Master Thesis

Do you feel more satisfied if it will be asked differently?

The influence of question tone, question wording & scale wording on self-reported life satisfaction

By

Rahul Mangroe

Student number: 373246rm

Supervisor: Martijn Hendriks

Second reader: Spyridon Stavropoulos

Master of Science

in

Behavioural Economics

Date final version: December 2018,

Acknowledgements

Throughout this research process, I have enjoyed the help and support of many kind people around me. Some of them I would like to mention here.

I would like to express my deepest gratitude and appreciation to my supervisor Martijn Hendriks. His guidance, clear instructions and willingness to respond on my many questions helped me to write this thesis. Furthermore, I would like to thank the second reader Spyridon Stavropoulos for reading and evaluating my work.

I also thank my friends for supporting and motivating me. Finally, I would deeply want to thank my mother and my father for taking care for me all the time. They have always supported me to reach my goals and making my dreams come true. At last, I thank my little brother to whom I always try to be a role model.

Abstract

Many countries measure life satisfaction, but they do it in different ways. This thesis empirically investigates whether the question tone, the question wording and the question scaling has an influence on the self-reported life satisfaction. This research uses self-collected data. The sample size of this research is 664. The data for the experiment is non-parametric. Tests that are used are the Mann-Witney U test and the Kruskal-Wallis test (for medians), the Brown Forsythe test (for distributions) and the Bootstrap (for correlations). The question tone, question wording and the question scaling do not influence the self-reported life satisfaction. A possible explanation for this could be due to a small sample size. Limitations for this study are that data is mostly collected at the Erasmus University. Also, there is only a between-subjects design. With a within-subject design it is easier to detect differences across levels of the independent variables.

Contents

1. Introduction

Since ancient times, people are interested in the factors that contribute to their well-being. Thousands of years ago the ancient Greek philosophers already debated about the well-being of individual persons within the society. Subjective well-being (SWB) can be defined as people's cognitive and emotional evaluations of their lives (Diener et al., 2002). The use of the term SWB became more popular because it was used as a parallel to hedonic well-being (Ibid.). Three components should be taken into account when measuring SWB. These components are life satisfaction, positive affect and negative affect (Andrews, 1974). According to Park & Seligman (2005), life satisfaction refers to people's global evaluation of the quality of their life.

Over the past decades, more countries started measuring SWB and life satisfaction. Many of these countries developed frameworks for measuring aspects of well-being (Global Happiness Council, 2018). SWB measurements capture important quality of life elements; they also provide a reflection on inequalities and offer a tool to measure well-being over time. In addition, Layard (2005) states that the use of SWB for economic and social policy purposes has been increasing. Governments, for example, use SWB measures to gain knowledge about the society to make changes to their pursued policies when required. SWB measures give governments the option for building up policies that are evidence-based and also create a shared understanding of what contributes to better lives (Global Happiness Council, 2018). It could foster further public debate about life satisfaction and could be used as a tool for promoting evaluation of the impact of specific policy programs on people's lives. For instance, academics and the media also use SWB measures to give information about the society and to understand the society better. It is therefore important to know what the purpose of the life satisfaction measures is. The Organisation for Economic Co-operation and Development (OECD) (2013) provides four main purposes. An SWB measure can complement other outcome measures. For example, migrants in a country can give different measures on aspects of life concerning life satisfaction than the 'local' inhabitants. A measure can also help to improve the understanding of the drivers of SWB: critical aspects of people's well-being can be identified. The third main purpose is that measures assist in evaluating policies and play an indirect though important role in cost-benefit analyses. The final main purpose is to identify potential policy problems as the measures give insights in human behavior.

Different measures of SWB in life satisfaction research could result into different results. The problem, in practice, is that SWB measures indeed do differ across countries. This could give significantly biased results when SWB data from different countries are compared (Dolan, 2012). There is no international consensus in the definition of SWB. This makes measuring SWB a challenging task: most conclusions about the measures are drawn from empirical research where the measurement of life satisfaction is done by questionnaires. In general, the measures have to be valid and accurate. In practice however, there can be measurement errors. This could lead to wrong conclusions: it can make the relationship between two components either stronger or weaker. Approximately one quarter of the variance in a subjective measure might be due to systematic sources of measurement errors (Podsakoff et al., 2003). Other measurement errors could be due to the formulation of the measures (questions). For example, a combined question tone (positive and negative) generates longer, though also more negative answers than just a positive or a negative tone (Brennan, 1997). Although there are some concerns with the SWB measures, Diener et al. (2013) state that on the whole, SWB can be measured in a sufficient and accurate manner.

Although SWB measurements (life satisfaction in particular) play an increasingly important role in public policy and research about the quality of life, there are some concerns among economists about the validity and reliability of such measures. One of these concerns is that the precise questioning, for example the phrasing, will influence the results. This paper addresses this concern by looking at the extent to which survey item wording influences the reported life satisfaction. Table 1 provides some examples of similar life satisfaction questions used in different questionnaires.

Table 1: Examples of different survey questions

Type of survey	Question
European Social Survey	All things considered, how satisfied are you with your life as a whole nowadays?
World Values Survey	All things considered, how satisfied are you with your life as a whole these days?
SOEP	How satisfied are you with your life, all things considered?
British household panel	How dissatisfied or satisfied are you with your life overall?
Latinobarometro	Generally speaking, would you say you are satisfied with your life?

There is not much literature concerning survey item wording. It would be inconvenient if the life satisfaction medians and correlates would depend on how life satisfaction is measured. This concern motivates the research goal of this paper. One purpose is to create awareness among policy makers and researchers to what extent the formulation of life satisfaction measures influence the self-reported life satisfaction of people. It is investigated to what extent the formulation of life satisfaction measures influences the life satisfaction that people report. Specifically, the paper explores to what extent question tone, question wording and scale wording are influencing the medians, distributions and correlates of self-reported life satisfaction. To meet these research goal, a questionnaire-experimental approach is utilized.

The structure of this paper is as follows: in Section 2, the theoretical framework of the research is described. A more detailed overview of SWB and the main causes for measurement errors are given and it is checked in more detail what kind of research has been done for each subpart of the research question. The data and methodology are covered in Section 3. Results are reported in Section 4. Section 5 discusses and concludes.

