Soc Indic Res (2011) 102:497–515
DOI 10.1007/s11205-010-9688-2

Happiness Scale Interval Study. Methodological Considerations

W. M. Kalmijn · L. R. Arends · R. Veenhoven

Accepted: 21 July 2010 / Published online: 24 August 2010

Abstract The Happiness Scale Interval Study deals with survey questions on happiness, using verbal response options, such as 'very happy' and 'pretty happy'. The aim is to estimate what degrees of happiness are denoted by such terms in different questions and languages. These degrees are expressed in numerical values on a continuous [0,10] scale, which are then used to compute 'transformed' means and standard deviations. Transforming scores on different questions to the same scale allows to broadening the World Database of Happiness considerably. The central purpose of the Happiness Scale Interval Study is to identify the happiness values at which respondents change their judgment from e.g. 'very happy' to 'pretty happy' or the reverse. This paper deals with the methodological/statistical aspects of this approach. The central question is always how to convert the frequencies at which the different possible responses to the same question given by a sample into information on the happiness distribution in the relevant population. The primary (cl)aim of this approach is to achieve this in a (more) valid way. To this end, a model is introduced that allows for dealing with happiness as a latent continuous random variable, in spite of the fact that it is measured as a discrete one. The [0,10] scale is partitioned in as many contiguous parts as the number of possible ratings in the primary scale sums up to. Any subject with a (self-perceived) happiness in the same subinterval is assumed to select the same response. For the probability density function of this happiness random variable, two options are discussed. The first one postulates a uniform distribution within each of the different subintervals of the [0,10] scale. On the basis of these results, the mean value and variance of the complete distribution can be estimated. The method is described, including the precision of the estimates obtained in this way. The second option assumes the happiness distribution to be described as a beta distribution on the interval [0,10] with two shape parameters (α and β). From their estimates on the basis of the primary information, the mean value and the variance of the happiness distribution in the population can be estimated. An illustration is given in which the method is applied to existing measurement results of 20 surveys in The Netherlands in the period 1990–2008. The results clarify our recommendation to apply the model with a uniform distribution

W. M. Kalmijn · L. R. Arends (✉) · R. Veenhoven
Erasmus University, Rotterdam, The Netherlands
e-mail: arends@fsw.eur.nl

within each of the category intervals, in spite of a better validity of the alternative on the basis of a beta distribution. The reason is that the recommended model allows to construct a confidence interval for the true but unknown population happiness distribution. The paper ends with a listing of actual and potential merits of this approach, which has been described here for verbal happiness questions, but which is also applicable to phenomena which are measured along similar lines.

Keywords Happiness · Measurement · Ordinal scales · Probability distribution · Beta distribution · Parameter estimation · Scale interval approach

Abbreviations

df	Degree(s) of freedom (Section 5)
HSIS	Happiness scale interval study (Section 1)
MIV	Mid-interval value(s) (Section 3)
p.d.f.	Probability density function (Section 5)
Prob	Probability (Section 5)
WDH	World database of happiness (Section 1)

1 The Measurement of Happiness

Happiness is typically measured by self-report and cross-national studies on happiness mostly use single questions. An example of such a frequently used question is: "Taking all things together, how would you say things are these days–would you say you are… ?" The respondent is requested to make a choice out of e.g. four possible ratings:

☐ ''unhappy'' (R_1)
☐ ''not too happy'' (R_2)
☐ ''pretty happy'' (R_3)
☐ ''very happy'' (R_4)

In this example, happiness is rated by the respondent on a 4-step verbal rating scale. In this context, the possible ratings are referred to as 'categories'. This term stems from the name "the method of successive categories", as is in use for the above method of measurement among psychometricians; see e.g. Guildford (1954, Chap. 10).

In the World Database of Happiness (Veenhoven 2010), further abbreviated WDH, a set of one question and all admissible responses to that question is referred to as a "measure of happiness", previously as "item". A great many (about 1250 by the end of 2009) of alternative measures which have been reported as used in at least one survey or other study, are gathered, not only verbal ones, but also numerical, pictorial scales using 'smilies' and other graphical scales.

In most of them, the respondent has to select one out of a limited number of discrete ratings. In the above example, the four possible responses are denoted as R_1, R_2, R_3 and R_4 respectively. In general we shall use the symbol R_j for the j-th response, being a member of a set of k possible alternatives, written as $\{R_j \mid j = 1(1)k\}$; in the above example $k = 4$.

The notation $j = 2(1)5$ means that the variable j ranges from 2 to 5 with steps of size 1, so in this case $j = 2$, $2 + 1 = 3$, $3 + 1 = 4$ or $4 + 1 = 5$.

In this example, R_1 corresponds to the most unhappy situation and R_k to the happiest one. This is the most frequently occurring choice and in this paper, we will assume that this choice has been made. In the case of a scale with R_1 as the happiest situation, a simple reversion of the order of the code numbers will enable the application of the methods described in this paper.

Questions of the above type are presented to members of a sample from a population, e.g. some nation, in order to obtain information about the happiness situation in that population. The happiness distribution of such a community is defined as the probability distribution of the individual happiness values of all members of this community. This distribution is unknown, but its parameters should be estimated from the frequency distribution of the individual happiness values in the sample that represents that population. The average value and the standard deviation can be estimated from the corresponding frequency distribution parameters of the k responses $\{R_j\}$ in the sample that represents the society of the study.

The basic results in this type of investigation are the counted absolute frequencies $\{n_j\}$ at which members of that sample with size N select one out of the k alternatives $\{R_j \mid j = 1(1)k\}$. Respondents who report "Don't know" or who do not make any choice are ignored in this context.

From these absolute frequencies, we can compute the k relative frequencies $\{f_j := n_j/N\}$ and the k cumulative relative frequencies $\{F_j \mid j = 1(1)k\}$, which in the above example are defined as

$$F_1 := f_1$$

$$F_2 := f_1 + f_2$$

$$F_3 := f_1 + f_2 + f_3, \quad \text{and}$$

$$F_4 := f_1 + f_2 + f_3 + f_4 (= 1)$$

while the symbol ":=" means "is defined as". In general $F_j := \sum_{i=1}^{j} f_i$

So, the total basis information can be summarized as $\{N;\ F_j \mid j = 1(1)k\}$ under the condition $0 \le F_1 \le F_2 \le \cdots \le F_{k-1} \le F_k = 1$.

