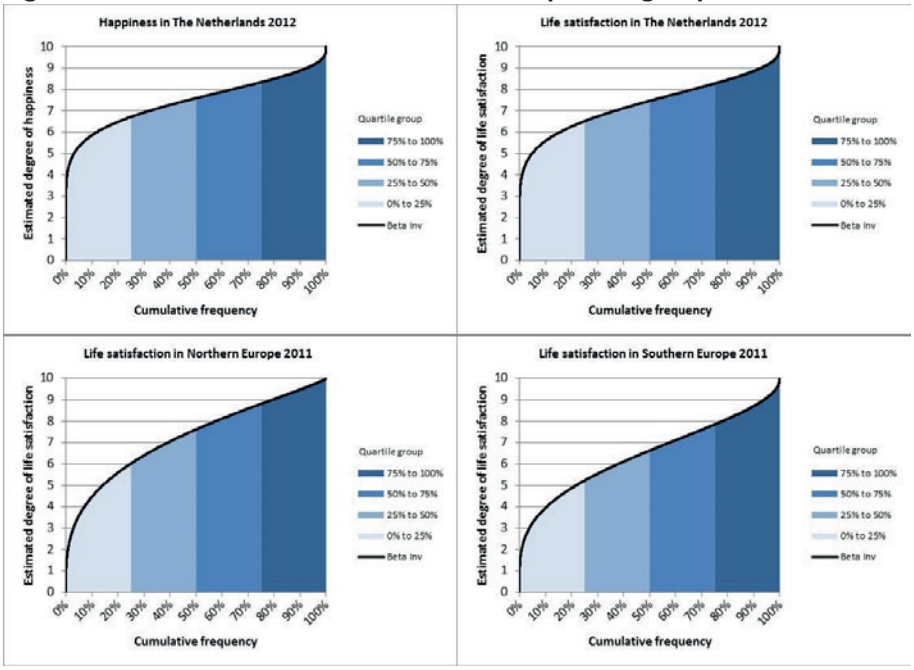
# Appendix F Means for subgroups derived from a beta distribution

Suppose that the parameters $\alpha$ and $\beta$ of the estimated beta distribution for happiness or life satisfaction in a population are given. The mean happiness or life satisfaction within subgroups of this population can accordingly be estimated by making use of the inverse beta distribution. The parameters of the best fitting beta distribution for life satisfaction in The Netherlands in 2012 that we estimated for the measurement in which a numerical scale was used are: $\alpha = 8.31$ and $\beta = 3.05$. The value of the inverse of this beta distribution for the 25th percentile point is equal to 6.5, meaning that the estimated value of life satisfaction is 6.5 or lower for the subgroup of the 25% least satisfied people in the population. In a similar way the inverse of the beta distribution at the 75th percentile point is equal to 8.3, meaning that the value of life satisfaction for the subgroup of the 25% most satisfied people is at least 8.3.

The parameters of the other reference distributions used in Ch. 9 are $\alpha = 9.60$ and $\beta = 3.28$ for happiness in The Netherlands, $\alpha = 3.21$ and $\beta = 1.21$ for life satisfaction in Northern Europe, and $\alpha = 3.95$ and $\beta = 2.17$ for life satisfaction in Southern Europe.

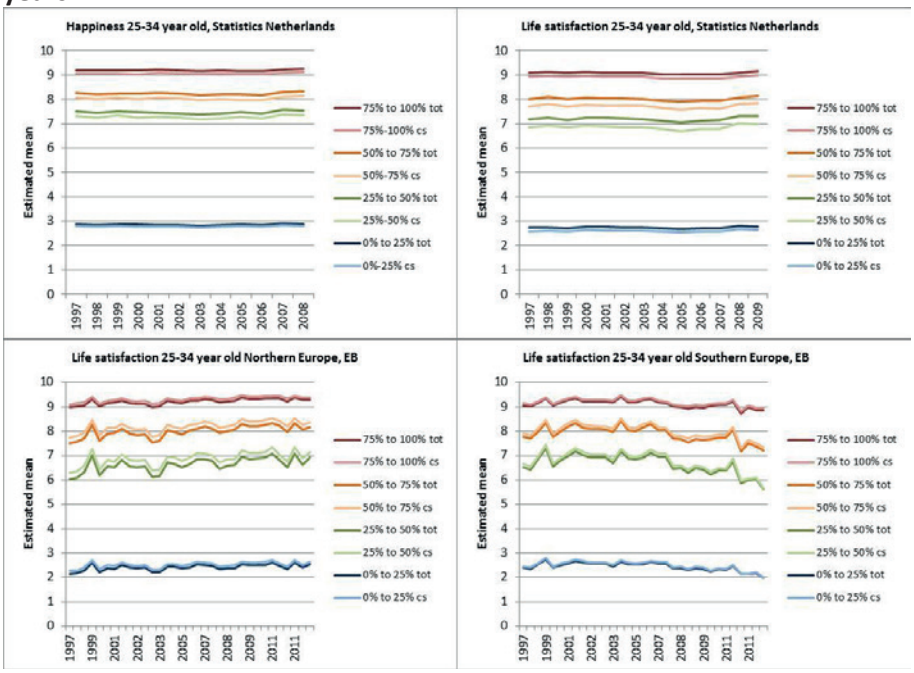
The inverse beta distributions of the four reference distributions used in Ch. 9 are shown in Fig. F.1. In each graph of Fig. F.1 we have also depicted four quartile groups. The most left of the quartile groups in each graph represents the least happy or satisfied people in the population and the most right quartile group represents the most happy or satisfied people. It can, for example, be seen in Fig. F.1 that the degree of life satisfaction in 2012 within the second quartile group in The Netherlands varies from 6.5 to 7.5. In the corresponding subgroup in Southern Europe the degree of life satisfaction varies from 5.2 to 6.6.

**Figure F.1 The inverse beta distribution and quartile groups**




We calculated the value of the inverse beta distribution for 25.000 equidistant points in each quartile group. The average of this large number of values for a quartile group is an estimate of the mean happiness or life satisfaction in this group. The estimates of the mean life satisfaction in The Netherlands calculated in this way for the four quartile groups are equal to respectively 5.6, 7.0, 7.9 and 8.8. We have estimated the means for the quartile groups of all the demographic categories we have distinguished in this paper, based on both the beta distribution estimated using category-specific reference boundaries and by using the reference boundaries for the general population. We did not find noteworthy differences in the evolution of the trends depending on the reference boundaries used. This is illustrated for the age group from 25-34 years old in Fig. F.2. We used the abbreviation ‘cs’ to denote the use of the category-specific boundaries to estimate the means and the abbreviation ‘tot’ to denote the use of the boundaries for the general population to estimate the means for the boundaries.

**Figure F.2 Means of quartile groups of the population aged from 25-34 years**



# Appendix G Can a piecewise linear distribution be used instead of a beta distribution?

## G.1 An attempt to make things easier

The need to estimate a beta distribution in the application of the Continuum Approach might be considered as a drawback of the Reference Distribution Method, since the technique needed for this estimation is not part of the standard tooling most researchers in this area have at their disposal. Things would be easier if use could be made of a reference distribution which could be derived in a more straightforward way than performing a maximum likelihood estimation as is required to obtain a beta distribution based on a discrete primary scale distribution. Within the Continuum Approach, Kalmijn (2010, p. 129) introduced not only a fully continuous model for the distribution in the population on the basis of the beta distribution, but also a semi-continuous model.

In this semi-continuous model it is assumed that each response option of a discrete primary scale corresponds to an interval in which all values the happiness feeling can adopt are equally likely. In other words, within each bounded interval representing a response option, happiness is uniformly distributed with its own constant probability density. All intervals together span the continuum from 0 to 10. The cumulative distribution of this semi-continuous model is piecewise linear, consisting of a straight diagonal line for each response option of the primary scale as can be seen in Fig. G.1 in Sec. G.2.

One advantage of this semi-continuous model over the use of a beta distribution, however, is that it is very easy to derive once the boundaries of the interval for each response option are known, as we have remarked above. The means obtained by application of the semi-continuous model are equal to the means that would be obtained if the Weighted Average Approach was applied to the mid-interval values of the intervals (Kalmijn, 2010, p. 205). In the Dutch Scale Interval Study we found that if the means after scale transformation are based on the Weighted Average Approach they are nearly the same as the means which would be obtained by application of the fully continuous model as can be seen in Tab. 10 in Sec. 6.4.

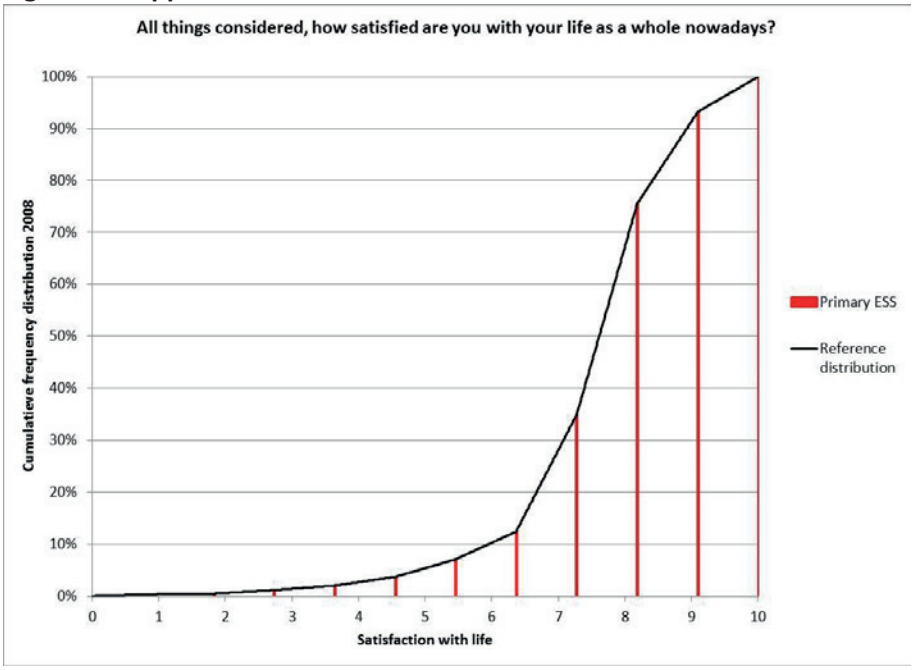
This observation in Tab. 10 gave rise to the idea to investigate whether replacing the fully continuous model (the beta distribution) by the semi-continuous model as described above in the Reference Distribution Method, would give similar results as obtained for the pooling of time

series which we described in Ch. 10. In that case it would be easier to apply the Reference Distribution Method.

## G.2 The Reference Distribution Method and the semi-continuous model

If we want to avoid having to estimate beta distributions and instead combine the Reference Distribution Model with the semi-continuous model, we first need to derive a piecewise linear reference distribution. In analogy with what we did in Sec. 6.4, we derive a reference distribution from the life satisfaction item taken from the European Social Survey (ESS). Similar to what we presented in Fig. 16, the cumulative frequency distribution for the ESS scale in 2008 and the piecewise linear distribution based on it are shown in Fig. G.1.