2. Theoretical framework

2.1 SWB self-reports and measures

The 'life as a whole' is the union of different life domains. To understand the concept of life domains, Sirgy (2012) introduced a system that reflects a hierarchy of psychological concepts. This hierarchy is shown in Figure 1. Life satisfaction is on top, domain satisfaction in the middle and satisfaction of events within a specific domain is at the bottom of the hierarchy. This is the cognitive structure of the hierarchy. A person can have affective experiences that are related to, for example education, work, social life health et cetera. Because of this, the memory can also be divided into different life domains, which can also be organized in a hierarchy. This is shown on the left side in Figure 1. Feelings about life overall are on top. In the middle, there are feelings about certain life domains. Each life domain is divided into life events within the domain. An example of cognitive evaluation of a life domain is the statement: 'I feel good about my family life.' This is stored in the memory. Affective evaluations are the emotions a person experiences for a certain domain outcome. One needs to understand the concept of domain hierarchy to understand the strategies people use to report their SWB.

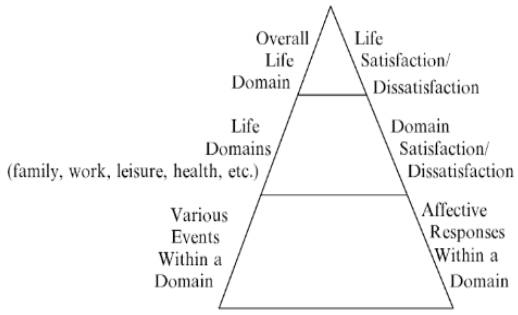

Figure 1: The domain hierarchy

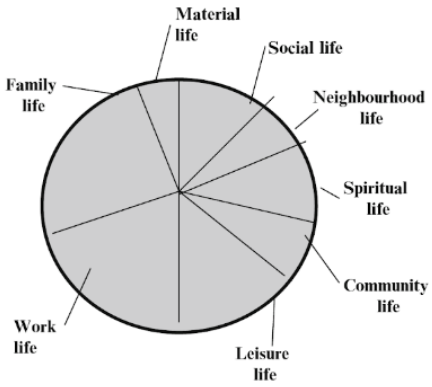

Figure 2: A graphic representation of domain salience

Domains in which a person has given much effort to attain positive affective evaluations are more likely to stand higher in the salience hierarchy. All life domains vary in salience: some life domains may be more important than others for a person. Figure 2 shows an example of how a person can value certain life domains. This person has high values on the domains work, family and leisure.

2.1.1 How do people answer life satisfaction questions

Response biases can have a large effect on the validity of measures in general. People make use of heuristics to answer certain questions. This can lead to people having their own response style if the heuristic is used repeatedly. Suppose that, for example, a question contains vague

words. In that case, the respondent might use his own interpretation for the words to give an answer. Another example is when the respondent's answer is not in the list of possible answers in a (multiple choice) question. The respondent then could give the answer closest to his true answer, but his honest answer is unfortunately not measured. The order of the questions also has an effect: when the most 'important' question is asked in the middle of the questionnaire, it is likely that the respondent gives less serious responses to the questions thereafter.

2.2 The validity and reliability of life satisfaction measures

Life satisfaction measures have to be reliable: the same results should be produced in repeated experiments. A tool to measure reliability of SWB measures is the test-retest correlation: when the same question is asked twice in a research, the respondent should (approximately) give the same answer. It is, however, a little bit harder to use this as a test for reliability for SWB because the answers of the respondents are rather influenced by the scale than by the time lag (Hoorn, 2008).

Another area where SWB measures can differ is across countries: cultural and linguistic factors can affect SWB ratings. However, there is not much evidence on this area (Hoorn, 2009). In general, language does not seem to be a problem. Ouweneel and Veenhoven (1991) compare self-reports from bilingual countries with reports from countries where the languages are from. They did not find significant differences. Helliwell and Barrington-Leigh (2010) did research on Canada and compared their results with the rest of the world. Patterns of SWB are very different across Canada than across the world, looking at individuals, societies, provinces and nations. However, they show that the results would not differ much if a country has the same set of life circumstances, making the measurements reliable.

Cultural characteristics appear to be more important factors and possible sources of biases (Hoorn, 2009). Scollon et al. (2004) did research on five different cultural groups of students. All three SWB areas (positive and negative emotions and life satisfaction) were assessed with three different methods. For nearly all measures, cultural differences were a source of difference between the groups. Research on countries in the Pacific Rim from Diener et al. (1995) shows that there was a difference between reported SWB when the Pacific Rim is compared to the United States, but cultural differences were not an explanation for this. The measures do not have to be biased on the other hand because SWB is differentially structured in terms of its predictors (cultural characteristics in this case) (Hoorn, 2009).

There are some other factors that influence the reported SWB. For example, 20-40% of SWB variance is due to measurement errors and external factors, like the weather. Variance from external factors can be minimized but not fully eliminated and therefore it has to be accepted that the measures can be biased, though not systematically. Variance from measurement errors can be minimized by using large (many respondents) and long (many days) datasets.
Such factors do not make the measurement systematically biased. This is an advantage of the methods stated above (Hoorn, 2008)

Much literature has been written about the validity of SWB measures. Diener et al. (2013) find evidence by combining results from many studies, concluding that self-reported life satisfaction measures correlate with other types of well-being measures that do not depend on reports from respondents, i.e. measures that do not significantly correlate with measures taken from family or friends of the respondent. Furthermore, measures from groups of respondents who have 'good' life circumstances differ significantly from measures of groups of respondents who have less 'good' life circumstances. Other studies where the respondents were twins show that the measures always have a positive correlation. These are all arguments in favour of life satisfaction self-reports being valid. However, there are no universal criteria the measure has to satisfy to be valid (OECD, 2013).

In the research of Rodgers et al. (1988), respondents evaluated several domains of their life. Life satisfaction measures were compared with scores obtained from different measures. More than a half of the total variance seems to be valid variance. 10% is method variance and one third is due to measurement errors et cetera. The measurement variance can be minimized by taking errors that can be made in the research into account. Podsakoff et al. (2003), discuss common method variance (common method bias particularly) extensively. Many potential sources of common method bias are provided. They divided the sources in four groups. Method effects produced by: (i) a common source or rater: the respondent providing the measure of the independent variable(s) and the dependent variable is the same person, which can cause correlation between these variables. (ii) Item characteristics: the way items (questions in this research) are presented to respondents. (iii) Item context: the interpretation the respondent has on the item and (iv) measurement context: time, location and media used to get the measures.
Even though it is possible that causes from all four groups can appear in any given study, the focus in this paper will be mainly on the second group. The causes that are most likely to be encountered in these groups are given in the following subsections.