The central issue in this paper is how to convert this information into valid and useful information on the population that is represented by the sample in which the measurements have been performed. There are two major problems in this.

The first one is that happiness, as it is measured above, is always a variable at the ordinal level of measurement. It is common practice to replace the various $\{R_j\}$ with the corresponding j-value as a code, but these k code numbers are essentially ordinal numbers. This implies that it is not admissible to subject them to addition, multiplication or other arithmetical operations, which are applied in the calculation of average values, standard deviations and other current descriptive statistics; such operations are defined on cardinal numbers only. So we have to find a solution for the "cardinalization problem": how to transform the ordinal code numbers into cardinal numbers?

A second major problem is that in happiness studies happiness is measured with different rating scales, which may even have different numbers of ratings. So there is a need to transform the happiness values as they are measured primarily to a common secondary rating scale. For this common secondary scale, a scale on the interval [0, 10] is the usual choice, where the upper end always represents the most happy situation and "0" the most unhappy one.

Since in practice the solutions of both major problems are interconnected, we shall discuss them jointly.

Plan of this Paper

In Sect. 2, we shall discuss some of the methods in which the cardinalization problem is solved in practice together with the transformation to a common secondary scale. As one of the ways-out, the international "Happiness Scale Interval Study" (HSIS) is proposed. For this approach, a model is presented in Sect. 3, where a continuous happiness variable is postulated, which is mapped onto a discrete scale of measurement. In Sect. 4, the underlying assumptions are specified into more detail. In Sect. 5, three possible models are described to convert the basis measurement information into information about the happiness distribution within the population that is assumed to be represented by the sample from which the observations have been obtained. In Sect. 6, we start with a brief description of how the HSIS runs in practice and what achievements have been realized until now. As an illustration, we present the results of the application to the happiness data in 20 Dutch surveys in the period 1990–2008. On that basis we recommend the application of a specific happiness distribution model, which is not the most attractive from a validity point of view, but which allows the construction of confidence intervals for the mean population happiness value in e.g. a nation. Moreover, this section lists the potential merits of the proposed approach. To what extent these expectations are empirically confirmed will be described in a separate paper.

2 The Cardinalization Problem

The traditional approach for the further condensation of the counted frequencies is to consider happiness as a discrete variable, which can adopt only a limited number (k) of different values, which number has been chosen by the investigator. As has been pointed out above, the responses are recorded as code numbers, R_j being recorded as a "rating $= j$".

For the subsequent processing, one has to solve the already mentioned cardinalization problem. Three alternatives will be discussed below:

(1) Simple cardinalization by direct stretching;
(2) Thurstone values and related approaches;
(3) The happiness scale interval approach.

There are more alternatives, but a discussion on these is outside the scope of this paper.

2.1 First Alternative: Simple Cardinalization and Linear Stretching

The most frequently occurring solution is to fully ignore (1) the label of the categories, e.g. "unhappy", and (2) the distinction between ordinal and cardinal numbers. Although the ratings $\{j\}$ are code numbers and hence are essentially ordinal numbers, they are treated as if they were cardinal. In that case, the various possible ratings are treated as equidistant numbers on a metric [1, k] scale, in our case integer numbers in the closed interval [1, 4]. Such a scale will be referred to as "pseudo-metric".

For comparing results obtained by using different scales, the results of the primary numerical scale are often subjected to 'direct rescaling' or 'stretching', which is a linear

transformation onto a common 'secondary' scale. This linear scale transformation, has been described in e.g. Veenhoven and Kalmijn (2005, Appendix C) and in Kalmijn (2010, Appendix B).

In the above example, the primary scale is a [1, 4] scale. For the common secondary scale, we select the [0, 10] scale as usual. Then the result of the [1, 4] scale transformation would be

$1 \rightarrow 0$
$2 \rightarrow 3{,}33$
$3 \rightarrow 6{,}67$
$4 \rightarrow 10$

The three underlying assumptions for such a linear scale transformation can be summarized for this example as

(a) $1 \rightarrow 0$, where "0" on the common secondary scale expresses feelings that are identical to the feelings corresponding to either the lowest or, for inverted scales, the highest rating on all primary scales, irrespective of the phrasing of that category,
(b) $k = 4 \rightarrow 10$ in a similar way, and
(c) the primary scale is 'metric', i.e. the k ratings are considered to represent equidistant happiness intensity feelings, and so are the corresponding secondary values.

2.2 Second Alternative: Thurstone Values and Related Approaches

A possible alternative might be to request all members of a panel to place k marks on a line, one for each of the possible responses, e.g. "Please place a mark on this line, at the position of which you feel the most appropriate for the judgment 'pretty happy', irrespective of your personal happiness judgment". The 'upper' end (10) of the line represents the most happy conceivable situation of the respondent personally and the 'lower' end (0) the most unhappy conceivable one. For each category, the average position of those given by all panel members is adopted as the transformed position of that category on the [0, 10] scale.

Jones and Thurstone (1955) describe a method in which they presented 51 verbal qualifications to a panel of 905 respondents, who were requested to select the most appropriate appreciation rating on a 9-point Likert scale for each qualification separately. As a result, the 51 qualifications could be mapped on a common interval scale.

Ehrhardt has proposed to apply the basic idea of this method in a similar way to the WDH on the basis of expert ratings. In 1993 Veenhoven and twelve co-workers, all involved in happiness studies at the Erasmus University Rotterdam (NL), were asked independently to assign the number they considered the most appropriate for the position of ratings in the interval [0, 10] on a scale which was presented as continuous. This was done for each of 29 categories that were current in a number of verbal happiness measures in happiness research. Their average values obtained in this way are included in the WDH and referred to as "Thurstone values" although "Jones—Thurstone values" might have been more correct. On this basis, average values and standard deviations of samples are computed by simply replacing ordinal numbers of the categories with the corresponding Thurstone values.

In the WDH, an extensive use is made of this method, in particular for verbal scales with 3 or 4 possible ratings, for which the application of direct rescaling is highly debatable. Although these Thurstone values have been established for one specific

language (English), just like Jones and Thurstone did, it is current practice to apply them in the WDH to other ones as well.

A similar study was run by Bartram and Yelding (1973) among 166 adult regular London ITV-watchers. A number of their qualifications overlapped those of the Thurstone values; the absolute differences of the numerical values range from 0.1 to 0.7, which differences are of the same order of magnitude as the inaccuracy of those numbers.