**Figure G.1 Application of the semi-continuous model to the ESS item**

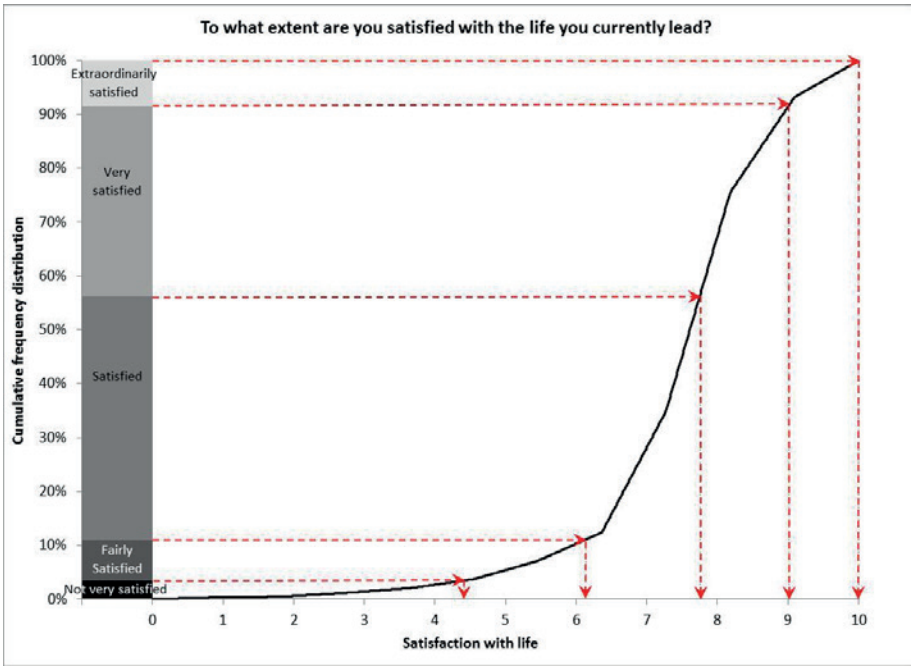



Contrary to what can be noticed in Fig. 16 for the reference distribution estimated by application of the fully continuous model, there are no gaps between the piecewise linear reference distribution which follows from the application of the semi-continuous model and the cumulative frequency

distribution for the ESS. The goodness-of-fit in this model is perfect, though with zero degrees of freedom.

In the same way as we illustrated by Fig. 18 in Sec. 7.2 we can derive reference boundaries from the piecewise linear reference distribution shown in Fig. G.1. The result for the life satisfaction item of CBS and the frequency distribution in 2008 is shown in Fig. G.2.

**Figure G.2 Deriving reference boundaries from a piecewise linear reference distribution**




Using the a piecewise linear distribution as a reference, we find reference boundaries for the CBS item that are equal to consecutively, 4.41, 6.13, 7.75, 9.01 and 10.0. The last three of these reference boundaries are almost equal to the reference boundaries to be found by making use of the beta distribution that fits best to the ESS item as a reference. These latter boundaries are equal to consecutively 4.78, 5.73, 7.77, 9.04 and 10.00. The reference boundaries for the two response options at the lower end of the scale differ considerably with respect to the reference distribution used.

We derived reference boundaries from piecewise linear reference distributions based on the ESS items on happiness and life satisfaction for the same items as in Sec. 10.3.2 according to the conversion scheme shown

in Fig. 30 in that section. The reference boundaries we have found this way are presented in Tab. G.1 and G.2.

The reference boundaries define the intervals representing the response options of each item on the 0-10 continuum. If we apply the Weighted Average Approach to the mid-interval values of these intervals for each frequency distribution in the time series of the corresponding item, we find the same population means that would be found if the semi-continuous model had been applied. The time series of the population means are depicted in Fig. G.3 and G.4. A glance at these two figures makes clear that it is un-admissible to average the means per year of measurements. Making use of a piecewise linear reference distribution turns out not to be a suitable way to make survey results comparable. Differently from using a best fitting beta distribution as a reference, using a piecewise linear distribution as a reference fails to overcome discontinuities in time series. The question “Can a piecewise linear distribution be used instead of a beta distribution?” must thus be answered negatively.

**Table G.1 Reference boundaries derived from piece wise linear reference distributions for happiness items**

| | Round 1 | | | | Round 2 | | | | | Round 3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Reference distribution** | **ESS 2008** | | | **ESS 2006** | **CBS 1997** | | **DHS 1993** | | **WVS 1981** | **EB 1986** | |
| **Response option** | ***CBS 5p-v*** | ***SCP 5p-v*** | ***DHS 5p-v*** | ***WVS 4p-v*** | ***CBS 5p-v*** | ***SCP 5p-v*** | ***CBS 5p-v*** | ***SCP 5p-v*** | ***EB 3p-v*** | ***CBS 5p-v*** | ***SCP 5p-v*** |
| - Very happy | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 |
| - Happy | 8.3 | 8.7 | 8.4 | | 8.0 | 8.0 | 7.7 | 7.7 | | 8.4 | 8.4 |
| - Quite happy | | | | 7.8 | | | | | | | |
| - Pretty happy | | | | | | | | | 7.0 | | |
| - Neither happy nor unhappy | 6.4 | 6.4 | 6.7 | | 6.1 | 6.0 | 5.4 | 5.4 | | 3.9 | 3.9 |
| - Not too happy | | | | | | | | | 5.2 | | |
| - Not very happy | 4.8 | 4.8 | | 5.4 | 4.1 | 3.9 | 4.2 | 4.2 | | 1.4 | 1.4 |
| - Unhappy | 2.8 | 2.6 | 4.3 | | 1.0 | 2.0 | 2.4 | 2.5 | | 0.2 | 0.2 |
| - Very unhappy | | | 2.2 | | | | | | | | |
| - Not at all happy | | | | 2.1 | | | | | | | |

**Table G.2 Reference boundaries derived from piece wise linear reference distributions for life satisfaction items**

| | Round 1 | | | | Round 2 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Reference distribution** | **ESS 2008** | | | | **CBS 1997** | | **EB 1993** | | **EB 1986** | **EB 1977** | |
| **Response option** | ***EB 4p-v*** | ***CBS 5p-v*** | ***WVS 10p-n*** | ***SCP 10p-n*** | ***SCP 5p-v*** | ***CBS 5p-v*** | ***CBS 5p-v*** | ***SCP 5p-v*** | ***CBS 5p-v*** | ***CBS 5p-v*** | ***SCP 5p-v*** |
| - Extraordinarily satisfied | | 10.0 | | | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 |
| - Very satisfied | 10.0 | 9.0 | | | 8.0 | 7.7 | 9.1 | 9.2 | 9.4 | 9.5 | 9.5 |
| - Satisfied | 7.6 | 7.7 | | | 6.0 | 5.6 | 7.1 | 7.2 | 7.5 | 7.9 | 7.8 |
| - Fairly satisfied | 4.2 | 6.1 | | | 4.0 | 4.1 | 5.3 | 5.4 | 5.4 | 5.6 | 5.6 |
| - Not very satisfied | | 4.4 | | | 2.0 | 2.4 | 4.0 | 4.1 | 4.0 | 4.7 | 4.6 |
| - Not at all satisfied | 1.4 | | | | | | | | | | |
| - 10[25] | | | 10.0 | 10.0 | | | | | | | |
| 9 | | | 9.0 | 9.0 | | | | | | | |
| 8 | | | 8.0 | 8.2 | | | | | | | |
| 7 | | | 6.8 | 7.1 | | | | | | | |
| 6 | | | 5.5 | 5.9 | | | | | | | |
| 5 | | | 4.5 | 4.6 | | | | | | | |
| 4 | | | 3.4 | 3.2 | | | | | | | |
| 3 | | | 2.4 | 2.0 | | | | | | | |
| 2 | | | 0.8 | 1.6 | | | | | | | |
| 1[26] | | | 0.4 | 0.6 | | | | | | | |

[25] In the WVS labelled with 'Satisfied' and in the SCP survey labelled with 'Completely satisfied'.
[26] In the WVS labelled with 'Dissatisfied' and in the SCP survey labelled with 'Completely dissatisfied'.

**Figure G.3 Converted time series for happiness in The Netherlands**

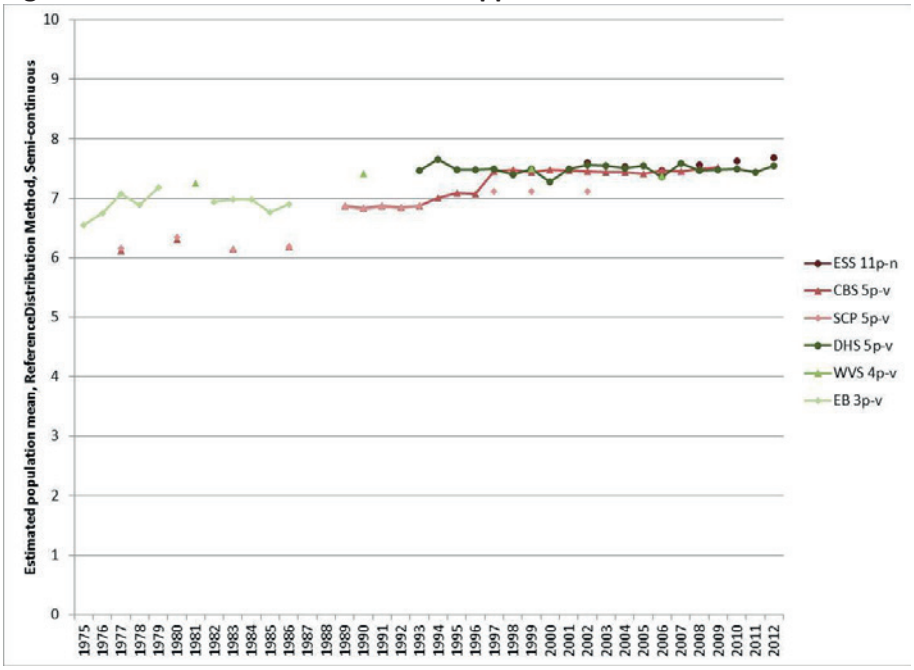
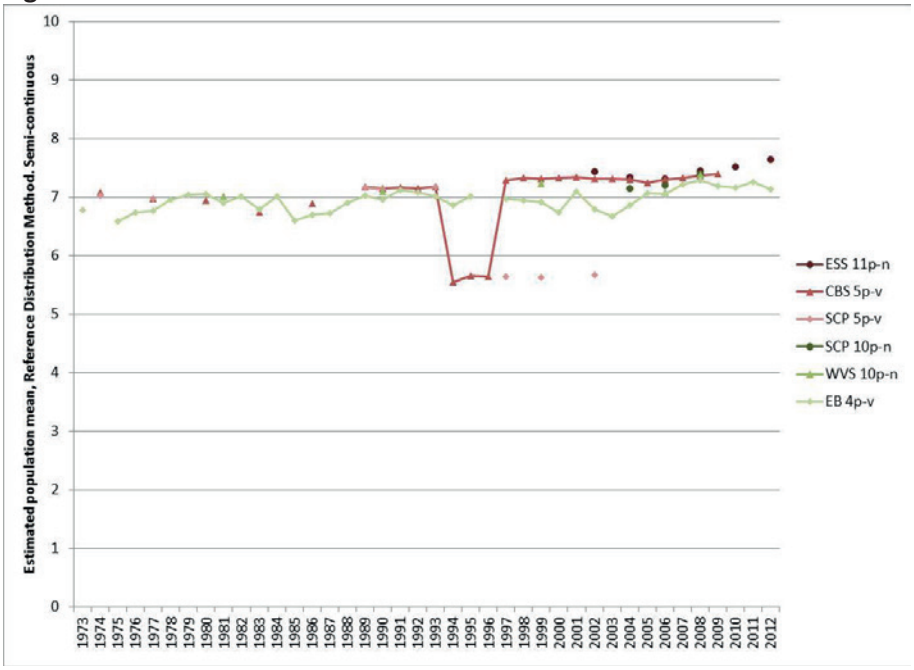