2.3 The influence of item wording on life satisfaction scores

2.3.1 Effects of question tone

The influence of question tone will be researched by changing the tone of the life satisfaction question. There will be a positive tone: “All things considered, how satisfied are you with your life as a whole these days?” To measure the negative tone the word “satisfied” is replaced by “dissatisfied”. For the neutral tone “satisfied or dissatisfied” is used. Besides changing the tone there is also an open wording. Hereby the words “do you feel about” is used instead of “satisfied”.

To understand the usage of different tones, different theories will be briefly discussed. There are five plausible perspectives that could confirm the influence of question tone on life satisfaction. Those perspectives are: framing, anchoring and priming, item demand characteristics and item social desirability.

First of all framing occurs when (small) changes of a question give different results with respect to the original question (Chong and Druckman, 2007)**.** Framing the life satisfaction question positively, neutrally or negatively could give other results according to the framing theory. Framing is a powerful tool in, for example, politics and the media. President Trump gave a rally and during that rally he insulted some minority groups. Some newspapers could defend the performance of Trump by framing his words as right of free speech. Other newspapers could reject Trump’s point of view and could frame his statements as a threat for maintaining social order. Readers of different newspapers are encouraged to emphasize certain considerations above others in evaluating a specific event (Chong and Druckman, 2007).

The second perspective, anchoring, means that initial pieces of information provide a benchmark for participants that affects their judgements (Tversky & Kahneman, 1974). By using words as ‘dissatisfied’, participants will automatically think about things that could be better in their lives. With words like ‘satisfied’ the opposite occurs. Anchoring also means that subjects make estimates by starting from an initial value that is adjusted to yield the final answer (Tversky & Kahneman, 1974). Most of the time these adjustments are typically insufficient, because often subjects estimate a result based on an incomplete computation. For example, anchoring effects arise when subjects in an experiment are asked to estimate various quantities, stated in percentages. For instance, the percentages of African countries in the United Nations.

In this experiment the subjects got different starting points, and those starting points had a marked effect on estimates (Tversky & Kahneman, 1974).

Thirdly, when people are exposed to certain words, this can influence their behaviour. This is called priming (Voicu, 2015). Questions that contain the words 'how do you feel' may give other results than questions using words like 'satisfied' or 'dissatisfied'. *Item priming effects* refer to the fact that the positioning of the predictor (or criterion) variable on the questionnaire can make that variable more salient to the respondent and imply a causal relationship with other variables. Sgroi et al. (2010) give an example where the respondent can give a lower rating on life satisfaction when the question asked prior to the question about life satisfaction is about the recent experience of a negative life event. This question (temporarily) influences the level of life satisfaction. However, they find that priming in general does not influence SWB much. This is also a form of *context-induced mood*: the question (or set of questions) induces a mood for responding to the remainder of the questionnaire. Deaton and Stone (2016) conclude that only the respondents who give negative answers to such questions lower their reported SWB, so context-induced mood could influence life satisfaction measurements.

Finally, Podsakoff et al. (2003) also provide some causes that could have some effects on question tone. *Item social desirability* refers to the fact that items may be written in a way that reflects more socially desirable attitudes, behaviours, or perceptions. The fourth and fifth perspectives (*item demand characteristics* and *item social desirability*) refer to the fact that items may convey hidden cues as to how to respond on them (to be socially acceptable). The research of Rasinski (1989) is a good example for these causes. Respondents are asked if they thought the government is spending too much money, too little money or about the right amount of money on issues like environment, health, education, infrastructure, social security and recreation, but also on crime, drug addiction, defense, foreign aid and welfare. Conclusions are that minor wording changes can alter the meaning of a survey question: some questions on issues (e.g. aid to the poor, social security) were worded in such a manner that a positive result would occur if the government spending is increased. For example, one question was asked using 'solving the problem', while another was asked using 'assistance', which is a more neutral tone. There were more positive results for 'solving the problem' than there were for 'assistance'. It could be that in the context of life satisfaction measures changes in wording also play a role. For example, the respondent could share the opinion that the government is spending too little on these issues (item demand characteristics).

Secondly, education is not necessarily correlated with the reaction on specific question wording (Schuman and Presser, 1977). They did an extensive research on question wording (although they did more research on the form of questions, rather than the wording), initially reasoning that poorly educated respondents would be triggered by emotionally toned words or by the presence or absence of a response alternative. Better educated respondents would understand the general point of a question and would not be affected by emotionally toned words or by the meaning of responses. This turns out to be different. Education did not seem related with how respondents reacted to different types of question wording.

Some empirical research has been done on the role of question tone in surveys in general, and also more specifically in relationship to SWB surveys. For example, Smith (1987) did research on question tone with welfare as question theme. He demonstrated that when the word 'welfare' is used in a question, it produces more negative responses than when the words 'assistance for the poor' or 'caring for the poor' are used. When measuring the support of welfare state, it is better to use 'poor' in the questions than 'welfare' because 'welfare' brings up associations with other concepts, for example government budget waste.

After researching relevant theories and previous literature the following hypothesis comes forth (with sub-hypotheses):

H_1: The question tone of life satisfaction measures influences self-reported life satisfaction

H_{1A}: Median life satisfaction is highest when the question tone is positive and lowest when the question tone is negative

H_{1B}: The question tone of life satisfaction measures influences the dispersion of self-reported life satisfaction

H_{1C}: The question tone of life satisfaction measures influences the correlates of self-reported life satisfaction

2.3.2 Effects of question wording

To examine whether question wording has an influence on the results or not, there will be two kinds of comparisons. The first comparison will be a time frame vs. no time frame comparison. The time frame question will be as follows: "All things considered, how satisfied are you with your life as a whole these days?" In the question that has no time frame the phrase "these days"

is left away. The second comparison is made by using equivalent word terms. The term "All things considered" will be replaced by "overall".

Not much empirical research is done considering the way question wording will be researched in this paper. However, we could link some causes from Podsakoff et al. (2003) to the methods used in this research and provide empirical results on new areas. The first comparison (time frame vs. no time frame) can give *item characteristic effects*. Properties of some words can influence the way the respondent interprets the question. With a time frame like "these days", the respondent would think of more recent events in his life. Without a time frame, the respondent would think about events in the further past that he can still remember.
On the other hand, the life 'as a whole' is a single-item scale. This is how SWB was measured when researchers began to find it interesting. A disadvantage of such a scale is that is becomes less reliable over time (Diener, 1984).