It should be noted that the procedure according to Ehrhardt was not completely identical to that of Jones and Thurstone, nor of that of Bartram and Yelding, since Ehrhardt engaged experts vs. the 905 non-experts of Jones and Thurstone and the 166 of Bartram and Yelding.

2.3 Objections to the Above Approaches

The procedures for measuring happiness and their underlying assumptions as has been described above were not at all uncontested, but as long as no suitable alternatives are available, this has hardly any consequences. At least four objections emerge at (ir)regular intervals.

An obvious criticism with respect to the simple cardinalization concerns the equidistance assumption, lacking any evidence for small *k*-values. The Thurstone and related methods claim to resolve this problem.

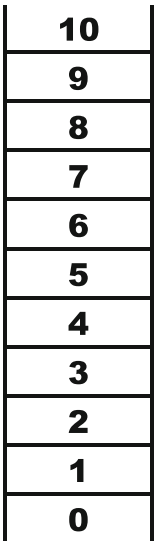
As a second, there is a validity problem in the approach in which happiness is measured as a discrete variable in its relationship to happiness as a psychological concept. The respondent has to make a forced choice out of a limited number of alternatives. However, if we consider happiness as the intensity of something in a subject's personal situation, it is obvious to look for a continuous variable rather than to a discrete one. If we managed to construct some variable that is related to happiness as measured above and that is continuous at the same time, this would improve the validity, at least in this respect.

The third class of objections especially concerns the verbal happiness ratings scales. Differences between e.g. "unhappy", "not too happy" and "extremely unhappy" are ignored as long as they refer to a lowest category of the scale. Moreover, in the comparison of studies in different nations, the usual assumption is that for Spanish people "feliz" has exactly the same significance or meaning as "happy" has for the British. Besides, it is questionable whether this meaning is the same for the Australians and for the (i.e. all) US citizens. As long as we are unable to demonstrate the existence of differences in this respect, we simply use to declare them non-existent.

Finally, there is a problem caused by the fact that happiness was measured by self-response, not only in different languages, but also by using scales with structural differences. Not all of them have equal numbers of possible ratings. Examples are known in which the same verbal expression is part of two or more scales with different values of *k*. It is most doubtful to assume that such an expression has identical significances in these different contexts.

The above objections to this practice do not concern all scales to the same extent. There exists a type of scales, known as the "Best-Worst Ladder Scales", that meets reasonably well all three underlying assumptions for direct rescaling. As an example, we mention the adapted version of Cantril's self-anchoring ladder rating of life (Cantril 1946; Kilpatrick and Cantril 1960). The respondent is presented with Fig. 1 and with the question: "Here is a picture of a ladder. The '10' at the top of the ladder means the best possible life you can imagine. The '0' at the bottom of the ladder means the worst possible life you can imagine.

Fig. 1 Cantrill's ladder scale



On which place of the ladder is your life as a whole? Please mark the number that best corresponds with how you feel about your life now."

On the other hand, violation of the assumptions is presumably rather strong for verbal scales that has been described in Sect. 1. Especially for relative small values of k, say for $k \leq 4$, we strongly dissuade linear scale transformation.

2.4 Third solution: The Happiness Scale Interval Study (HSIS)

In order to encounter a number of the above problems, Veenhoven (2009) has started his International "Happiness Scale Interval Study". In this study, local judges are requested to partition the total [0, 10] continuum into k intervals in such a way, that each of them corresponds to one of the k possible response ratings. In the example, each panel member has to identify his or her subjective boundary between "unhappy" and "not too happy", as (s)he sees that boundary, and (s)he is expected to do so irrespective of one's own happiness. More details are given in Sect. 6.

The proposed approach does not pretend to solve all problems concerning measurement of happiness, nor that of life satisfaction etc., but it (cl)aims at reducing at least a number of them, especially the above ones.

3 The Model Underlying the Happiness Scale Interval Study

The model underlying the Happiness Scale Interval Study postulates the existence of a variable, here denoted H, that—in this application—expresses the intensity of the feelings of happiness of a respondent. In this description, we will deal with the application of the model to the measurement of happiness, but it is equally applicable to the measurement of life satisfaction or some other related subjective self-judgment of the respondent's hedonic situation.

To this variable H the following properties are assigned:

I. H is postulated to be a variable, measured at the metric level of measurement and expressed as a real number in the closed interval [0, 10].

II. the value $H = 0$ represents the respondent's subjectively worst conceivable situation with respect to his or her happiness, whereas $H = 10$ represents the subjectively best conceivable situation. This choice excludes the possibility of any H-value outside the [0, 10] interval.

III. H is an intensity variable and is a strictly increasing continuous function of the happiness intensity as experienced by the respondent: if a person at the moment t_2 feels happier than at the moment t_1, then $h_2 > h_1$, where h_1 and h_2 are the H-values at t_1 and t_2 respectively;

IV. the variable H is a latent variable. It is unobservable as such, but can be mapped by the respondent onto a set of k different verbal, numerical or pictorial observable ordered qualifications (ratings) $\{R_j \mid j = 1(1)\ k\}$, k being a natural number, usually $k \leq 12$. The order of the qualifications is assumed to be unambiguous;

V. the interval [0, 10] can be partitioned into k contiguous subintervals, each of which being defined as the subset of H-values that are mapped to the same image. All these intervals are right-hand closed half open intervals, except the closed interval including the value $H = 0$;

VI. the above mapping is monotonous, while the subinterval with the largest H-values is mapped as the happiest qualification R_k.

VII. the variable H is a random variable; within a population, it has a probability distribution: different individuals in that population will have a happiness which is represented by generally different H-values.
In general, different populations will have different probability distributions of H. These are of the same type, but have different values of the parameters.

VIII. except for $H = 0$ and $H = 10$, the H-values of the subinterval boundaries are subjective, since the interpretation of the possible responses is subjective as well. This applies especially to verbal qualifications, which may have a strong cultural component. Not only the language/nation combination will influence their interpretation, but also conditions as social class, age etc.; moreover the emotional value of terms may shift over time. Therefore, in linking H-values to qualifications, especially the verbal ones, some degree of variability in the results is to be expected.

As an example, we consider the next situation (Fig. 2).