**Figure G.4 Converted time series for life satisfaction in The Netherlands**



The eye-catching low values for the mean satisfaction with life in the period 1994-1997 shown in Fig. G.4, are entirely due to the effect of the difference in survey mode during this period compared to earlier and later years. We have given the given the frequency distributions on the primary scale for life satisfaction in the period from 1993 to 1997 in Tab. G.3 to make this clear. The mode of surveying was changed in 1994 and in 1997. We recall from Sec. 10.3.1 that in the latter year the change was introduced using a split-half measurement with half of the respondents being required to fill in a paper-and-pencil questionnaire and the other half being interviewed.

**Table G.3 Frequency distribution life satisfaction CBS, 1993-1997**

| | To what extent are you satisfied with the life you currently lead? | | | | |
|---|---|---|---|---|---|
| Year | Not very satisfied | Fairly satisfied | Satisfied | Very satisfied | Extraordina-rily satisfied |
| 1993 | 3.85% | 8.38% | 33.72% | 38.25% | 15.81% |
| 1994 | 6.20% | 11.78% | 34.80% | 35.42% | 11.81% |
| 1995 | 4.84% | 11.76% | 34.76% | 35.67% | 12.97% |
| 1996 | 5.07% | 10.13% | 36.79% | 35.97% | 12.05% |
| 1997 | 5.10% | 8.95% | 36.71% | 36.87% | 12.37% |
| 1997 | 3.44% | 9.26% | 46.71% | 32.59% | 8.00% |

It can be seen in Tab. G.3 that the frequency by which the option 'Satisfied' was chosen in 1997 after the mode change was 10 percentage points higher than before the mode change. The options 'Very satisfied' and 'Extraordinarily satisfied' were chosen much more frequently in 1993 than in 1994. These differences in frequency distributions cause the low estimated population means in the period 1994 – 1997 and make also clear that the combination of the Reference Distribution Method with the semi-continuous model falls short to deal with mode changes. The mode change was less strong for happiness than for life satisfaction as is shown in Tab. G.4 and therefore has less impact on the conversion results.

**Table G.4 Frequency distribution happiness CBS, 1993-1997**

| | To what extent do you consider yourself a happy person? | | | | |
|---|---|---|---|---|---|
| Year | Unhappy | Not so happy | Neither happy nor unhappy | Happy | Very Happy |
| 1993 | 0.39% | 2.09% | 9.12% | 63.35% | 25.05% |
| 1994 | 0.70% | 3.48% | 11.36% | 63.79% | 20.67% |
| 1995 | 0.50% | 3.00% | 10.77% | 63.41% | 22.33% |
| 1996 | 0.44% | 3.15% | 10.27% | 64.86% | 21.29% |
| 1997 | 0.25% | 3.17% | 10.03% | 65.00% | 21.55% |
| 1997 | 0.51% | 2.29% | 8.72% | 67.88% | 20.60% |

## List of abbreviations used

| | |
|---|---|
| AUWI | Australian Unity Wellbeing Index |
| EB | Eurobarometer |
| CBS | Statistics Netherlands |
| DHS | Dutch Household Survey |
| EDAC | European Data Center for Work and Welfare |
| ESS | European Social Survey |
| HSIS | Happiness Scale Interval Study |
| NRC | National Research Council |
| NZGSS | New Zealand General Social Survey |
| OECD | Organisation for Economic Cooperation and Development |
| POLS | Permanent Onderzoek Leefsituatie |
| %SM | Percentage of Scale Maximum |
| SCP | The Netherlands Institute for Social Research |
| SWB | Subjective well-being |
| WDH | World Database of Happiness |
| WVS | World Values Survey |

## Glossary of terms used

**Anchor point**
Response option at an end of a discrete scale.

**Beta distribution**
A continuous distribution on a two-sided bounded continuum with shape parameters $\alpha$ and $\beta$.

**Continuum Approach**
Method to estimate a population mean and standard deviation of the continuous cumulative distribution function that fits best to the points on a common, bounded continuum at which response options for a given response scale transit from one to another combined with the frequency distribution of the primary response scale.

**Degree of appreciation**
Valuation of a verbal response option in the context of the scale by an interval on a bounded continuum.

**Extreme**
The boundary of the continuum where the range of values of a continuous distribution falls.

**Happiness Scale Interval Study**
Study to look at survey questions on happiness using verbal response options with the intent to determine consistently what degrees of happiness are denoted by the labels of these options when the questions are asked in different languages and different or cultural subpopulations are addressed.

**(In)comparability problem**
The incomparability of survey results due to differences in response scales used.

**Interpreters' bias**
Phenomenon that an adverb used in the label of a response options is not necessarily always translated into the same adverb in another language.

**Item**
A survey question and its corresponding response options.

**Linear Stretch Method**
Transformation method by which a discrete primary scale is transformed in such a way that the rank of the 'lowest' response option is always projected onto 0 and

the ‘highest’ one onto the highest value of the range, and all the intermediate options are given equally distanced numbers in between.

**Numerical scale**
Response scale with a discrete number of response options, with only the anchor points verbally labelled and all other response options labelled with a number.

**Rank Method**
Method according to which the sample mean is calculated as the weighted average of the ranks of the response options using the relative frequencies as weights.

**Rating scale**
Discrete response scale to rate the response to a survey question.

**Reference boundary**
Boundary between two response options of a discrete primary response scale derived from a reference distribution for the frequency distribution in the same year as the year the reference distribution is based on.

**Reference distribution**
Continuous cumulative distribution based on an appropriate primary scale and frequency distribution in a given year to be used as a reference to derive reference boundaries for the response options of a discrete primary scale given the frequency distribution for this primary scale in the same year as the year the reference distribution is based on.

**Reference Distribution Method**
Method according to which the boundaries between the response options of a primary scale are derived from a reference distribution given the frequency distribution for this primary scale in the same year as the year the reference distribution is based on and for samples from the same population.

**Scale Interval Method**
Method according to which judges assess the points on a common, bounded continuum in which verbal response options for a given response scale transit from one to another. A population mean and standard deviation are estimated using the Continuum Approach.

**Scale Interval Recorder**
Web-based instrument to be used for the assessment of the points on a common, bounded continuum at which verbal response options for a given response scale transit from one option to another.

**Semantic Judgement of Fixed Word Value Method**
Method according to which fixed values are assigned to the labels of verbal response options, based on the rating by experts of a series of qualifications that can be given to verbal response options on a common numerical scale. The sample mean is calculated by means of the Weighted Average Approach.

**Scale transformation**
Transformation of the response options of a primary scale into numerical secondary ratings, usually on a common range.

**Verbal scale**
Response scale with a discrete number of response options labelled with text.

**Weighted Average Approach**
Method according to which the sample mean is calculated as the weighted average of the values assigned to the response options of a discrete scale using the relative frequencies as weights. Generalization of the Rank Method, by not requiring that the numbers assigned to the response options are equal to the ranks of these response options to calculate a sample mean.

**World Database of Happiness**
Archive of research findings on subjective enjoyment of one's life as a whole. The procedure to estimate the population mean and standard deviation presented in this thesis were developed in the context of this project.

## References


Andrews, F.M. & Withey, S.B. (1976).
*Social Indicator of Well-Being, Americans' Perceptions of Life Quality*
Plenum Press, New York

Bălţătescu, S. (2002).
*Problems of transforming scales of life satisfaction*
Euromodule workshop. Berlin. http:--worlddatabaseofhappiness.eur.nl-hap_bib-freetexts-baltatescu_sm_2002A.pdf

Bartikowski, B., Kamei, K., & Chandon, J.L. (2010).
*A verbal rating scale to measure Japanese consumers' perceptions of product quality*
Asia Pacific Journal of Marketing and Logistics 22(2), 179-195
Emerald Group Publishing Limited, doi 10.1108/13555851011026935

Beuningen, J. van, Houwen, K. van der, & Moonen, L. (2014).
*Measuring well-being, An analysis of different response scales*
Discussion Paper, Statistics Netherlands
Available at: http:--www.cbs.nl-NR-rdonlyres-FF644A99-580C-4D7E-B214-BEE54A947D46-0-Measuringwellbeing.pdf

Brulé, G. & Veenhoven, V. (2014).
*The '10-excess' phenomenon in responses to survey questions on happiness*
Accepted with minor revisions for publication in Social Indicators Research,
http://www2.eur.nl/fsw/research/veenhoven/Pub2010s/2014k-full.pdf

Cummins, R. A. (1997).
*The Comprehensive Quality of Life Scale – Intellectual-Cognitive Disability (ComQol-I5)*
5th edition, School of Psychology, Deakin University, Melbourne

Cummins, R.A. & Gullone, E. (2000).
*Why we should not use 5-point Likert scales: The case for subjective quality of life measurement*
Proceedings, Second International Conference on Quality of Life in Cities (pp.74-93). Singapore: National University of Singapore

Cummins, R.A. (2003).
*Normative life satisfaction: measurement issues and homeostatic model*
Social Indicators Research 64, 225-240.

Cummins, R. A. (2009).
*Australian Unity Wellbeing Index, Survey 21*
Report 21.0, May 2009, Australian Centre on Quality of Life, Deakin University, Figure 2.12

Danaher, P.J. & Haddrell, V. (1996).
*A comparison of question scales used for measuring customer satisfaction*
International Journal of Service Industry Management, 7(4), 4-26.