The second comparison (using equivalent words) has to do with *item ambiguity*. Ambiguous items allow respondents to respond to them systematically using their own heuristic or respond to them randomly. The "all things considered" question wording could give different results than the "overall" question wording, because both phrases could be interpreted different, while the same is asked.

After researching relevant theories and previous literature the following hypothesis comes forth (with sub-hypotheses):

H_2: The question wording of life satisfaction measures influences self-reported life satisfaction

H_{2A}: The question wording of life satisfaction measures influences the median of self-reported life satisfaction

H_{2B}: The question wording of life satisfaction measures influences the dispersion of self-reported life satisfaction

H_{2C}: The question wording of life satisfaction measures influences the correlates of self-reported life satisfaction

2.3.3 Effects of scale wording

The last thing that will be researched is the scale wording of the life satisfaction questionnaire. All scales will be from 0 to 10. The common scale that is used is as follows: 0 = completely

dissatisfied; 10 = completely satisfied. To examine whether the scale wording has an effect on the results, the option 'completely dissatisfied' will be replaced by 'not at all satisfied' and 'totally dissatisfied'. These scale wordings are used to create unipolar ('not at all satisfied') and bipolar ('totally dissatisfied') scales. Afterwards the results of the different scale wordings will be compared.

The actual difference between unipolar and bipolar scaling in this experiment is the fact that bipolar gives a two-way response whereas the unipolar scale gives a one-way response. If the scale wording 'totally dissatisfied' is seen as 0 and 'totally satisfied' as 10, a respondent can indicate whether he is not satisfied (scaling 5), which indicates neutrality or he can indicate that he is totally dissatisfied. This is the difference with using the phrase 'not at all satisfied', where the responder cannot indicate if he is dissatisfied.

Davern and Cummins (2006) researched whether scaling bipolar/unipolar has an influence on participants. A bipolar response format is anchored by bipolar affective antonyms. It is a two-way scale and a midway score of neutral resides midway between these options, whereas a unipolar format or one-way response scale involves only a single affect dimension (Davern & Cummins, 2006). There is an important difference between these models in their prediction regarding the relationship between pleasant and unpleasant affect. In a bipolar continuum an individual may not be happy or sad at the same time, so theoretically these feelings are correlated with each other. On the other hand, some scholars argue that positive and negative affect are independent of each other and therefore will not be strongly correlated in general. Therefore, a unipolar response scale also measures life dissatisfaction, or subjective ill-being. Furthermore, participants clearly experienced greater difficulty in rating life dissatisfaction using the two-way dissatisfaction-satisfaction response scale (Davern and Cummins, 2006).

The most widely used measurement instrument in social, behavioral and psychological research are rating scales with labeled endpoints (Schwarz et al., 1991). Although there is only little empirical research found about scale wording regarding to life satisfaction questionnaires, some researchers are suggesting that scale wording has an influence on results of surveys in general (Schwarz et al., 1991). Respondents could for instance be confused by the disambiguate meaning of scale labels. As a result, this could possibly lead to different interpretations of a question item and the respondents' different subjective scale anchors. According to Schwarz et al. (1991) even the most unambiguous words show a range of meaning, or a degree of semantic

flexibility that is constrained by the particular context in which these words occur. When, for example, respondents in a survey were asked to rate their success in life, the meaning of success is left for the interpretation of the respondents. There is no common ground in being unsuccessful, successful and the like. Depending on how respondents interpret the term, the numeric values changed the meaning of the endpoint labels, resulting in different responses (Schwarz et al., 1991). Furthermore, the response alternatives that are provided to respondents do constitute a source of information that respondents actively use in determining their task and in constructing a reasonable answer.

The comparison between 'completely' and 'totally' would be interesting to investigate to test if using synonyms has an influence on the results. The expectation is that the influence of these differences in scale wording would not be significant. On the other hand, the results of the wordings 'completely'/'totally' in comparison with 'not at all satisfied' can give highly significant results. That is because of the framing effect. 'Completely'/'totally' sounds more extreme. Therefore, participants would be less likely to fill in 0 or 10 than in questions where 'not at all satisfied' is used instead.

After researching relevant theories and previous literature the following hypothesis comes forth (with sub-hypotheses):

H_3: The polarity but not the use of synonyms in answering scales of life satisfaction influences self-reported life satisfaction

H_{3A}: The polarity but not the use of synonyms in answering scales of life satisfaction influences the median of self-reported life satisfaction

H_{3B}: The polarity but not the use of synonyms in answering scales of life satisfaction influences the dispersion of self-reported life satisfaction

H_{3C}: The polarity but not the use of synonyms in answering scales of life satisfaction influences the correlates of self-reported life satisfaction

3. Data & methodology

3.1 Data

The data for this research is collected by many parties. First of all, the self-collected data is from students on the campus of the Erasmus University. Students filled in the survey on the phone or on the computer. Other students filled in the survey after receiving the link of the survey. The number of respondents that filled in the survey was 71 (N=71). Besides this, there was more data available for this research.

Another student collected data through social media platforms. Furthermore, students that followed the minor "Quality of Life and Happiness Economics" also filled in the survey. There were also participants from a Massive Open Online Course (MOOC), namely 'deception detox'. At last, students of a guest lecture of Martijn Hendriks filled in the survey. The total sample size for this research is 664.

3.2 Design of the experiment

As mentioned in the introduction, the need for an experiment regarding the way life satisfaction questions are formulated is high. Governments tend to put more priority in measuring life satisfaction. If this is not done at the same way, the results could possibly not be compared with each other. A good way of doing this is with an experiment.

The advantage of the experiment is that it isolates the effect of survey item wording by holding all other circumstances constant and randomly assign people to experimental conditions. Also, this experiment can be done multiple times in order to see if the results are valid or not. To test if whole countries are effected by the way life satisfaction questions are formulated, costs more time and could be more expensive.

Table 2 shows which names for each survey are used such that they can be easily referred to.