In this model, there is a one-to-one correspondence between each of the k presented different qualifications R_j and one of the intervals of $\{h\}$. The upper boundary of the jth subinterval will be denoted as b_j and this half-open subinterval as $(b_{j-1},\ b_j]$, with $j = 1(1)k$, $b_o = 0$ and $b_k = 10$. For convenience reasons, the set $\{(b_{j-1},\ b_j], j = 1(1)k\}$ is assumed to include the closed interval $[b_o,\ b_1]$ as well. The values $\{b_j;\ j = 1(1)k - 1\}$ are also referred to as 'cut points'; however, this term is usually extended to include also the values $b_o = 0$ and $b_k = 10$. We shall use the terms "boundary values" and "cut points" as synonyms.

In this way, there is also a one-to-one relation between each qualification R_j and the mid-interval value (further abbreviated MIV) of the jth interval, which is defined as $m_j := \frac{1}{2}(b_{j-1} + b_j)$.

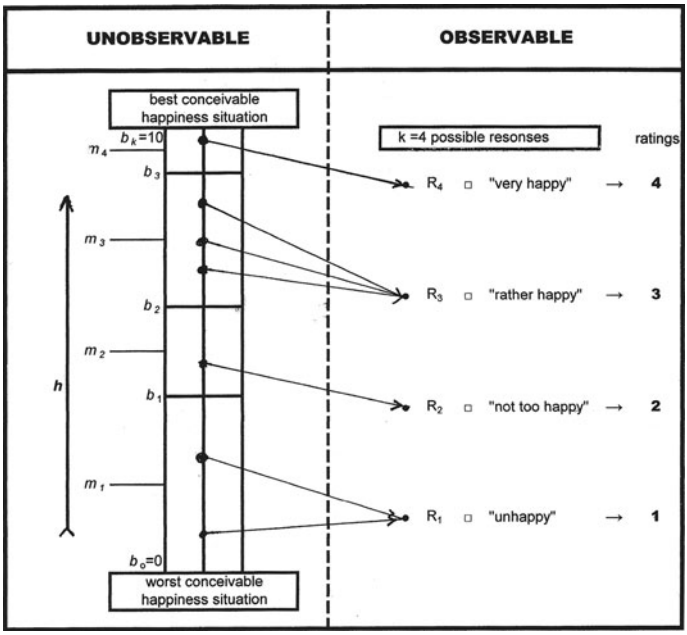


Fig. 2 Representation of model for happiness scale interval study

4 Further Assumptions of the Model

In an ideal world, there would be complete consensus about the H-values of all subinterval boundaries. However, under VIII in the previous section, it has already pointed out why individual opinions on the same boundary are expected to differ.

Each panel member is requested to report the value of H at which in his personal opinion a shift ought to be made towards a "more happy judgment category". The average value of these judgments is adopted as the estimated cut point position to be used in the application phase later on.

The basic assumption of this approach is that every respondent in the application phase with $R = R_j$ will report this rating on the basis of his happiness feeling which corresponds to an H-value in the interval $(b_{j-1}, b_j]$. However, it is conceivable that for some respondent in the sample $b_j < H_i < (b_j)_i$, where $b_j =$ the estimated cut point position as obtained in the construction phase, $(b_j)_i =$ his personal opinion on the position of the boundary between the j-th and the $(j + 1)$-th interval and H_i his personal happiness value. This respondent will report "R_j", and in this way the observed frequency of the j-th category is overestimated. This bias may, however, be compensated by an other respondent to whom $(b_j)_i < H_i < b_j$. Unless the distribution of individual opinions around their average value is very skewed, the net bias is assumed to be negligible and we will make this assumption, at least for the moment.

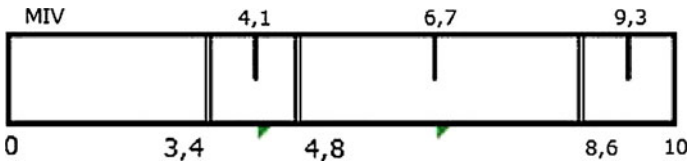
Two identical phrasings, but within different items, are judged in the HSIS separately and independently within each item. This practice was not applied to the determination of the Thurstone values nor to similar other approaches. The proposed practice is justified in the comparison of the mid-interval values (MIV) of the judgment "very satisfied" within two different items of the WDH as an example. Item coded O-SLW/c/sq/v/5/p raises the

question "All things together, how satisfied are you with your life as-a-whole these days?" with five response categories: completely satisfied/very satisfied/satisfied/not very satisfied/not at all satisfied. In item O-SLS/c/sq/v/3/a it is asked: "How satisfied are you with the way you are getting on now ?" with three response categories: very satisfied/all right/not at all. On a [0, 10] scale, the MIV of "very satisfied" for these different questions with different alternatives were 7.6 and 8.9 respectively, which demonstrates that the other categories and their phrasings should not be ignored.

Intuitively, one might expect that the average result of all respondents in the determination of the Thurstone and related values, whether or not done by experts, is a good estimate for the MIV as defined in the HSIS. The answer to the question whether this expectation is correct is negative, at least in general. The reason is that the k MIV are not mutually independent. They have to satisfy a simple criterion which can be described as follows: write down the supposed MIV in descending order of magnitude and connect them with alternating minus and plus signs, starting with a minus sign. Then the result in the case of a [0, 10] scale should be equal to 5. In the case for $k = 4$, one gets $m_4 - m_3 + m_2 - m_1 = 5$. If the 'alternating sum' $\neq 5$, the $\{m_j\}$ cannot be considered to be MIV. This proof of this rule is to be found in Kalmijn (2010, Appendix F3).

After substitution of (the positions of) some set of four marks in the above equation, the 'alternating sum' will in general $\neq 5$, and in that case these four average positions$\{m_j\}$ cannot be considered to be a set of unbiased estimates of the MIV. In case of modest departures from this condition, some adjustment procedure of the marks position may be a 'solution' to deliver a more or less valid estimation of the MIV. In practice, however, it appears that it is rather exceptional when acceptable results are obtained along these lines.

Consequently, generally speaking, Thurstone values cannot be considered as pseudo-MIV, since usually they do not satisfy our criterion that their alternating sum equals the value 5. This is easily demonstrated for the scale example in Sect. 1. The Thurstone values of the four responses in the WDH have been agreed to be {0.6; 4.1; 6.7; 9.3}. Since $9.3 - 6.7 + 4.1 - 0.6 = 6.1 \neq 5$, the set of Thurstone values of this item clearly does not satisfy our MIV criterion, in this particular case not even approximately. This can also be demonstrated by the graphical representation below. Suppose that all Thurstone values are MIV, and that at least the largest three of them are correct. Then the boundary values are {0; 3.4; 4.8; 8.6; 10} Consequently the smallest Thurstone value in this case should be 1.7 and not 0.6.