Davis, R.E., Couper, M.P., Janz, N.K., Caldwell, C.H., & Resnicow, K. (2010).
*Interviewer effects in public health surveys*
Health Education Research 25(1), 14-26, Oxford University Press, doi:10.1093-her-cyp046

Deaton, A. (2012).
*The financial crisis and the well-being of Americans.*
Oxford Economic Papers 64(1), 1-26.

DeJonge, T. (2009).
*The state of play in measuring SWB in the Netherlands*
Paper for the OECD meeting Measuring subjective well-being: an opportunity for National Statistical Offices?, Florence, Italy, Available at: http:--www.isqols2009.istitutodeglinnocenti.it-Content_en-Measuring%20SWB%20in%20The%20Netherlands%20v2RGED%20with%20cover%20page.pdf

DeJonge, T. de, Hupkens, C., & Bruggink, J.W. (2009).
*Living a happy, healthy and satisfying life. Background paper for the 3rd World Conference of the OECD in Busan, South Korea*, Available at http:--www.oecd.org-dataoecd-63-3-43705841.pdf?contentId=43705842.

DeJonge, T., Veenhoven, R., & Arends, L.R. (2014a).
*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, Springer, doi: 10.1007-s11205-013-0335-6

DeJonge, T., Veenhoven, R., Kalmijn, W.M., & Arends, L.R. (2014b).
*Stability of boundaries between response options of response scales*
*Does 'very happy' remain equally happy over the years?*
Social Indicators Research. Published online: 3 September 2014, Springer, doi: 10.1007/s11205-014-0735-2

DeJonge, T., Veenhoven, R., & Arends, L.R. (2015a).
*'Very Happy' is Not Always Equally Happy on the Meaning of Verbal Response Options in Survey Questions*
Journal of Happiness Studies 16(1), 77-101, Springer, doi:10.1007-s10902-013-9497-9

DeJonge, T., Veenhoven, R., Kalmijn, W.M., & Arends, L.R. (2015b).
*Pooling time series based on slightly different questions about the same topic Forty years of survey research on happiness and life satisfaction in The Netherlands*
Social Indicators Research. Published online: 21 February 2015, Springer, doi: 10.1007/s11205-015-0898-5

DeJonge, T., Veenhoven, R., Moonen, L., Kalmijn, W.M., Beuningen, J. Van, & Arends, L.R. (2015c).
*Conversion of verbal response scales. Robustness across demographic categories*
Social Indicators Research. Published online: 21 February 2015, Springer, doi: 10.1007/s11205-015-0897-6

Diener, E., Emmons, R.A., Griffin, S., & Larsen, R.J. (1985).
*The Satisfaction With Life Scale*
Journal of Personality Assessment 49, 71-75

Diener, E. & Diener, C. (1996).
*Most people are happy*
Psychological Science 7, 181-185

Dijkgraaf, R. (2008).
*Blikwisselingen*
Publisher Bert Bakker, Amsterdam

Dolan, P. & Kavetsos, G. (2012).
*Happy Talk: Mode of Administration Effects on Subjective Well-Being*
CEP Discussion Paper No. 1159, London School of Economics and Political Science, http:--cep.lse.ac.uk-pubs-download-dp1159.pdf
(assessed May 2014)

European Commission (2012a).
European Opinion Research Group (EORG), Brussels. GESIS Data Archive, Cologne.
*Eurobarometer 57.1 (Mar-May 2002),* ZA3639 Data file Version 1.0.1, doi:10.4232-1.10949
*Eurobarometer 58.1 (Oct-Nov 2002),* ZA3693 Data file Version 1.0.1, doi:10.4232-1.10953
*Eurobarometer 60.1 (Oct-Nov 2003)*, ZA3938 Data file Version 1.0.1, doi:10.4232-1.10958

*Eurobarometer 62.0 (Oct-Nov 2004)*, ZA4229 Data file Version 1.1.0, doi:10.4232-1.10962

European Commission (2012b).
TNS Opinion & Social, Brussels [Producer]. GESIS Data Archive, Cologne.
*Eurobarometer 63.4 (May-Jun 2005)*, ZA4411 Data file Version 1.1.0, doi:10.4232-1.10968
*Eurobarometer 64.2 (Oct-Nov 2005)*, ZA4414 Data file Version 1.1.0, doi:10.4232-1.10970
*Eurobarometer 65.2 (Mar-May 2006)*, ZA4506 Data file Version 1.0.1, doi:10.4232-1.10974
*Eurobarometer 66.1 (Sep-Oct 2006)*, ZA4526 Data file Version 1.0.1, doi:10.4232-1.10980
*Eurobarometer 66.3 (Nov-Dec 2006)*, ZA4528 Data file Version 2.0.1, doi:10.4232-1.10982
*Eurobarometer 67.2 (Apr-May 2007)*, ZA4530 Data file Version 2.1.0, doi:10.4232-1.10984
*Eurobarometer 68.1 (Sep-Nov 2007)*, ZA4565 Data file Version 4.0.1, doi:10.4232-1.10988
*Eurobarometer 70.1 (Oct-Nov 2008)*, ZA4819 Data file Version 3.0.2, doi:10.4232-1.10989
*Eurobarometer 71.2 (May-Jun 2009)*, ZA4972 Data file Version 3.0.2, doi:10.4232-1.10990
*Eurobarometer 72.4 (Oct-Nov 2009)*, ZA4994 Data file Version 3.0.0, doi:10.4232-1.11141
*Eurobarometer 73.4 (May 2010)*, ZA5234 Data file Version 2.0.1, doi:10.4232-1.11479
*Eurobarometer 76.2 (2011),* ZA5566 Data file Version 1.0.0, doi:10.4232-1.11388
*Eurobarometer 76.3 (2011)*, ZA5567 Data file Version 1.0.0, doi:10.4232-1.11448

European Commission (2013).
TNS Opinion & Social, Brussels [Producer]. GESIS Data Archive, Cologne.
*Eurobarometer 69.2 (Mar-May 2008)*, ZA4744 Data file Version 5.0.0, doi:10.4232-1.11755
*Eurobarometer 71.1 (Jan-Feb 2009)*, ZA4971 Data file Version 4.0.0, doi:10.4232-1.11756
*Eurobarometer 74.2 (2010)*, ZA5449 Data file Version 2.2.0, doi:10.4232-1.11626
*Eurobarometer 75.3 (2011)*, ZA5481 Data file Version 2.0.0, doi:10.4232-1.11768
*Eurobarometer 75.4 (2011)*, ZA5564 Data file Version 3.0.0, doi:10.4232-1.11769
*Eurobarometer 77.3 (2012)*, ZA5612 Data file Version 1.0.0, doi:10.4232-1.11558
*Eurobarometer 77.4 (2012)*, ZA5613 Data file Version 2.0.0, doi:10.4232-1.11697
*Eurobarometer 78.1 (2012)*, ZA5685 Data file Version 1.0.0, doi:10.4232-1.11706

Fischer, J.A. (2009).
*Subjective Well-being as Welfare Measure: Concepts and Methodology*
MPRA Paper No. 16619, 2009, München, Germany, http:--mpra.ub.uni-muenchen.de-16619-

Frijters, P., Johnston, D.W., & Shields, M.A. (2008).
*Happiness Dynamics with Quarterly Life Event Data*
IZA Discussion Paper No. 3604

Guven, C., Senik, C., & Stichnoth, H. (2011).
*You can't be happier than your wife. Happiness Gaps and Divorce*
Paris School of Economics, Working Paper No. 2011-01, halshs-00555427

Hazelrigg, L.E. & Hardy, M.A. (2000).
*Scaling the semantics of satisfaction*
Social Indicators Research 49(2), 147-180, doi: 10.1023-A:1006937713249

Houwen, K. van der & Moonen, L. (2014).
*Allochtonen en geluk*
Bevolkingstrends 2014 (10), Statistics Netherlands

Howard, G.S., Ralph, K.M., Gulanick, N.A., Maxwell, S.E., Nance, D.W., & Gerber, S.K. (1979).
*Internal Invalidity in Pretest-Posttest Self-Report Evaluations and a Re-evaluation of Retrospective Pretests*
Applied Psychological Measurement 3(1), 1-23, doi: 10.1177-014662167900300101

Hull, C.L. (1922).
*The Conversion of Test Scores into Series Which Shall Have Any Assigned Mean and Degree of Dispersion*
Journal of Applied Psychology 6(3), 298-300

Jones, L.V. & Thurstone, L.L. (1955).
*The Psychophysics of Semantics. An Experimental Investigation*
The Journal of Applied Psychology 39 (1), 31-36

Kalmijn, W.M. (2010).
*Quantification of Happiness Inequality*
PhD-thesis, Erasmus University Rotterdam, The Netherlands: Ipskamp Drukkers, Enschede. Available at http:--repub.eur.nl-res-pub-21777-

Kalmijn, W.M., Arends, L.R., & Veenhoven, R. (2011).
*Happiness Scale Interval Study, Methodological Considerations.*
Social Indicators Research 102(3), 497-515, doi: 10.1007-s11205-010-9688-2

Kalmijn W.M. (2012).
*Happiness is not Normally Distributed. A comment to Delhey and Kohler*
Social Science Research 41 (1), 199-202, doi: 10.1016-j.ssresearch.2011.11.008

Kalmijn, W.M. (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, doi: 10.1007-s11205-011-9943-1

Katz, D. (1942).
*Do Interviewers Bias Poll Results?*
The Public Opinion Quarterly, 6(2), 248-268, Oxford University Press on behalf of the American Association for Public Opinion Research, http:--www.jstor.org-stable-2745023

Lee, H., Kim, K.D., & Shin, D.C. (1982).
*Perceptions of Quality of Life in an Industrializing Country: The Case of the Republic of Korea*
Social Indicators Research 10, 297-317, doi: 10.1007-BF00301097

Lim, H.E. (2008)
*The Use of Different Happiness Rating Scales: Bias and Comparison Problem?*
Social Indicators Research 87, 259-267, Springer, doi: 10.1007-s11205-007-9171-x

Lodge, M. (1981).
*Magnitude scaling : quantitative measurement of opinions*
Beverly Hills, London: Sage Publications, Series Quantitative applications in the social sciences, ISSN 0149-192X ; 07-025, doi: 10.4135-9781412984874

Market Directions (2014).
*Discussion Paper on Scales for Measuring Customer Satisfaction*
http://www.marketdirectionsmr.com/images/SurveyScales.pdf (accessed November 2014)

Mazaheri, M. & Theuns, P. (2009).
*Effects of Varying Response Formats on Self-ratings of Life-Satisfaction.*
Social Indicators Research 90, 381-395. doi 10.1007-s11205-008-9263-2

McDowell, I. (2006).
*Measuring Health. A Guide to Rating Scales and Questionnaires. Third Edition*
Oxford University Press,
http://a4ebm.org/sites/default/files/Measuring%20Health.pdf

McKennel, A.C. & Andrews, F.M. (1980).
*Models of cognition and affect in perceptions of well-being*
Social Indicators Research 8, 127-155