Table 2: Questions and framings overview

Version	Framing	Question
V1	Positive (satisfied)	All things considered, how satisfied are you with your life as a whole these days? (0=completely dissatisfied; 10=completely satisfied)
V2	Neutral (satisfied or dissatisfied)	All things considered, how satisfied or dissatisfied are you with your life as a whole these days? (0=completely dissatisfied; 10=completely satisfied)
V3	Open wording (do you feel about)	All things considered, how how do you feel about your life as a whole these days? (0=completely dissatisfied; 10=completely satisfied)
V4	Negative (dissatisfied)	All things considered, how dissatisfied are you with your life as a whole these days? (0=completely dissatisfied; 10=completely satisfied)
V5	Framing (overall)	Overall, how satisfied are you with your life as a whole these days? (0=completely dissatisfied; 10=completely satisfied)
V6	Scaling (not at all/completely)	Overall, how satisfied are you with your life as a whole these days? (0=not at all satisfied; 10=completely satisfied)
V7	Time frame (life)	All things considered, how satisfied are you with your life? (0=completely dissatisfied; 10=completely satisfied)
V8	Time frame (life), Scaling (totally/totally)	All things considered, how satisfied are you with your life? (0=Totally dissatisfied; 10=Totally satisfied)

The survey consists of numerous questions. For this research only a few of these questions will be used. First of all, there are 8 survey versions, where the life satisfaction question is different. Furthermore, there are some treatment variables that are measured by survey questions. In Table 3, a quick overview of all the variables is shown.

Table 3: Dependent and independent variables

Dependent variables	Independent variables
V1 Positive (satisfied)	Children
V2 Neutral (satisfied or dissatisfied)	Partner
V3 Open wording (do you feel about)	Level of education
V4 Negative (dissatisfied)	Employment
V5 Framing (overall)	Gender
V6 Scaling (not at all/completely)	Age
V7 Time frame (life)	Income
V8 Time frame (life)Scaling (totally/totally)	

As said earlier, there are 8 versions of the life satisfaction question. The first version is the way the life satisfaction is measured in the World Values Survey (scaled from 0-10 instead of 1-10). Versions 2-4 are used to measure the question tone. These questions are variants to test the effect of the question tone. Version 5 is how the life satisfaction question is recommended by the OECD report but with the World Values Survey scale. Version 6 is the same as version 5

but with the OECD scales. Version 7 is the HILDA survey but with the World Values Survey scale. Version 8 is the same as version 7, only with the HILDA scales.

3.3 Hypothesis testing

3.3.1 Comparisons to test the hypotheses

Before testing the hypotheses, the descriptive statistics for each survey item will be checked. This is done to see whether the participants of the different survey versions have the same characteristics. For a good overview the distribution of the life satisfaction will be compared per survey. This is all done by running descriptive statistics. Afterwards each hypothesis will be tested.

The first hypothesis that will be tested is: *"H1: The question tone of life satisfaction measures influences self-reported life satisfaction."* Firstly, the positive (V1), the neutral (V2), the negative (V4) and the open wording (V3) questions will be jointly compared. Secondly, these four questions will be compared with each other individually. The aspects that will be compared are the means and the standard deviations.

The second hypothesis that will be tested is: *"H2: The question wording of life satisfaction measures influences self-reported life satisfaction."* Hereby V1, V5 and V7 will be compared with each other. These three questions are exactly the same except for the fact that in V5 the term 'overall' is used instead of 'all things considered', and the fact that there is a different time frame in V8 ('with your life' instead of 'with your life these days'). Because of the uniqueness of these three questions, these will not be jointly tested.

The third hypothesis that will be tested is: *"H3: The polarity but not the use of synonyms in answering scales of life satisfaction influences self-reported life satisfaction."* Hereby V1, V6 and V8 will be compared with each other. The reason for choosing these three questions is because the scale wordings are different for each of them. These versions will also be jointly compared.

3.3.2 Tests

To assign which tests will be used for rejecting the hypotheses or not, it has to be determined whether the data is parametric or not. Therefore, it has to be tested whether the data has a normal distribution. To do so, the Kolmogorov-Smirnov test is used. The results of this test

states that the data is non-parametric (all P-values of the results of the test were 0,000). Therefore, it can be concluded that the data is non-parametric.

To test whether the medians of the questions that will be compared (as explained in the previous paragraph) are significantly different some tests will be used. First, the Mann-Whitney U test tests whether there is a difference between two groups. The Kruskal Wallis test is to compare more than two groups jointly. Both tests are used when the groups are independent from each other, which is the case.
The second thing that will be tested to answer the hypotheses is whether the standard deviations significantly differ from each other or not. To test this, the Brown Forsythe test will be used. Results for the medians and distributions are significant if the P-value is smaller than 0,05.
The test is whether the correlates differ between the different versions. In order to test this, it will be researched if the correlates between the variables and the life satisfaction questions differ significantly from each other. In order to do so, the confidence intervals of the correlations will be calculated by performing a bootstrap. Bootstrapping is the fact that from a certain sample, each time a random number is chosen. This is done numerous times. From this the median will be made and that will be the confidence interval.

4. Results

This section presents the results of the analysis that is made in order to determine to reject the hypotheses or not. To get a better understanding of the data, the descriptive statistics will be reviewed. Afterwards, the medians and distributions will be examined for each sub-question. In Table 4 an overview is given of the mean, median and standard deviation of each survey version.

Table 4: Descriptive statistics of N-value, mean, median and std. deviation

	V1	V2	V3	V4	V5	V6	V7	V8
N	84	86	92	86	87	55	88	86
Mean	6,95	7,01	6,75	7,03	7,18	7,29	7,25	7,19
Median	7,00	7,00	7,00	7,00	7,00	7,00	8,00	8,00
Std. Deviation	1,605	1,538	1,948	1,662	1,506	1,286	1,783	1,712

4.1 Descriptive statistics

There is no significant difference between the different survey versions and the socio-demographic characteristics. Nevertheless, there will be a review of the descriptive statistics of the survey versions.

As seen in Table 5 there are no major differences in the correlates in comparison with the different surveys. Approximately 20% to 30% of the participant speaks English as first language. In all surveys approximately 90% of the participants have no children.

In version 3 of the surveys there is quite a different distribution between having a partner or not (53% has a partner and 47% does not have a partner). In version 5 there is an approximately equal division between participants that have a partner and participants that have not. In the other versions, approximately 40% does not have a partner. Although there is a difference between versions 3, 5 and the other versions, this is not a reason to assume that there are statistically significant socio-demographic differences in having a partner or not. This is illustrated at the end of this section.