5 Conversion of the Sample Data to Information About the Population Happiness Distribution

The happiness distribution of a community is defined as the probability distribution of the individual H-values of members of that community. This population probability distribution is unknown, but it can be estimated from the frequency distribution of the individual H-values in the sample that represents that population. The expected or mean value and the standard deviation can be estimated from the corresponding frequency distribution parameters of the k responses $\{R_j\}$ in the sample that represents the community of the study.

If the variable H is assumed to be a random variable, it will have a cumulative distribution function, denoted as $G(h)$:= Probability $\{H \leq \mathrm{h}\}$. This $G(h)$ is a monotonically nondecreasing function of h with $G(-\infty) = 0$ and $G(\infty) = 1$.

In the case H is assumed to be a discrete random variable, $G(h)$ is a step function with k steps, one at each value h that H can adopt, the size of the j-th step being $\mathrm{Prob}\{H = h_j\}$.

If however H is assumed to be a continuous variable, $G(h)$ is a continuous function. Now we define:

$$g(h) := \frac{dG(h)}{dh}$$

provided it exists, which derivative is called the probability density function (p.d.f.) of H. Whether or not g(h) exists depends on the further assumptions made on $G(h)$.

We will discuss three possible models, which have been represented in Fig. 3. Under the model described in Sect. 3, it is assumed that each respondent with a happiness feeling corresponding to any H-value in the interval $(b_{j-1}, b_j]$ will respond as R_j. However, all we know is the number of respondents with R_j, but it is unknown which H-value in the interval $(b_{j-1}, b_j]$ belongs to each of them. Therefore, we have to make assumptions on the unknown distribution of H over [0, 10], more precisely, over each of the k intervals $\subset$ [0, 10]. The three models differ in these underlying assumptions.

I. In model I, it is assumed that all respondents giving the same response $\mathrm{R_j}$ are equally happy and have the same H-value, for which the MIV of the jth interval is the obvious one to be selected. These k responses are the only ones available, not only for the sample members, but also in the population as a whole. In other words, the population probability distribution of H is assumed to be discrete with only k possible H-values.
II. The variable H is assumed to be continuous and has a distribution which is uniform over each of the k intervals.
III. The variable H is assumed to be a continuous variable with a beta distribution. From the observations, estimates of the two model parameters α and β are calculated, Subsequently, estimates of the mean and the variance of the distribution are calculated on the basis of these estimates of α and β.

A more detailed description of the three models will be given below.

An important property of any estimator is whether it is biased or not. If θ is a parameter or a function of one or more parameters of a probability distribution of some random variable, and is estimated by a statistic $\hat{\theta}$, then the bias of $\hat{\theta}$ with expectation $\mathrm{E}(\hat{\theta})$ is defined as the difference $\mathrm{E}(\hat{\theta}) - \theta$, where θ is either a scalar or a vector, and $\hat{\theta}$ will be accordingly. It should be emphasized that a bias is defined only if the distribution of the statistic is known and that it depends on which type of probability distribution is adopted for the random variable. Hence the same statistic, which is an unbiased estimator for some parameter in model I and/or II may not necessarily be unbiased for the same parameter in e.g. model III.

5.1 Model I: The Discrete Approach

One way-out could be to locate all respondents in the middle of the interval and to use the MIV as an estimate of the H-value of all of them.

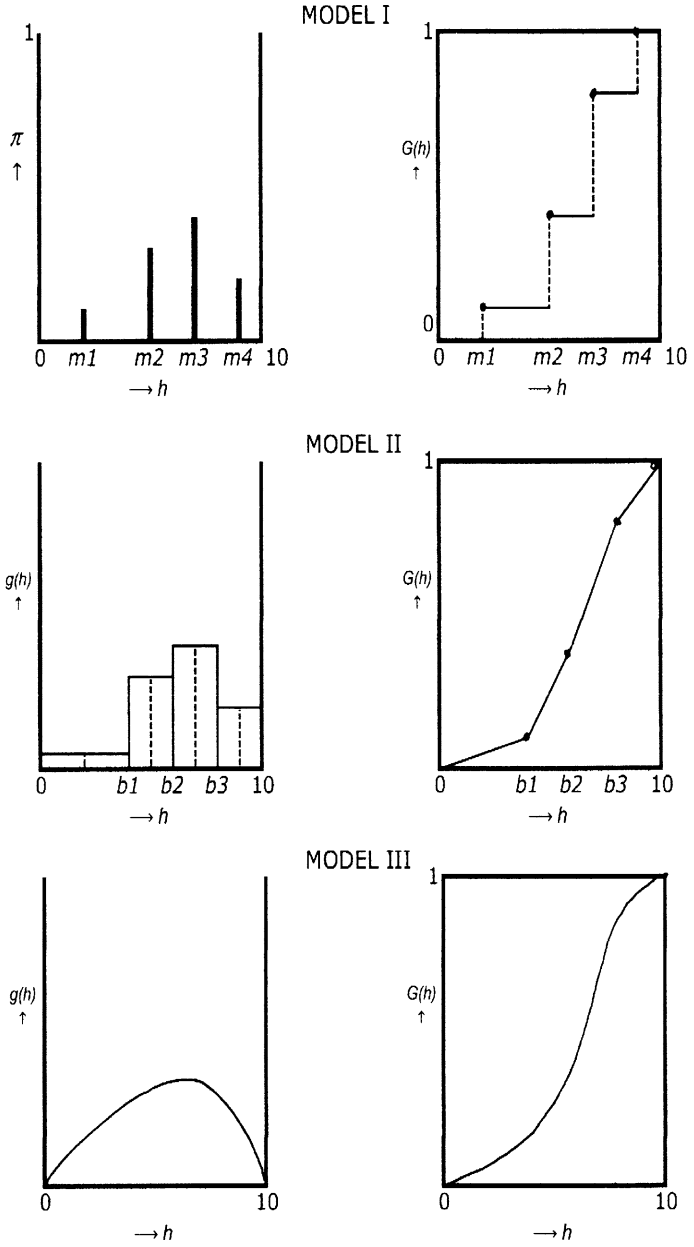


Fig. 3 Probability values and densities (left) and cumulative Probabilities (right) for $h \in [0,10]$ in three models: I (discrete distribution), II(semi-continuous distribution) and III (beta distribution), all on the basis of a four-point rating scale

This approach is rather similar to the traditional one and yet considers happiness as a discretely distributed variable. The essential difference is the replacement of the transformed code number of the categories with the empirical MIV, but the conversion of

sample results into information on the population happiness distribution follows identical lines.