Michalos, A.C., Zumbo, B.D. (1999).
*Public Services and the Quality of Life*
Social Indicators Research 48, 125-156, doi: 10.1023-A:1006893225196

Michalos, A.C. (2003).
*Policing Services and The Quality of Life*
Social Indicators Research 61, 1-18, doi: 10.1023-A:1021259917948

Michalos, A.C., Zumbo, B.D. (2003).
*Leisure Activities, Health and the Quality of Life*
In Michalos, A.C.(Ed)., Essays on the Quality of Life (pp. 217-238). Kluwer, 2003, Dordrecht, The Netherlands

Michalos, A.C. & Orlando, J.A. (2006).
*A Note on Student Quality of Life*
Social Indicators Research 79, 51-59, doi: 10.1007-s11205-005-2404-y

MORI (2002).
*Public Service Reform. Measuring & Understanding Customer Satisfaction*
MORI Social Research Institute, London, https://www.ipsos-mori.com/DownloadPublication/1202_sri_local_gov_public-service_reform_measuring_and_understanding_customer_satisfaction_042002.PDF

National Research Council (2013).
*Subjective Well-Being: Measuring Happiness, Suffering, and Other Dimensions of Experience*. Panel on Measuring Subjective Well-Being in a Policy-Relevant Framework. In A.A. Stone and C. Mackie (Eds). Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

OECD (2013).
OECD Guidelines on Measuring Subjective Well-Being, OECD Publishing. http:--dx.doi.org-10.1787-9789264191655-en

Sangster, R.L., Willits, F.K., Saltiel, J., Lorenz, F.O., & Rockwood, T.H. (2001)
*The effect of numerical labels on response scales*,
Article presented at the Annual Meeting of the American Statistical Association, Atlanta, GA, http:--www.bls.gov-osmr-pdf-st010120.pdf

Saris, W.E. & Andreenkova, A. (2001)
*Following Changes in Living Conditions and Happiness in Post Communist Russia: The Russet Panel*
Journal of Happiness Studies 2, 95-109, doi:10.1023/A:1011579608121

Saris, W.E. & Gallhofer, I.N. (2007).
*Design, evaluation, and analysis of questionnaires for survey research*
Publisher Hoboken, New York, USA, Wiley-Interscience, Wiley series in survey methodology, ISBN 978-0-470-11495-7, e-ISBN 978-0-470-16519-5

Scherpenzeel, A. (1999).
*Why use 11-point scales?*
*Documentation of the Swiss Household Panel*
http:--forscenter.ch-en-our-surveys-swiss-household-panel-documentationfaq-methods-varia- (accessed May 2014)

Schmitt, H., Scholz, E., Leim, I., & Moschner, M. (2008).
*The Mannheim Eurobarometer Trend File 1970-2002 (ed. 2.00).*
European Commission [Principal investigator]. GESIS Data Archive, Cologne. ZA3521 Data file Version 2.0.1, doi:10.4232-1.10074

Schwartz, C.E. & Sprangers, M.A.G. (1999).
*Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research*
Social Science & Medicine 48, 1531-1548

Schwartz, C.E., Ahmed, S., Sawatzky, R., Sajobi, T., Mayo, M., Finkelstein, J., Lix, L., Verdam, M.G.E., Oort, F.J., & Sprangers, M.A.G. (2013).
*Guidelines for secondary analysis in search of response shift*
Quality of Life Research 22, 2663 - 2673, doi:10.1007-s11136-013-0402-0

Schwarz, N., Knauper, B., Hippler, H.J., Noelle-Neumann, E., & Clark, W. (1991).
*Rating Scales: Numeric Values May Change the Meaning of Scale Labels.*
*The Public Opinion Quarterly,* 55, 570-582. http:--www.jstor.org-stable-2749407

Senik, C. (2013).
*The French Unhappiness Puzzle: the Cultural Dimension of Happiness*
Université Paris-Sorbonne and Paris School of Economics
http:--www.parisschoolofeconomics.eu-docs-senik-claudia-new_version_french_october25_2013.pdf

Storm, C., Jones, C., & Storm, T. (1996),
*Aspects of meaning in words related to happiness*
Cognition & Emotion 10, 279-302, doi: 10.1080-026999396380259

Studer, R. & Winkelmann, R. (2014).
*Reported happiness, fast and slow,*
Social Indicators Research 117(3), 1055-1067, doi: 10.1007-s11205-013-0376-x

The Leisure Development Center. (1980).
*Survey of Values in 13 Countries. Table book for the 1980 International Conference on Human Values.*
The Leisure Development Center, Tokyo
HTTP:--worlddatabaseofhappiness.eur.nl-hap_bib-freetexts-~leisure development centre_1980.pdf

Veenhoven, R. (1984).
*Databook of Happiness.*
Publisher Reidel, Dordrecht, The Netherlands
http:--www2.eur.nl-fsw-research-veenhoven-Pub1980s-84b-con.htm

Veenhoven, R. (1993).
*Happiness in nations, subjective appreciation of life in 56 nations, 1946-1992.*
Studies in Social-Cultural Transformation, No. 2, Risbo, Erasmus University Rotterdam, Netherlands

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

Veenhoven, R. (2008).
*The International Scale Interval Study*.
In V. Møller & D. Huschka (Eds*),* Quality of Life in the new millennium: 'Advances in quality-of-life studies, theory and research', Part 2: Refining concepts and measurement to assess cross-cultural quality-of-life (pp. 45-58). Social Indicator Research Series, vol. 35, Dordrecht, The Netherlands: Springer Press

Veenhoven, R. (2011).
*World Database of Happiness, Example of a focused 'Findings Archive',*
RatSWD, Working Paper Series, Working Paper, No. 169, http:--www.ratswd.de-download-RatSWD_WP_2011-RatSWD_WP_169.pdf

Veenhoven, R. (2015a).
*Measures of Happiness*
World Database of happiness, Erasmus University Rotterdam
Available at: http:--www1.eur.nl-fsw-happiness-hap_quer-hqi_fp.htm

Veenhoven, R. (2015b)
*Happiness in Nations*
World Database of Happiness
Available at: http:--www1.eur.nl/fsw/happiness/hap_nat/nat_fp.php?mode=1

Ventegodt, S. (1995)
*Liskvalitet I* Danmark
Forskningscentrets Forlag, Copenhagen, Denmark

Ventegodt, S. (1996).
*Liskvalitet hos 4500 31-33-arige. (The Quality of Life of 4500 31-33-Years-Olds)*
Forskningscentrets Forlag, Copenhagen, Denmark

Voorpostel, M., Tillmann, R., Lebert, F., Weaver, B., Kuhn, U., Lipps, O., Ryser, V.A., Schmid, F., & Wernli, B. (2009).
Swiss Household Panel User Guide (1999 - 2008)
Swiss Foundation for Research in Social Sciences
Available at: http:--aresoas.unil.ch-DataWeb-SHP_USER_GUIDE_W1_W10.pdf

Wierzbicka, A. (2004).
*'Happiness' in cross-linguistic & cross-cultural perspective*
Daedalus 133, 34-43, doi: 10.1162-001152604323049370

# DIFFERENT SURVEY QUESTIONS ON THE SAME TOPIC
## How to make responses comparable?

## Summary

Survey data are often used for comparison purposes, such as comparisons across nations or comparisons over time. Ideally, this would require equivalent questions and equivalent responses options to these questions. Yet there is a lot of variation in the response scales used, which, for example, differ in the number of response options used and the labelling of these options. This difference in items[27] is no problem when surveys are analysed separately, but it limits the comparability of findings gathered in different surveys that used different items for the same topic. This reduces our accumulation of knowledge and calls for methods to transform ratings on different scales to attain comparable results and to correct for effects of changes in measurements and other influencing factors.

Conventional methods to transform ratings on different response scales to a common one, such as the commonly used Linear Stretch Method, fall short to overcome the comparability problem caused by the non-uniformity of survey items. The weaknesses of these early transformation methods also appear when the transformed scores are compared to average ratings on 0-10 numerical scales in the same country in the same year. The shortcomings of conventional methods instigated the development of new techniques, which will be discussed in this thesis.

This thesis is divided into four parts. In the first part we give a comprehensive description of the comparability problem and why conventional methods fall short to solve this problem. Each of the three parts that follow focuses on a successive innovation to improve the comparability of survey findings with different survey items.

**Part 1: The incomparability problem and conventional approaches**
The first part starts with an introduction to the incomparability problem and an overview of the diversity in survey items. Throughout this thesis use is made of a number of survey items that constitute several time series on happiness and life satisfaction for the Dutch population and an outline is given in part 1. This is followed by a description and illustration of the problem of incomparability of these time series based on findings taken

[27] The survey question and its corresponding response options.

from different surveys. Next two conventional methods for scale transformation are described: the Linear Stretch Method and the Semantic Judgement of Fixed Word Value Method. For both methods we explain why they fall short to solve the comparability problem and conclude that these shortcomings require further innovations.

**Part 2 Innovation 1: The Happiness Scale Interval Study**

To counter the shortcomings of the conventional scale transformation methods the Happiness Scale Interval Study (HSIS) was started. This study was set up to look at survey questions on happiness using verbal response options, such as 'Very happy' and 'Pretty happy' with the intent to determine what degrees of happiness are denoted by such terms when based in different questions and languages. In the HSIS persons who are referred to as 'judges' are asked to rate the degree of happiness denoted by each of the verbal response options in the context of the full item, using the web-based Scale Interval Method. The HSIS and the Scale Interval method are described in detail in Ch. 2.

The Scale Interval Method offers some interesting applications which allow a view on the size of the comparability problem. For the two applications described in this thesis, use is made of the results of several studies that have been conducted under the umbrella of the HSIS. These studies and the recruitment of judges employed for them are described in Ch. 3.

In the first application of the Scale Interval Method the question is addressed whether response scales which appear to be equivalent, can also be considered to be equivalent when interpreting and mutually comparing survey results. Does 'Very happy', for instance, denote the same degree of well-being as 'Very satisfied' in an otherwise equivalent rating scale? This problem has been studied for equivalent response scales for happiness and life satisfaction with response options labelled in Dutch or in Spanish. The results are described in Ch. 4. We found no differences between these topics in the degree of appreciation assigned to response options labelled in Dutch, but found considerable differences if the response options were labelled in Spanish. We conclude that language and culture are influential factors when it comes to whether equivalent response options for happiness and life satisfaction are appreciated equally. The interpretation of the scales by respondents therefore has to be examined and discussed carefully in advance, before mutually comparing survey results for two different topics. The Scale Interval Recorder can be used for that purpose.

The focus in the second application of the Scale Interval Method was on the meaning of verbal response options in survey questions. We used the Scale Interval Method to get insight in the size of the comparability problem. Application of the Scale Interval Method to commonly used survey questions on happiness in Dutch language reveals considerable differences. The implications of this for research synthesis are discussed in Ch. 5.