Among the variable 'gender' there are also some differences. Overall, the participants are around 40% female and 60% male. In version 5, 70% is male. In version 6 the distribution between male and female is equal. Not much people find that their health status is bad or poor. The gross of the participants finds their health status satisfactory or good. Approximately 20% to 30% states that their health is very good.

Only a small group of the participants did not study further than primary and/or secondary education. Most of the participants attended tertiary education.

In order to analyze the income level of the participants, they are divided in seven groups of different income levels. The income level is an income that participants earn monthly. Because there were people from different countries, there also are different currencies. Therefore, the currencies are calculated to the euro such that the results can be compared. As seen in the table below, most participants earn between €0 and €5000. People that earn more are outliers or are participants that filled in their yearly income.

Most of the participants are between 18 and 30 years old. The reason for this is because of the fact that most participants of the survey are students. The participants that are between 31 and 60 years old are approximately 10% of the people that participated.

Most of the participants are already working full-time (between 50% to 65%). This is in contrast with the fact that most of the participants are students. An explanation for this is that most of the students that participated are finishing their education and/or already work full-time.

Table 5: Dependent variables

		V1	V2	V3	V4	V5	V6	V7	V8
English first language	Yes	21%	26%	20%	28%	22%	22%	21%	26%
	No	79%	74%	80%	72%	78%	78%	79%	74%
Children	Yes	10%	10%	12%	5%	8%	9%	7%	12%
	No	90%	90%	88%	95%	92%	91%	93%	88%
Partner	Yes	38%	40%	53%	39%	52%	44%	40%	45%
	No	62%	60%	47%	61%	48%	56%	60%	55%
Gender	Male	41%	46%	32%	43%	30%	52%	41%	37%
	Female	59%	54%	68%	57%	70%	48%	59%	63%
Health Status	bad	1%	1%	0%	0%	1%	0%	1%	0%
	poor	8%	7%	3%	3%	6%	6%	1%	1%
	satisfactory	22%	21%	20%	26%	22%	26%	23%	24%
	good	49%	44%	49%	47%	51%	37%	48%	45%
	very good	19%	26%	29%	24%	19%	31%	27%	30%
Education	primary	3%	1%	1%	1%	0%	2%	2%	0%
	secondary	33%	31%	30%	25%	22%	44%	24%	31%
	tertairy	62%	61%	66%	68%	74%	48%	70%	61%
	vocational	3%	7%	3%	5%	4%	6%	4%	8%
Income	0-1500	44%	40%	39%	43%	45%	54%	41%	33%
	1500-3000	7%	7%	9%	8%	7%	9%	1%	13%
	3000-5000	8%	6%	5%	8%	7%	2%	9%	7%
	5000-10000	6%	4%	4%	8%	8%	6%	6%	7%
	10000-40000	20%	30%	27%	15%	24%	17%	23%	24%
	40000-80000	11%	11%	8%	11%	7%	7%	11%	10%
	>80000	6%	3%	8%	8%	4%	6%	9%	7%
Occupation	unemployed	15%	15%	9%	17%	12%	9%	9%	14%
	part time	26%	21%	24%	22%	21%	24%	21%	23%
	full time	51%	57%	61%	55%	59%	57%	65%	63%
	self employed	7%	4%	4%	4%	8%	6%	6%	0%
	retired	1%	3%	1%	1%	0%	4%	0%	0%
Age	0-17	1%	1%	0%	1%	0%	9%	2%	0%
	18-21	33%	38%	29%	39%	43%	44%	41%	38%
	22-30	53%	46%	57%	46%	46%	31%	46%	47%
	31-60	11%	14%	12%	13%	9%	13%	11%	12%
	>60	1%	1%	1%	0%	1%	2%	0%	3%

With a Kruskal Wallis test it is examined whether there are socio-demographic differences between the different versions or not. From Table 6 it can be concluded that there are no significant differences between the eight versions and the socio-demographic characteristics.

Table 6: Test for socio- demographic differences

	Health	English first language	Children	Partner	Education	Gender	Age	Household income	Self employed	Retired	Employed
Kruskal-Wallis	5,391	2,667	3,415	6,725	9,168	9,928	5,221	7,319	6,581	7,228	4,844
P value	0,612	0,914	0,844	0,458	0,241	0,193	0,633	0,396	0,474	0,406	0,679

4.2 Results of question tone

To research whether people are influenced by the way the life satisfaction question is formulated or not, the influences of the question tone will be discussed firstly. There are four different ways to set a different question tone. There is a positive tone (V1), a neutral tone (V2),

an open wording tone (V3) and a negative tone (V4) . The output of all the tones are jointly tested with a Kruskall-Wallis test for medians and a Brown-Forsythe test for distributions.

The P-value for the Kruskall-Wallis test is 0,271. This implies that the medians of the four question tones jointly do not differ significantly. The P-value of the jointly Brown-Forsythe test is 0,429. This implies that the distribution of the four question tones jointly do not differ significantly.

After testing the four tones jointly, the questions are compared individually. For the comparisons of the medians, the Mann-Witney U test is used. For the distribution comparison the Brown-Forsythe test is used.

As illustrated in Table 7, there are no P-values that are significant. This implies that there are no medians and no distributions that differ significantly from each other. This contradicts with H1 (the hypothesis of the question tone).

Table 7: Results for medians and distributions tests for the question tone

	P-values	
Versions	Mann-Witney U	Brown-Forsythe
V1 & V2	0,990	0,806
V1 & V3	0,569	0,451
V1 & V4	0,946	0,742
V2 & V3	0,555	0,320
V2 & V4	0,945	0,924
V3 & V4	0,485	0,294

4.3 Results of question wording

The second factor that is researched to answer the research question is whether the question wording influences the self-reported subjective well-being. It will be tested if choosing different wordings like ‘overall’ instead of ‘all things considered’ influences the results (V1 in comparison with V5). It will also be tested whether time frames influence the results (V1 in comparison with V7). At last, V5, where there is no time frame but the word choice ‘overall’, will be compared with V7 where there is a time frame.

Table 8 reports the P-values of the Mann-Witney U test and the Brown- Forsythe test. As seen, there are no significant P-values. This implies that there are no medians and distributions that differ significantly between the question wording groups. This rejects H2 (the hypothesis of the question wording). Though there are no significant P-values, the comparison between V1 and V7 is close to statistical significance at the 5% level.