In the traditional approach, it is very unusual to specify the probability distribution in the population explicitly. Implicitly, the situation in the population is assumed to be structurally identical to that of the sample, but with larger size only. The same assumption is made in this model I. This means that this population probability distribution is assumed to be a discrete polytomous distribution with $2k$ parameters, k for the probabilities $\{\pi_j | 0 \leq \pi_j \leq 1, j = 1(1)k, \sum \pi_j = 1\}$, and k for the mid-interval values, $2(k-1)$ of which parameters being independent. The parameters $\{\pi_j\}$ are defined as $\pi_j :=$ the probability that an individual, 'selected' at random from the population, will report R_j. They are estimated as the k relative frequencies in the sample. In that case the sample mean is an unbiased estimator of the mean happiness of the population probability distribution. The second moment about the mean of the sample is made an unbiased estimator of the population variance by the application of Bessel's correction, i.e. by replacing the denominator n with $n - 1$. Its square root is underestimating the value σ of the population systematically, but since this estimator is consistent, usually the sample size is sufficiently large to neglect this bias.

In this model, the cumulative probability distribution $G(h) := \text{Prob}\{H \leq h\}$ is a step function with a step of size π_j at $H = b_j$ for $j = 1(1)\ k - 1$, where at each step the value of $G(h)$ is the higher one.

5.2 Model II: The 'Semi-Continuous' Model

A second alternative is to assume that all H-values in an interval are equally likely, i.e.to assume a uniform distribution of H over each of the k intervals separately. In that case, consecutive points in the cumulative distribution plot with co-ordinates $(b_{j-1}, G(b_{j-1}))$ and $(b_j, G(b_j))$ are connected by straight line segments, making $G(h)$ a broken line with kinks in all cut points where $H = b_j$. At these H-values, $G(h)$ is not differentiable, so there $g(h)$ does not exist. Consequently, in this approach $g(h)$ is a step function with steps in $H = b_j$ for all $j = 1(1)k$ and horizontal lines of different elevations in between. In other words: at each cut point, the probability density is changing stepwise to remain constant until the next boundary/step.

As long as no explanation can be offered for such steps at a number of points, all selected by the investigator, such a model is not very satisfactory. A sufficiently realistic model should at least satisfy the condition that its p.d.f. is continuous over the complete interval (0, 10). We refer to the model II as "semi-continuous", since it assumes the happiness variable H to be continuous, while its probability density function is not.

Just like the model I, the model II has $2k - 2$ parameters. As long as no better alternative is available, we have to accept this model. The consequences of this assumption for the estimation of the population mean and variance have been described in Kalmijn (2010, Appendix F1), including those for the precision of these estimators.

5.3 Model III: The Beta Distribution as Continuous Model

Because the model II is not satisfactory in all respects, there is at least one alternative to be considered. This is known as the beta distribution, which has a continuous density function of a random variable in a closed interval with finite boundaries (see e.g. Kendall and Stuart 1977; 35 and 46).

As applied to our situation, the cumulative distribution function is defined by:

$$dG(h) = [10 \cdot B(\alpha, \beta)]^{-1} h^{\alpha-1}(10-h)^{\beta-1}\, dh,$$

in which B(α, β) is the complete beta function with parameters α and β, defined as:

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

This model of the beta distribution has only two parameters, α and β, which are positive real numbers; they are usually referred to as the two shape parameters of the distribution. This number of parameters is considerably smaller than in the models I and II, because in this model, there are no categories at all in the population distribution. The density function $g(h)$ is continuous over the complete domain, finite and positive for all $h \in (0, 10)$ and zero outside the interval [0, 10]. All relevant properties and other information on this application of the beta distribution have been summarized in Kalmijn (2010, Appendix H), most of which can be found in various textbooks on calculus and statistics and/or in other public sources, e.g. Gupta and Nadarajah (2004).

In applying this distribution as the model, the empirical frequency information, available as $\{F_j \mid j = 1(1)k\}$, is compared to the corresponding values of $G(b_j)$, minimizing the differences between F and G jointly. The value of G is dependent on both α and β for all $\{b_j \mid j = 1(1)k-1\}$.

The comparison of F and G is possible and meaningful only at $k-1$ values of H $\{b_j \mid j = 1(1)k-1\}$, since the equations $F(0) = G(0) = 0$ and $F(10) = G(10) = 1$ are trivial. The situation can be considered as one with a screen before the cumulative distribution function G(h), which is observable only through one of the $k-1$ very narrow windows at $H = b_j$ ($j = 1(1)k-1$). From these $k-1$ comparisons, the two model parameters $\{\alpha, \beta\}$ are to be estimated, leaving $k-3$ degrees of freedom (df).

For $k = 3$, there is always a unique solution with a perfect fit.

For $k = 2$, the number of solutions for this underdetermined situation is infinite.

For $k \geq 4$, we have an overdetermined situation and in general there will be no perfectly fitting distribution, so we have to look for the 'best fitting' solution.

If one has found this distribution, it would be possible to a apply a 'goodness- of-fit test' (see e.g. Cramér 1974; 416–424). For this situation, K. Pearson has proposed a test statistic, which is based on the multinomial distribution of N respondents over k possible responses and which is defined as

$$\sum_{j=1}^{k} \frac{\left(n_j - En_j|\mathrm{H_o}\right)^2}{(En_j|\mathrm{H_o})}$$

where $En_j|\mathrm{H_o}$:= the expected value of $\mathrm{n_j}$ under the null hypothesis $\mathrm{H_o}$ that the estimated distribution is a perfect representation of the actual distribution in the population. Under $\mathrm{H_o}$ and under some additional conditions, Pearson's statistic is approximately distributed as chi-square (χ^2) with in our case $k-3$ degrees of freedom (df). These conditions are that $k > 3$, that N is not too small and that responses with $En_j|\mathrm{H_o} \leq 5$ are 'pooled' with an adjacent response, which is obviously done at the cost of the number of df due to the effective reduction of k. Such a test in other than comparative situations is well debatable from the point of view of standard statistical test theory.