In part 2 we also compare the results of the Scale Interval Method to those of the conventional methods and conclude that an additional innovation is necessary to solve this comparability problem. This second innovation is the topic of part 3.

**Part 3 Innovation 2: The Continuum Approach**

The comparability problem is partly due to the variety of response scales that has developed over time caused by the use of discrete scales. If in the conventional scale transformation methods or the Scale Interval Method a discrete primary scale is transformed, the resulting secondary scale is still discrete. Although the use of discrete scales in survey research is often practically motivated, a more valid approach is to consider the existence of a latent continuous variable underlying the survey variable, the distribution of which is estimated using the survey item and the response to it.

The Continuum Approach was developed with the notion that happiness is to be treated as a continuous variable. In the Continuum Approach the shape parameters $\alpha$ and $\beta$ of the best fitting beta distribution are estimated on basis of the cumulative frequencies and the values on the continuum from 0 to 10 of the boundaries between the response options of the primary scale. The mean $\mu$ based on the parameters of this best fitting beta distribution is considered to be an estimator for the mean happiness in the population.

An outline of the Continuum Approach is given in Ch. 6. In that chapter we also show that application of the Continuum Approach in combination with the assessments of the boundaries between response options obtained using the Scale Interval Method does not solve the comparability problem entirely and that an third innovation is required.

**Part 4 Innovation 3: The Reference Distribution Method**

With the Reference Distribution Method an attempt is made to deal with the fact that, for a given year and a given population, one would expect the distribution means after scale transformation for similar questions about happiness asked in different representative surveys to be approximately the

same irrespective of the primary response scales used. This provides the basis for another approach to transformation.

The Reference Distribution Method for making survey data comparable builds heavily on the Scale Interval Method. Basically the two methods are identical except that in the Reference Distribution Method the boundaries between the response options of the primary scale are derived from a reference distribution instead of being derived from the assessments by judges by means of on the Scale Interval Recorder.

This Reference Distribution Method is introduced in Ch. 7. In this chapter a comparison is also made between the means estimated by application of the Continuum Approach for the two methods mentioned of deriving boundaries between the response options. From this comparison we conclude that the Reference Distribution Method performs better than the Scale Interval Method with respect to solving the comparability problem.

When the Continuum Approach is applied to the time series of a survey which has remained unchanged over time, the boundaries between the response options are assumed to remain unchanged over time. We inspect whether this assumption is realistic in Ch. 8. Using time series of survey items from the Eurobarometer, surveys of Statistics Netherlands and from the Dutch Household Survey, we conclude that the research question can be answered affirmative.

A switch from a verbal scale to a numerical scale causes a severe problem for trend analyses, due to the incomparability of the old and new measurements. The Reference Distribution Method can deal with this comparison problem. In Ch. 9 we address the question whether the boundaries between response options derived for the general population can be used for demographic categories to produce reliable, extended time series to monitor differences in trends among these categories. We conclude that this is possible and that it is not necessary to derive boundaries for each demographic category separately.

Lastly we applied the Reference Distribution Method to pool time series of happiness and life satisfaction which span a time period of almost 40 years. In Ch. 10 we conclude that in the past 40 years the Dutch have become slightly happier and satisfied with their lives.

**Note to the contents of this thesis**

This thesis is to a large extent a compilation of the following papers which have been written as part of the PhD-research.

DeJonge, T., Veenhoven, R., & Arends, L.R. (2014a).
*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, Springer, doi: 10.1007-s11205-013-0335-6

Parts of this paper have been used in Ch. 1, 2, 6 and 7.

---

DeJonge, T., Veenhoven, R., Kalmijn, W.M., & Arends, L.R. (2014b).
*Stability of boundaries between response options of response scales*
*Does 'very happy' remain equally happy over the years?*
Social Indicators Research. Published online: 3 September, 2014, Springer, doi: 10.1007/s11205-014-0735-2

The main part of Ch. 8 is based on this paper and parts of this paper have been used in Ch. 3 and 7.

---

DeJonge, T., Veenhoven, R., & Arends, L.R. (2015a).
*'Very Happy' is Not Always Equally Happy on the Meaning of Verbal Response Options in Survey Questions*
Journal of Happiness Studies 16(1), 77-101, Springer, doi:10.1007-s10902-013-9497-9

The main part of Ch. 5 is based on this paper and parts of this paper have been used in Ch. 1 and 3.

---

DeJonge, T., Veenhoven, R., Kalmijn, W.M., & Arends, L.R. (2015b).
*Pooling time series based on slightly different questions about the same topic*
*Forty years of survey research on happiness and life satisfaction in The Netherlands*
Social Indicators Research. Published online: 21 February 2015, Springer, doi: 10.1007/s11205-015-0898-5

The main part of Ch. 10 is based on this paper and parts of this paper have been used in Ch. 1.

---

DeJonge, T., Veenhoven, R., Moonen, L., Kalmijn, W.M., Beuningen, J. Van, & Arends, L.R. (2015c).
*Conversion of verbal response scales. Robustness across demographic categories*
Social Indicators Research. Published online: 21 February 2015, Springer, doi: 10.1007/s11205-015-0897-6

The main part of Ch. 9 is based on this paper.

# VERSCHILLENDE ENQUÊTEVRAGEN OVER HETZELFDE ONDERWERP
# Hoe kunnen antwoorden vergelijkbaar gemaakt worden?

## Samenvatting (summary in Dutch)

Enquêtegegevens worden vaak gebruikt voor vergelijkingsdoeleinden, zoals voor onderlinge vergelijking van landen of voor een vergelijking door de tijd. Idealiter zijn dan zowel de vragen als de bijbehorende responsopties equivalent. De variatie in de responsschalen die gebruikt worden is echter groot. Responsschalen verschillen bijvoorbeeld in het aantal responsopties en in de bewoordingen waarmee de responsopties gelabeld zijn. Dit verschil in items[28] is geen probleem als enquêtes los van elkaar geanalyseerd worden, maar beperkt de vergelijkbaarheid van resultaten van enquêtes die onderling verschillen in de items die gebruikt zijn voor een en hetzelfde onderwerp. Dit beperkt de mate waarin we kennis kunnen verrijken en combineren. Dit vraagt om methoden om scores die met verschillende schalen gemeten zijn, vergelijkbaar te maken en om te corrigeren voor effecten van veranderingen in de wijze van meten en veranderingen in andere invloedsfactoren.

Conventionele methoden, zoals de algemeen gebruikte methode van Linear Stretch, om scores die met verschillende responsschalen gemeten zijn te transformeren naar een universele schaal , schieten tekort om het probleem van onvergelijkbaarheid vanwege de non-uniformiteit van responsitems op te lossen. De zwakte van deze vroege methoden voor schaaltransformatie blijkt ook uit een vergelijking van getransformeerde scores met gemiddelde scores op numerieke 0-10 schalen in hetzelfde land en hetzelfde jaar. De tekortkomingen van de conventionele methoden vormen de aanleiding voor de ontwikkeling van nieuwe technieken die in dit proefschrift besproken worden.

Dit proefschrift is onderverdeeld in vier delen. In het eerste deel beschrijven we uitgebreid het probleem van onvergelijkbaarheid en gaan we in op de oorzaken waardoor conventionele methoden tekort schieten om dit probleem op te lossen. In elk deel dat volgt ligt het accent op een volgende innovatie om de vergelijkbaarheid van enquêteresultaten die met

[28] De enquêtevraag en de bijbehorende responsopties.

verschillende items verkregen zijn te verbeteren.

**Deel 1: Het onvergelijkbaarheidsprobleem en conventionele benaderingen**

Deel 1 begint met een introductie tot het onvergelijkbaarheidsprobleem en geeft een overzicht van de diversiteit in enquête-items. Op verschillende plaatsen in dit proefschrift, wordt gebruik gemaakt van enquête-items die elk de basis vormen voor een tijdreeks van het geluk of de tevredenheid van de Nederlandse bevolking. We geven een overzicht van deze tijdreeksen in deel 1 en laten aansluitend zien waarom deze tijdreeksen, die op uitkomsten van verschillende enquêtes zijn gebaseerd, niet met elkaar te vergelijken zijn. Daarna volgt een beschrijving van twee conventionele methoden voor schaaltransformatie: de 'Linear Stretch Method' en de 'Semantic Judgement of Fixed Word Value Method'. Voor beide methoden leggen we uit waarom ze niet geschikt zijn om het probleem van onvergelijkbaarheid op te lossen en concluderen we dat verdere innovaties nodig zijn om de tekortkomingen van deze methoden te overbruggen.

**Deel 2 Innovatie 1: The Happiness Scale Interval Study**

De Happiness Scale Interval Study (HSIS) is gestart als antwoord op de tekortkomingen van de conventionele schaaltransformatiemethoden. De HSIS richt zich op enquêtevragen over geluk met verbale responsopties zoals 'Zeer gelukkig' en 'Redelijk gelukkig'. De studie beoogt te bepalen aan welke mate van geluk deze termen uitdrukking geven afhankelijk van de vraag die eraan voorafgaat en de taal waarin de vraag en de bijbehorende responsopties geformuleerd zijn. Voor deelname aan de HSIS worden personen benaderd die worden aangeduid als 'beoordelaars'. De opdracht die zij krijgen is om voor een reeks enquête-items met verbale responsschalen voor elke responsoptie aan te geven welke mate van geluk ermee correspondeert binnen de context van het item als totaal. Het instrument daarvoor is de Schaalintervalrecorder waar zij via het internet toegang toe krijgen. De HSIS en de Schaalintervalmethode worden in detail besproken in hoofdstuk 2.

De Schaalintervalmethode kent enkele interessante toepassingen om een beeld te kunnen krijgen van de omvang van het probleem van onvergelijkbaarheid. Voor de twee toepassingen die in dit proefschrift worden beschreven is gebruikt gemaakt van de resultaten van een aantal studies die zijn uitgevoerd onder de noemer van de HSIS. Een beschrijving van deze studies en de inzet van beoordelaars daarin staat in hoofdstuk 3.

De eerste toepassing van de Schaalintervalmethode is gericht op de vraag of responsschalen die equivalent lijken, ook als equivalent beschouwd mogen worden bij de interpretatie en onderlinge vergelijking van enquêteresultaten. Drukt 'Zeer gelukkig' bijvoorbeeld dezelfde mate van welzijn uit als 'Zeer tevreden' gegeven een responsschaal voor geluk die verder equivalent is aan de schaal voor tevredenheid? Dit probleem is onder de loep genomen voor equivalente responsschalen voor geluk en tevredenheid met responsopties met een label in het Nederlands of in het Spaans. De resultaten daarvan zijn beschreven in hoofdstuk 4. We vonden geen verschillen tussen geluk en tevredenheid voor wat betreft de mate van waardering die wordt toegekend aan responsopties met een label in het Nederlands. Voor responsopties met een label in het Spaans vonden we echter aanzienlijke verschillen. We concluderen dat taal en cultuur invloedsfactoren zijn als het gaat om de vraag of equivalente responsopties voor geluk en tevredenheid gelijk gewaardeerd worden. Hoe respondenten responsschalen interpreteren dient daarom zorgvuldig onderzocht en bediscussieerd te worden, alvorens enquêteresultaten voor twee verschillende onderwerpen onderling te vergelijken. De Schaalintervalrecorder kan voor dat doel gebruikt worden.