Table 8: Results for medians and distributions tests for the question wording

	P-values	
Versions	Mann-Witney U	Brown-Forsythe
V1 & V5	0,449	0,332
V1 & V7	0,072	0,251
V5 & V7	0,285	0,791

4.4 Results of scale wording

The last factor that is researched to answer the research question is whether the scale wording influences the self-reported subjective well-being. Hereby V1 (0 = completely dissatisfied; 10 = completely satisfied), V6 (0 = not at all satisfied; 10 = completely satisfied) and V8 (0 = totally dissatisfied; 10 = totally satisfied) are jointly tested. Afterwards these three versions are tested separately.

The P-value for the Kruskall-Wallis test is 0,358. This implies that the medians of the three scale wordings jointly do not differ significantly. The P-value of the jointly Brown-Forsythe test is 0,928. This implies that the distribution of the three scale wordings jointly do not differ significantly.

Table 9 reports the P-values of the Mann-Witney U test and the Brown- Forsythe test. As seen, there are no significant P-values. This implies that there are no medians and distributions that differ significantly between scale wording groups. This contradicts with H3 (the hypothesis of the scale wording).

Table 9: Results for medians and distributions tests for the scale wording

	P-values	
Versions	Mann-Witney U	Brown-Forsythe
V1 & V6	0,353	0,806
V1 & V8	0,173	0,451
V6 & V8	0,661	0,742

4.5 Correlation analysis

The analysis of the correlates illustrates whether the determinants of life satisfaction depend on the way the question is formulated. To test for this, the correlations are calculated with the Spearman's rank. This is a test that is used to calculate correlations of non-parametric data. The confidence intervals are calculated via bootstrapping. If all the confidence intervals for each determinants overlap, it can be concluded that the correlations do not differ significantly, and the determinants of life satisfactions have no influence on the self-reported life satisfaction.

As seen in Table 10, the confidence intervals for all determinants overlap. As an example, we see that for all the eight versions of 'English first language', the lower and upper bounds are overlapping with each other.

Table 10: Results for correlations and confidence intervals[1]

		V1	V2	V3	V4	V5	V6	V7	V8
English first language	Correlation	-0,006	0,051	0,121	-0,088	-0,092	0,009	-0,140	-0,143
	Lower	-0,239	-0,203	-0,092	-0,287	-0,308	-0,246	-0,356	-0,375
	Upper	0,204	0,309	0,303	0,114	0,139	0,276	0,091	0,086
Children	Correlation	-0,002	-0,107	0,000	0,189	0,166	0,086	0,087	-0,211
	Lower	-0,297	-0,331	-0,185	-0,004	-0,008	-0,249	-0,136	-0,437
	Upper	0,301	0,128	0,175	0,347	0,327	0,384	0,316	0,066
Partner	Correlation	0,102	0,202	0,001	-0,034	-0,080	0,001	0,208	0,150
	Lower	-0,146	-0,044	-0,238	-0,251	-0,301	-0,268	0,004	-0,100
	Upper	0,332	0,435	0,232	0,182	0,137	0,266	0,397	0,384
Gender	Correlation	0,120	-0,064	0,080	0,040	-0,086	0,051	0,026	-0,045
	Lower	-0,124	-0,308	-0,136	-0,197	-0,305	-0,227	-0,220	-0,265
	Upper	0,364	0,199	0,287	0,277	0,133	0,328	0,262	0,179
Health status	Correlation	0,279	0,349	0,276	0,203	0,446	0,065	0,162	0,341
	Lower	0,050	0,123	0,074	-0,043	0,258	-0,224	-0,053	0,121
	Upper	0,485	0,543	0,470	0,423	0,613	0,345	0,365	0,554
Education	Correlation	-0,051	0,185	-0,183	-0,094	0,021	0,037	0,073	-0,087
	Lower	-0,304	-0,054	-0,380	-0,333	-0,182	-0,234	-0,146	-0,304
	Upper	0,207	0,423	0,027	0,127	0,216	0,316	0,302	0,127
Income	Correlation	0,028	0,116	-0,068	0,178	0,216	-0,047	-0,158	0,160
	Lower	-0,191	-0,130	-0,296	-0,056	-0,017	-0,316	-0,362	-0,079
	Upper	0,254	0,362	0,172	0,404	0,409	0,240	0,059	0,383
Employment	Correlation	-0,241	0,083	0,170	0,146	0,051	0,140	-0,178	0,093
	Lower	-0,491	-0,168	-0,045	-0,069	-0,145	-0,140	-0,386	-0,126
	Upper	0,014	0,320	0,393	0,378	0,260	0,409	0,045	0,320
Self employed	Correlation	-0,102	-0,051	-0,186	-0,126	-0,082	0,079	-0,055	
	Lower	-0,389	-0,263	-0,325	-0,277	-0,278	-0,099	-0,288	
	Upper	0,221	0,153	-0,031	-0,004	0,108	0,237	0,213	
Employed/unemployed	Correlation	0,120	0,007	0,074	-0,108	0,003	0,006	-0,127	0,023
	Lower	-0,145	-0,220	-0,184	-0,318	-0,207	-0,300	-0,333	-0,216
	Upper	0,337	0,229	0,316	0,147	0,210	0,344	0,095	0,268
Retired	Correlation	0,187	0,060	0,091	-0,205		0,036		
	Lower	0,156	-0,075	0,039	-0,370		-0,329		
	Upper	0,444	0,197	0,255	-0,203		0,419		
Age	Correlation	-0,193	0,040	-0,121	-0,090	0,110	-0,106	-0,071	-0,080
	Lower	-0,434	-0,213	-0,333	-0,315	-0,114	-0,380	-0,295	-0,317
	Upper	0,049	0,294	0,089	0,120	0,316	0,166	0,170	0,168

[1] Note that not all the correlations are in the middle of the confidence intervals. The reason for this is because the bias-corrected and accelerated bootstrap is used. This corrects the confidence intervals for bias and skewness. This tends to be closer to the true confidence interval (Hesterberg et al., 2003).

5. Conclusion

Measuring life satisfaction is getting more attention among governments, companies and other organizations that measure happiness. The aim of this thesis was to investigate whether different life satisfaction questions influence the self-reported life satisfaction. To answer this, there are eight versions of the life satisfaction questions. These eight questions differ regarding to the question tone, the question wording and the question scaling. To answer the research question, the medians, distributions and correlations are compared.