The two parameters of the beta distribution cannot be interpreted directly as a location and a dispersion parameters as is the case for e.g. the normal distribution. From the

relationship between α, β, μ and σ^2, the mean μ and the variance σ^2 of the distribution of H can be estimated by direct substitution of the estimates of the shape parameters α and β:

$$\mu = \frac{\alpha}{\alpha + \beta}$$

and

$$\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

In general, the values of the estimates obtained in this way will not be identical to those of the corresponding sample statistics. However, they may be more valid as they allow for the assumption of a continuous random variable H with a continuous p.d.f. over (0, 10).

The beta distribution also enables one to compute a potentially useful in a comparative study of nations, especially in relationship to other characteristics. It is the "percentage happy", which is defined in this context as the percentage of the society for which the happiness, expressed as the H-value, is closer to their most happy situation than to the most unhappy one, i.e. for which $H > 5$. In the above notation, this proposed statistic is defined as the estimate of $[1 - G(5)]\cdot 100\%$, and can be computed on the basis of the estimates of the parameters α and β. Since the value of this statistic is influenced by both the mean value and the variance of the distribution, it may be considered as a possible alternative to the 'Inequality-adjusted happiness' as has been described by Veenhoven and Kalmijn (2005).

6 Application and Merits of the Model

6.1 The HSIS in Practice

The application of the HSIS method is a two-step process. The first one is the scale construction phase by a panel as has been described by Veenhoven (2009), and the second is its application to characterize the happiness of a population by a sample of subjects using this scale. Note that we use the terms 'panel' and 'judges' for the scale construction phase and 'sample' and 'respondents' for the application phase as a contribution to strengthen the distinction—and the separation—of these two phases.

In the HSIS, the judges in the construction phase have to identify their personal opinions with respect to of the $k - 1$ cut points $\{b_j \mid j = 1\ (1)k - 1\}$, bearing in mind that $b_0 = 0$ and $b_k = 10$ are fixed. For a given measure of happiness, the values of the $k - 1$ boundaries or cut points have to be estimated as the average values reported by n panel members. Each of these judges has to specify the above mapping by indicating the b-values he feels to separate the consecutive categories, ignoring his personal happiness self-judgment.

In the second phase, the outcomes of the first phase are applied to the observed frequencies of the various categories as counted in a sample of N subjects from the relevant population. From these results, the sample mean and its happiness inequality are calculated, the latter being expressed in the standard deviation. These statistics are used to compute estimates of the parameters of the distribution of the variable H in the population represented by the study sample. As a matter of fact, both stages will contribute to the eventual inaccuracy of these estimates.

We have to emphasize that the application phase of the methods described in this paper is only applicable to samples of which the 'complete' empirical sample cumulative distribution $\{F_j \mid j = 1(1)k\}$ is known, albeit for k happiness values only. Knowledge of both the average value and the standard deviation of the sample happiness only is insufficient.

6.2 First Results

Since the start of the HSIS, a large amount of data has been gathered. Of the first harvest, 100 cases have been analyzed. The observations are also available http://worlddatabaseofhappiness.eur.nl/scalestudy/datafiles/first100cases.xls and the results have been described by Kalmijn (2010, Chap. VII).

These data has been delivered by 12 institutes and cover 9 different languages. In this context, a case is defined as the set of judgments on the cut points of a specific happiness measure (one leading question + k response categories), obtained within the same participating institute and the same session. The total number of happiness measures involved is 52, since several measures have been presented to judges in more than one institute.

6.3 Some Findings as Illustration

Five of these cases have been applied to 20 already existing happiness distribution data from Dutch surveys in the period 1980–2008. As an illustration, the results have been summarized in Table 1.

For each of the five cases, denoted A, B, C, D and E, we included the text in English of the leading question and all response categories. Each row below this description refers to one of the existing surveys, the year of which has been specified. In the next columns, the estimated mean values have been listed according to the different approaches. We start with the traditional approach (happiness as a discrete variable and equidistant ratings ranging from 0 to 10). Then follows the estimate obtained on the basis of Thurstone values. In the next column, we report the estimate according to the models I and II as described in Sect. 5; both models always give identical estimates for the mean happiness value. Moreover, we have calculated the estimates on the basis of the best fitting beta distribution. In Table 1, we recorded the difference between the latter estimate and the one according to the models I/II. Next there are two columns with estimates of the within-nation standard deviation, one according to the traditional method and the other one on the basis of model II. Finally, the right hand column gives the 95% confidence limits for the true, but unknown mean happiness value of the happiness of the Dutch population.

The number of judges in the panel was about 30, the sample size in the application phase varied between 1000 and 1500, except for case D, in which much larger samples were involved. For comparison reasons we considered the average happiness value measured by using numerical scales. In all those cases, the leading question was at least very similar and incidentally even identical to the one of the verbal scales. Over the total period 1990–2008 this estimate varied between 7.4 and 7.8 on a [0, 10] scale.

From this table, we conclude that there are substantial differences between the estimated mean values. These do not only depend on the text of the happiness measures and the number of categories, but also on the model according to which the observations have been processed. Moreover, the agreement with the above estimates on the basis of the use of numerical scales is not always excellent.

From a validity point of view, the model III on the basis of a beta distribution is the most attractive one, but it has one serious disadvantage: we are unable to estimate the

Table 1 Estimated mean values and standard deviations 1980–2008 in The Netherlands