In de tweede toepassing van de Schaalintervalmethode ligt het accent op de betekenis van de verbale responsoptions van enquêtevragen. We hebben de Schaalintervalmethode gebruikt om inzicht te krijgen in de omvang van het onvergelijkbaarheidsprobleem. Uit de toepassing van de Schaalintervalmethode voor gangbare enquêtevragen over geluk in het Nederlands blijkt dat er omvangrijke verschillen zijn. Wat dit impliceert voor de synthese van onderzoek wordt besproken in hoofdstuk 5.

In deel 2 vergelijken we de resultaten van Schaalintervalmethode ook met die van de conventionele methoden. De conclusie die we aan die vergelijking verbinden is dat nog een innovatie nodig is om een oplossing te vinden voor het onvergelijkbaarheidsprobleem. Deze tweede innovatie is het onderwerp van deel 3.

**Deel 3 Innovatie 2: The Continuum Approach**

Het onvergelijkbaarheidsprobleem is voor een deel toe te schrijven aan de variatie in responsschalen die in de loop der tijd is ontstaan door het gebruik van discrete schalen. Als een primaire schaal met een van de conventionele methoden of met de Schaalintervalmethode wordt getransformeerd, dan is de resulterende secundaire schaal nog steeds discreet. Hoewel het gebruik van discrete schalen in enquête-onderzoek vaak practisch gemotiveerd is, is een meer valide benadering om te

veronderstellen dat er een latente continue variabele bestaat die de basis is voor de variable in de enquête en waarvan de verdeling geschat kan worden door gebruik te maken van het enquête-item en de respons daarop.

De Continuum Approach is ontwikkeld vanuit het besef dat geluk behandeld moet worden als een continue variabele. In de Continuum Approach worden de vormparameters α en β van de best-passende betaverdeling geschat op basis van de cumulatieve frequenties en de waarden op het continuum van 0 tot 10 van de grenzen tussen de responsopties van de primaire schaal. Het gemiddelde μ dat gebaseerd is op de parameters van deze best-passende betaverdeling geldt als schatter voor het gemiddelde geluk in de populatie.

De contouren van de Continuum Approach zijn gegeven in hoofdstuk 6. In dat hoofdstuk laten we ook zien dat toepassing van de Continuum Approach in combinatie met de bepaling van de grenzen tussen responsopties met de Schaalintervalmethode het probleem van onvergelijkbaarheid niet helemaal oplost en dat een derde innovatie gewenst is.

**Deel 4 Innovatie 3: De Referentieverdelingmethode**

Met de Referentieverdelingsmethode wordt geprobeerd om tegemoet te komen aan het feit dat voor een gegeven jaar en een gegeven populatie de distributiegemiddelden na schaaltransformatie voor vergelijkbare vragen over geluk die gesteld zijn in verschillende representatieve enquêtes bij benadering hetzelfde moeten zijn, ongeacht de primaire responsschalen die gebruikt zijn. Dit vormt de basis voor een andere benadering om resultaten uit verschillende enquêtes onderling vergelijkbaar te maken.

De Referentieverdelingsmethode om gegevens over geluk vergelijkbaar te maken rust sterk op de Schaalintervalmethode. In beginsel zijn beide methodes identiek, behalve dat de grenzen tussen de responsopties van de primaire schaal in de Referentieverdelingsmethode afgeleid worden van een referentieverdeling in plaats van dat ze gebaseerd zijn op het oordeel dat beoordelaars hebben gegeven door toepassing van de Schaalintervalmethode.

Deze Referentieverdelingmethode wordt in hoofdstuk 7 geïntroduceerd. In dat hoofdstuk wordt ook een vergelijking gemaakt tussen de gemiddelden volgens schattingen met de Continuum Approach voor de twee genoemde methoden om grenzen tussen responsopties af te leiden. Op basis van deze vergelijking concluderen we dat de Referentieverdelingmethode betere resultaten geeft dan de

Schaalintervalmethode voor wat betreft de oplossing van het probleem van onvergelijkbaarheid.

Voor de toepassing van de Continuum Approach op een tijdreeks van een enquête die in de loop der tijd niet is veranderd, worden de grenzen tussen responsopties verondersteld gelijk te blijven in de overeenkomstige periode. In hoofdstuk 8 gaan we in op de vraag of deze aanname realistisch is. Op basis van tijdreeksen van resultaten voor enquête-items uit de Eurobarometer, enquêtes van het Centraal Bureau voor de Statistiek en uit de Nederlandse Huishoudens Enquête concluderen we dat deze vraag bevestigend beantwoord kan worden.

Een overgang van een verbale naar een numerieke schaal leidt tot grote problemen bij de analyse van trends vanwege de onvergelijkbaarheid van de oude en de nieuwe metingen. De Referentieverdelingsmethode kan met dit probleem overweg. In hoofdstuk 9 besteden we aandacht aan de vraag of grenzen tussen responsopties die afgeleid zijn voor de gehele populatie gebruikt kunnen worden voor demografische categorieën om zo betrouwbare, verlengde tijdreeksen te maken waarmee verschillen in trends tussen deze categorieën in beeld gebracht kunnen worden. We concluderen dat dit mogelijk is en dat het niet nodig is om voor elke demografische categorie afzonderlijk grenzen te bepalen.

Tot slot hebben we de Referentieverdelingsmethode toegepast om tijdreeksen voor geluk en tevredenheid te combineren die gezamenlijk een periode van bijna 40 jaar overbruggen. We concluderen in hoofdstuk 10 dat Nederlanders in de afgelopen 40 jaar iets gelukkiger en tevredener met hun leven zijn geworden.

**Opmerkingen bij de inhoud van dit proefschrift**

Dit proefschrift is in belangrijke mate een compilatie van de volgende artikelen die zijn geschreven als onderdeel van het promotieonderzoek.

DeJonge, T., Veenhoven, R., & Arends, L.R. (2014a).
*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, Springer, doi: 10.1007-s11205-013-0335-6

Delen van dit artikel zijn gebruikt in de hoofdstukken 1, 2, 6 and 7.

DeJonge, T., Veenhoven, R., Kalmijn, W.M., & Arends, L.R. (2014b).
*Stability of boundaries between response options of response scales*
*Does 'very happy' remain equally happy over the years?*
Social Indicators Research. Published online: 3 September, 2014, Springer, doi: 10.1007/s11205-014-0735-2

Het grootste deel van hoofdstuk 8 is gebaseerd op dit artikel en delen van dit artikel zijn gebruikt in de hoofdstukken 3 and 7.

---

DeJonge, T., Veenhoven, R., & Arends, L.R. (2015a).
*'Very Happy' is Not Always Equally Happy on the Meaning of Verbal Response Options in Survey Questions*
Journal of Happiness Studies 16(1), 77-101, Springer, doi:10.1007-s10902-013-9497-9

Het grootste deel van hoofdstuk 5 is gebaseerd op dit artikel en delen van dit artikel zijn gebruikt in de hoofdstukken 1 en 3.

---

DeJonge, T., Veenhoven, R., Kalmijn, W.M., & Arends, L.R. (2015b).
*Pooling time series based on slightly different questions about the same topic*
*Forty years of survey research on happiness and life satisfaction in The Netherlands*
Social Indicators Research. Published online: 21 February 2015, Springer, doi: 10.1007/s11205-015-0898-5

Het grootste deel van hoofdstuk 10 is gebaseerd op dit artikel paper en delen van dit artikel zijn gebruikt in de hoofdstuk 1.

---

DeJonge, T., Veenhoven, R., Moonen, L., Kalmijn, W.M., Beuningen, J. Van, & Arends, L.R. (2015c).
*Conversion of verbal response scales. Robustness across demographic categories*
Social Indicators Research. Published online: 21 February 2015, Springer, doi: 10.1007/s11205-015-0897-6

Het grootste deel van hoofdstuk 9 is gebaseerd op dit artikel.

# Dankwoord (Acknowledgement)

In 2009 mocht ik bij het CBS het thema ‘Welzijn van de Nederlandse bevolking’ opzetten. Een bezoek aan Ruut Veenhoven in Rotterdam om ideeën op te doen voor dat thema, leidde ertoe dat we gezamenlijk een schaalintervalstudie uitvoerden. Daarmee was de kiem gelegd voor het promotieonderzoek waaraan ik de afgelopen jaren met veel enthousiasme gewerkt heb. Ik prijs mijzelf gelukkig dat Ruut mij gevraagd heeft aan dit onderzoek te beginnen en dat hij er vertrouwen in stelde dat ik er iets van zou kunnen maken. Zijn aanwijzingen waren niet alleen inhoudelijk zeer waardevol, maar ook belangrijk om te leren wetenschappelijke artikelen schrijven.

Het was een luxe om naast Ruut ook Wim Kalmijn als co-promotor te hebben. Wim heeft er veel tijd in gestoken om mijn conceptartikelen grondig door te nemen en van nuttig en noodzakelijk commentaar te voorzien, vaak afgesloten met een opmerking in de trant van ‘de pot met zout voor de slakken is vrijwel leeg’. Waar mijn kennis van statistiek te kort schoot, wist hij vragen op dat terrein keer op keer op adequate wijze te beantwoorden. Zowel Ruut als Wim hebben mij het leven makkelijk gemaakt, doordat ik kon voortborduren op hun ideeën. Zonder de lancering van de Happiness Scale Interval Study door Ruut en de introductie van de Continuum Approach door Wim, was dit proefschrift er niet geweest.

Als eerdere co-promotor van het promotieonderzoek van Wim Kalmijn, lag het voor de hand dat ook Lidia Arends bij mijn onderzoek werd betrokken. Ik ben haar zeer erkentelijk dat zij in een voor haar zeer hectische tijd, de rol van promotor op zich wilde nemen. Het meer formele eindtraject van mijn promotie heeft zij op een voor mij prettige wijze in goede banen geleid.

Ik denk dat ik geboft heb dat Miranda Aldham-Breary mijn artikelen geredigeerd heeft. Zij heeft dit uiterst consciëntieus gedaan. Niet alleen werden de artikelen zelf daar beter van, maar ook mijn kennis en begrip van de Engelse taal.

Van de oud-collega’s bij het CBS wil ik er een aantal bij name noemen. Allereerst is dat Marly Odekerken. Haar steun had ik nodig om de eerder genoemde schaalintervalstudie uit te kunnen voeren. Zij zei die gelukkig volmondig toe toen ik met het voorstel daartoe kwam. Het was heel prettig om bij het CBS met Marleen Wingen samen te werken. Zij heeft gezorgd dat ik kon beschikken over de lange tijdreeksen voor geluk en tevredenheid van het CBS. De informatie over het split-half experiment van

het CBS en bijbehorende data heb ik gekregen van Linda Moonen en Jacqueline van Beuningen. Ik vond het erg leuk dat zij bereid waren om co-auteur te zijn van het artikel waarop het negende hoofdstuk van dit proefschrift gebaseerd is.

Dankzij Jeroen Boelhouwer van het SCP lukte het om ook een groot deel van zijn collega's als beoordelaars deel te laten nemen aan de schaalintervalstudie. Hij voorzag mij bovendien van de tijdreeksen voor geluk en tevredenheid van het SCP.

Hoewel ik maar een paar keer per jaar op de Erasmusuniversiteit kwam, waren dat altijd wel dagen om naar uit te kijken. De vriendelijke ontvangst door de vrijwilligers van de "Onderzoeksgroep Geluk" maakte dat ik mij altijd welkom voelde, waarvoor mijn dank.

Mijn jongste zus Nienke zei tot mijn grote vreugde direct 'ja' toen ik haar vroeg om paranimf te willen zijn en wilde pas daarna weten wat er daarvoor van haar verwacht werd. Mijn andere paranimf is mijn levenspartner Robert Oude Wolbers.

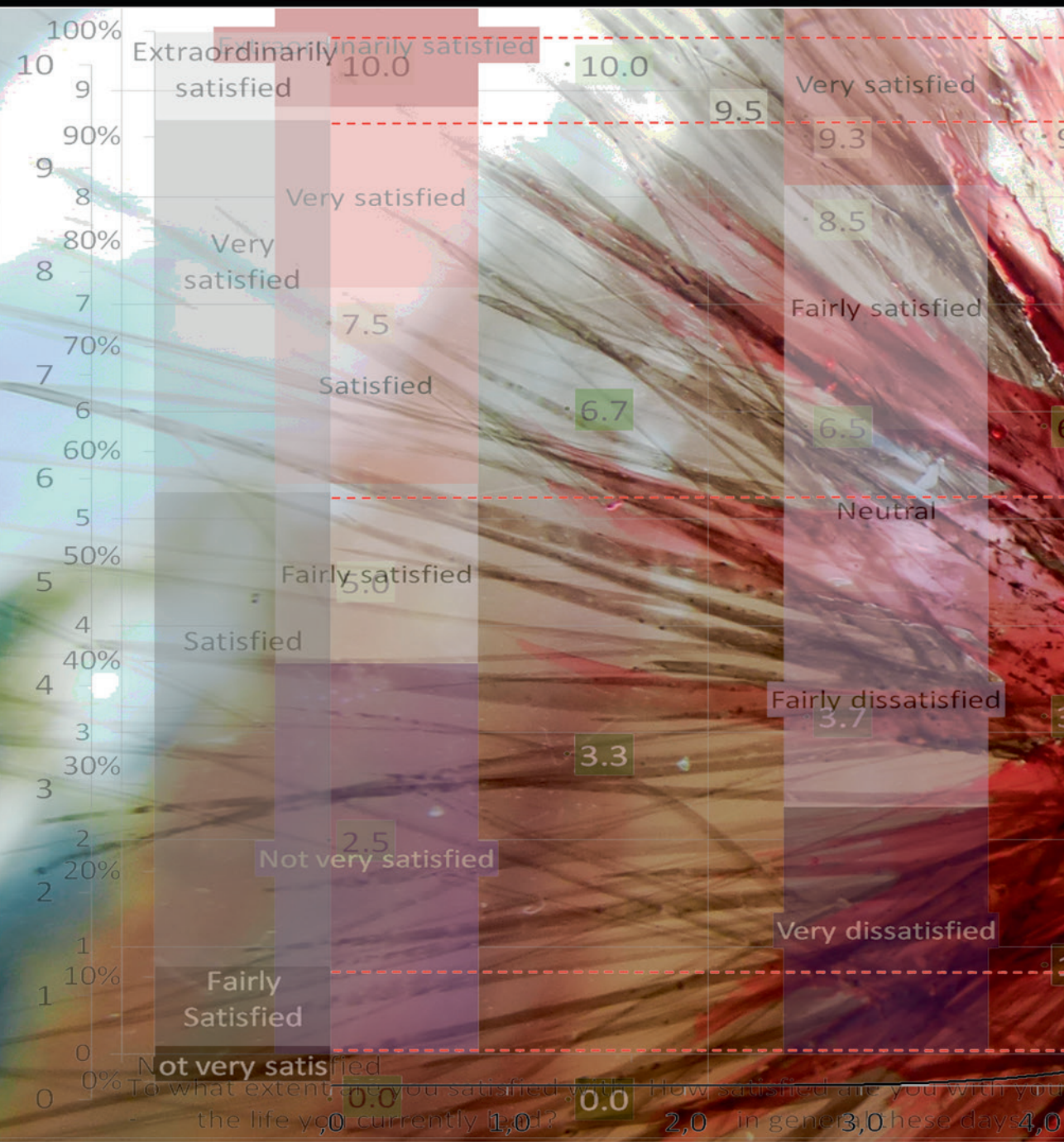
Aan Robert ben ik zonder twijfel de meeste dank verschuldigd. Zonder hem had ik niet de tijd en ruimte kunnen vrijmaken die nodig was om dit proefschrift te voltooien. Hij heeft bovendien de omslag van het boekje ontworpen. Met het resultaat daarvan, wat tenslotte als visitekaartje geldt, ben ik zeer tevreden. Ik ben heel blij om in Robert een vriend te hebben die mij laat zijn wie ik ben en die mij de vrijheid laat om de dingen te doen die voor mij belangrijk zijn.

# Curriculum Vitae (in Dutch)

Tineke de Jonge werd op 16 maart 1964 in Winschoten geboren. Zij studeerde wiskunde aan de Universiteit van Amsterdam waar zij in november 1987 haar doctoraalbul ontving. Aansluitend volgde zij de ontwerpersopleiding 'Wiskundige beheers- en beleidsmodellen' aan de Technische Universiteit Delft, die zij in maart 1990 afsloot met een Professional Doctorate in Engineering. Als onderdeel van deze opleiding werkte zij een jaar aan een ontwerpproject op het gebied van onderhoudsmodellering bij het Fysisch en Elektronisch Laboratorium van TNO in Den Haag. Daarna begon zij in de functie van projectingenieur bij het Waterloopkundig Laboratorium, thans deel van Deltares, waar zij van april 1990 tot augustus 1997 werkte, de eerste zes jaar in de Noordoostpolder, het laatste jaar in Delft. In augustus 1997 stapte zij over naar de dienst Infrastructuur, Verkeer en Vervoer van de gemeente Amsterdam om daar als beleidsonderzoeker te gaan werken. In februari 2002 verruilde zij deze functie voor die van adviseur informatievoorziening en onderzoek bij de Sociale Verzekeringsbank in Amstelveen. Daar kreeg ze alle ruimte om sociaaleconomisch onderzoek op het gebied van de volksverzekeringen te doen en om als adviseur te participeren in een aantal projecten in het buitenland op het gebied van pensioenstelsels. In de jaren bij de Sociale Verzekeringsbank voltooide zij ook de masteropleiding Beleidsonderzoek aan de Erasmusuniversiteit Rotterdam. In november 2008 veranderde zij weer van werkgever om als senior statistisch onderzoeker te gaan werken bij het Centraal Bureau voor de Statistiek in Heerlen. Vanuit die functie was zij eerst kwartiermaker en later trekker van de projecten voor de thema's 'Welzijn van de Nederlandse bevolking' en 'Belastingdruk Nederlandse Huishoudens'. Door het thema welzijn kwam zij in contact met Prof.dr. Ruut Veenhoven van de Erasmusuniversiteit in Rotterdam, die haar voorstelde om het promotieonderzoek te doen waarvan de resultaten in dit proefschrift beschreven zijn. Om hier tijd voor vrij te maken, begon zij in november 2011 als senior adviseur risicoverevening bij het College voor Zorgverzekeringen, thans Zorginstituut Nederland, in Diemen, veel dichter in de buurt van haar woonplaats. In mei 2013 verliet zij het College voor Zorgverzekeringen voor het UWV waar zij op dit moment als business controller werkt bij de divisie Werkbedrijf.

# Curriculum Vitae (in English)

Tineke de Jonge was born on 16 March 1964 in Winschoten. She studied mathematics at the University of Amsterdam where she received her master degree in November 1987. She continued her studies at the Delft University of Technology, where she followed the 2-years program 'Mathematical management and policy models' and graduated in March 1990 receiving a Professional Doctorate in Engineering. As part of this educational program in Delft, she worked on a technological design project in the field of maintenance modelling at the Defence, Safety and Security laboratory of TNO in The Hague. After graduating from Technical University Delft she started in the position of project engineer at Delft Hydraulics, at present part of Deltares, where she worked from April 1990 to August 1997, six years in the Noordoostpolder, and for the last year in Delft. In August 1997 she changed jobs and became policy researcher at the department for Infrastructure, Traffic and Transport of the municipality of Amsterdam. She left this job in February 2002 when she started working as an information and research advisor at the Social Insurance Bank in Amstelveen. In this job she was given ample time to conduct socioeconomic research in the field of social security and to participate as an advisor in a number of projects abroad with respect to pension systems. During the time that Tineke worked at the Social Insurance Bank she also took and graduated from a masters course in policy research at Erasmus University Rotterdam, The Netherlands. In November 2008 she became a senior statistical researcher at Statistics Netherlands in Heerlen. In this position she set up the program 'Well-being of the Dutch population' and a second program to study tax pressure in households paying Dutch taxes. She went on to be the project leader of these programs. The well-being theme brought her into contact with Prof. Ruut Veenhoven of Erasmus University Rotterdam, who proposed that she should do a PhD the results of which are presented in this thesis. To find sufficient time to do the PhD Tineke changed jobs and became a senior advisor risk adjustment at the National Healthcare Insurance Board in Diemen, The Netherlands, a job far closer to home. She left the National Healthcare Insurance Board in 2013 and is presently working as business controller at the Dutch Public Employment Services (UWV).


100%
10
9
90%
9
8
80%
8
7
70%
7
6
60%
6
5
50%
5
4
40%
4
3
30%
3
2
20%
2
1
10%
1
0
0%
0
Extraordinarily satisfied
10.0
Very satisfied
9.3
Very satisfied
8.5
Very satisfied
7.5
Fairly satisfied
6.5
Satisfied
6.7
Neutral
Fairly satisfied
5.0
Satisfied
Fairly dissatisfied
3.7
3.3
Not very satisfied
2.5
Very dissatisfied
Fairly Satisfied
Not very satisfied
0.0
To what extent are you satisfied with the life you currently lead?
How satisfied are you with your in general these days?
1,0
2,0
3,0
4,0