Case	Year	Estimated mean in different models				SD		CI95 model II
		Traditional	Thurstone	Models I and II	III-II difference beta distribution	Trad	II	CI for true unknown population mean happiness value
O-HL/c/sq/v/3/ab: Taking all things together, how would you say things are these days? Would you say you are…? (1) not too happy; (2) pretty happy; (3) very happy								
A	1982	6.9	7.6	7.1	+0.18	3.1	2.1	[6.7; 7.4]
	1983	7.1	7.7	7.2	+0.17	2.9	2.0	[6.8; 7.5]
	1984	6.9	7.6	7.1	+0.17	2.9	2.0	[6.8; 7.4]
	1985	6.7	7.5	6.9	+0.17	2.9	2.0	[6.6; 7.3]
	1986	7.0	7.6	7.1	+0.16	2.9	2.0	[6.8; 7.5]
O-HL/u/sq/v/4/a: Taking all things together, would you say you are…..? (1) not at all happy; (2) not very happy; (3) quite happy; (4) very happy.								
B	1981	7.7	7.7	7.5	+0.16	1.7	1.5	[7.2; 7.7]
	1990	8.0	8.0	7.7	+0.16	2.2	1.8	[7.5; 7.9]
	2006	7.9	7.9	7.6	+0.15	2.0	1.7	[7.3; 7.8]
	2006	7.9	7.9	7.6	+0.14	2.0	1.7	[7.4; 7.9]
O-SLL/u/sq/v/4/b: On the whole, how satisfied are you with the life you lead? (1) not at all satisfied; (2) not very satisfied; (3) fairly satisfied; (4) very satisfied.								
C	1981	7.8	7.5	7.5	+0.13	2.3	1.9	[7.3; 7.7]
	2000	7.9	7.6	7.5	+0.13	2.1	1.7	[7.3; 7.8]
	2006	8.1	7.7	7.7	+0.16	2.1	1.7	[7.5; 7.9]
	2007	8.3	7.8	7.8	+0.16	1.9	1.5	[7.6; 8.0]
	2008	8.2	7.8	7.8	+0.15	2.0	1.7	[7.6; 8.0]
O-SLL/c/sq/v/5/d: How satisfied are you with the life you currently lead? (1) not so satisfied; (2) fairly satisfied; (3) satisfied; (4) very satisfied; (5) extraordinary satisfied.								
D	1980	5.7	8.3	6.7	+0.09	2.5	1.9	[6.4; 7. 1]
	1997	5.8	8.5	6.8	+0.06	2.2	1.7	[6.5; 7.1]
	2000	5.9	8.5	6.8	+0.14	2.2	1.6	[6.5; 7.2]
	2002	5.8	8.5	6.8	+0.07	2.2	1.7	[6.5; 7.2]
	2004	5.8	8.5	6.7	+0.13	2.2	1.7	[6.4; 7.1]
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? (1) completely unhappy; (2) very unhappy; (3) fairly unhappy; (4) neither happy, nor unhappy; (5) fairly happy; (6) very happy; (7) completely happy.								
E	2002	7.1	7.3	7.3	+0.06	1.4	1.4	[7.1; 7.5]

inaccuracy of the estimates, at least on the basis of our present knowledge. As a consequence, we are unable to construct 95% confidence intervals for the true but unknown population mean value. The application of model II does not have this disadvantage. From the column III–II we learned that the difference between the fully continuous and the semi-continuous model is modest (<0.2) and that this difference is always well within the 95% confidence interval. Our final conclusion is that eventually the model II is to be preferred

over the beta distribution model III, since it does not only provide us estimates, but also information about their inaccuracy.

A more elaborate analysis and discussion is given by Kalmijn (2010, Chap. VII).

6.4 Potential Merits of the Scale Interval Approach

The main possible merits of the above approach–some of which are potential–can be summarized as follows:

(a) Improvement of the validity of the method in that sense that the proposed approach considers happiness no longer as a discretely distributed variable, but allows for its continuous nature. In this way, the method described in this paper is no doubt closer to reality and is to be considered more valid, so more relevant for social scientists than previously conventional methods were.
Moreover, as compared to the method of direct rescaling, the criticism on the latter method does not apply to the results obtained according to the scale interval approach. This especially includes the objections against the controversial treating of ordinal ratings as if they were cardinal, since in the proposed approach, no equidistance between the ratings is no longer assumed.

(b) A consequence could also be an improvement of correlational findings, at least in the validity perspective. Moreover, it is conceivable, at least theoretically, that this improvement of the validity of the happiness measurement may also result in higher numerical values of the association measures with conditions of happiness. Such an expectation would be based on the assumption that associations that are really present, may be blurred by the fact that happiness is measured in a suboptimal way rather than due to the fact that the associations are intrinsically insufficiently strong.

(c) Meta-analytical studies are almost always hampered by the problem that different findings that need to be combined arise from the application of different WDH items. It is to be expected that the results obtained according to the scale interval approach will be more reliable than those obtained according to previously current methods, so the method may seriously enlarge our meta-analytical opportunities. Similar considerations can be applied to the investigation of trends of happiness in nations or other societies.

(d) Finally, the method enables the opportunity to optimize the set of questions. Items with a relative large skipping rate, with a large interval width inequality and/or in which a relatively poor consensus about the positions of the boundaries has been observed within panels and/or between panels from different nations, are less suitable than those without these problems. All these observations could be good reasons to discontinue the application of these problematic happiness measures, although a number of studies will still remain where they have been applied in the past. In this way, the present approach may contribute to the standardization and improving the quality of measuring happiness.

In a next paper we will evaluate the application of this approach to a number of verbal scales and test to what extent the underlying assumptions and the model can be corroborated or not.

References

Bartram, P., & Yelding, D. (1973). The development of an empirical method of selecting phrases used in verbal rating scales: A report on a recent experiment. *Journal of the Market Research Society, 15*(2), 151–156.

Cantril, H. (1946). The intensity of an attitude. *Journal of Abnormal and Social Psychology, 41*, 129–135.

Cramér, H. (1974). *Mathematical methods of statistics* (13th ed.). Princeton, US: Princeton University Press.

Ehrhardt, J. (1993). Unpublished paper.

Guildford, J. P. (1936/1954). *Psychometric methods* (2nd ed.). New York/Toronto/London: McGraw-Hill Book Company.

Gupta, A. K., & Nadarajah, S. (Eds.). (2004). *Handbook of beta distribution and its applications*. New York, US: Marcel Dekker.

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*. Dissertation. Erasmus University Rotterdam (NL).

Kendall, M. G., & Stuart, A. (1977). *The advanced theory of statistics* (4th ed.). London/High Wycombe(GB): Ch. Griffin.

Kilpatrick, F. P., & Cantril, H. (1960). Self-anchoring scale: A measure of individuals' unique reailty world. *Journal of Individual Psychology, 16*, 158–173.

Veenhoven, R. (2009). International scale interval study: Improving the comparability of responses to survey questions about happiness. In Moller, V & Huschka, D. (Eds.), Quality of life and the millennium challenge: Advances in quality-of-life studies, theory and research (Vol. 35, pp. 45–58). Social Indicators Research Series, Springer, e-ISBN 978-1-4020-8569-7.

Veenhoven, R. (2010). World Database of Happiness (WDH): Continuous register of research on subjective appreciation of life. Erasmus University Rotterdam, The Netherlands. Available at: http://www.worlddatabaseofhappiness.eur.nl.

Veenhoven, R., & Kalmijn, W. M. (2005). Inequality-adjusted happiness in nations. *Journal of Happiness Studies, 6*(4), 447–449.