After the analysis it can be concluded that the medians, distributions and correlations do not differ significantly for the different question tones. So, having different question tones does not significantly influence the self-reported life satisfaction.

Similar results are obtained for question wordings. Using time frames and different words (‘all things considered’ vs ‘overall’) does not influence the self-reported life satisfaction. There are also no significant differences between the medians, distributions and correlates, regarding to the different question wordings. On the other hand, the time framing was almost significant for the median test. Therefore, with a bigger sample size, time framing could maybe influence the self-reported life satisfaction.

Like the question tone and the question wording, there are also no significant differences in the median, mean and correlates when using different scale wordings. This means using that polarity instead of synonyms in answering scales, has no influence on the self-reported life satisfaction.

It can be concluded that according to this research, the way the life satisfaction is formulated has no influence on the self-reported life satisfaction. A possible explanation for this can be the fact that there is a small sample size.

A practical implication is the fact that governments use different life satisfaction questions. Therefore, there are some comments that self-reported life satisfaction of different countries cannot be compared. Furthermore, respondents could possibly not give an honest answer to a question due to item wording bias. Explanations for this are the respondent not understanding the question due to the use of uncommon words, their answer not given as a choice (in multiple choice questions), the structure of the questions et cetera. According to this research, the way the life satisfaction question is formulated does not influence the result. Therefore, self-reported life satisfaction of different countries could be compared.

From this research it can be concluded that framing of the life satisfaction questions has no influence on the self-reported life satisfaction. This means that small changes of a question do not necessarily give different outcomes. The same applies for anchoring: relying on specific elements in a question does not influence the responses. Lastly, item characteristic effects could be present: the use of different time frames (e.g. ‘overall’ and ‘these days’) can influence the results.

This research has some limitations. First of all, it cannot be stated that the research is externally valid. The reason for this is the fact that participants that filled in the survey are mainly people from the Erasmus University. For further research, responses across the world with persons of different backgrounds should be collected. Furthermore, this experiment is only done between subjects. To make a further research more complete, it can also be done within subjects. This makes it easier to detect differences across levels of the independent variables.

References

Andrews, F. M. (279-299). Social indicators of perceived life quality. *Social Indicators Research*, 1974.

Andrews, F.M., & Crandall, R. (1976). The Validity of Measures of Self-Reported Well-Being. *Social Indicators Research*, 1-19.

Beegle, K., and Ravallion, M., & Himelein, K. (2012). Frame-of-Reference Bias in Subjective Welfare. *Journal of Economic Behaviour and Organization*, 556-570.

Brennan, M. (1997). The Effect of Question Tone on Responses to Open-Ended Questions. *Marketing Bulletin*, 66-72.

Chong, D., & Druckman, J.N. (2007). Framing Theory. *Annual Review of Political Science*, 103-126.

Davern, M., & Cummins, R. (2006). Is life dissatisfaction the opposite of life satisfaction? *Australian Journal of Psychology 58(1)*, 1-7.

Deaton, A., & Stone, A.A. (2016). Understanding Context Effects for a Measure of Life Evaluation: How Responses Matter. *Oxford Economic Papers, 68(4)*, 861–870.

Diener, E. (1984). Subjective Well-Being. *Psychological Bulletin, 95(3)*, 542–575.

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

Diener, E., Inglehart, R., & Tay, L. (2013). Theory and Validity of Life Satisfaction Scales. *Social Indicators Research, 112(3)*, 497-527.

Diener, E., Lucas, R.E., & Oishi, S. (2002). Personality, Culture, and Subjective Well-being: Emotional and Cognitive Evaluations of Life. *Annual Review of Psychology*, 403-425.

Diener, E., Suh, E.M., Smith, H., & Shao, L. (1995). National Differences in Reported Subjective Well-Being: Why Do They Occur? *Social Indicators Research, 34(1)*, 7-32.

DOLAN, P., & METCALFE, R. (2012). Measuring subjective wellbeing : Recommendations on measures for use by national governments. *Journal of Social Policy, 41(2)*, 409-428.

Durayappah, A. (2011). The 3P Model: A General Theory of Subjective Well-Being. *Journal of Happiness Studies, 12(4)*, 681-716.

Easterlin, R. (1974). Does Economic Growth Improve the Human Lot? Some Empirical Evidence. *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz. New York/London: Academic Press*, 89-125.

Ekman, P. D. (1990). The Duchenne Smile: Emotional Expression and Brain Physiology II. *Journal of Personality and Social Psychology, 58(2)*, 342-353.

Frey, B.S., & Stutzer, A. (2000). Happiness, Economy and Institutions. *Economic Journal, 110*, 918-938.

Helliwell, J.F, & Barrington-Leigh, C.P. (2010). Measuring and Understanding Subjective Well-Being. *Canadian Journal of Economics, 43(3)*, 729-753.

Hesterberg, T., Monaghan, S., Moore, S., Clipsone, A., & Epstein, R. (2003). *Bootstrap methods and permutation tests.* New York: W. H. Freeman and Company.

Hobfoll, S. (1989). Conservation of Resources. A New Attempt at Conceptualizing Stress. *The American Psychologist, 44(3)*, 513-524.

Hoorn, A.A.J. (2009). *Measurement and Public Policy Uses of Subjective Well-Being.* Nijmegen: Netherlands: Radboud University Nijmegen.

http://www.happinesscouncil.org/. (2018, 10 6). Retrieved from Global Happiness Council (GHC).

Larsen, R.J., Diener, E., & Emmons, R.A. (1985). An Evaluation of Subjective Well-Being Measures. *Social Indicators Research*, 1-18.

Layard, R. (2005). Rethinking public economics: The implications of rivalry and habit. *Economics and happiness, 1(1)*, 147-170.

Lyubomirsky, S., Sheldon, K.M., & Schkade, D. (2005). Review of General Psychology, 9(2). *Pursuing Happiness: The Architecture of Sustainable Change*, 111-131.

Meijman, T.F., & Mulder, G. (1998). Psychological Aspects of Workload. *Handbook of work and organizational psychology (2nd edition)*, 5-33.

Newman, D., & Diener, E. (2014). Leisure and Subjective Well-Being: A Model of Psychological Mechanisms as Mediating Factors. *Journal of Happiness Studies, 15(3)*, 555-578.

OECD. (2013). *OECD Guidelines on Measuring Subjective Well-being.* OECD Publishing.

Ouweneel, P., & Veenhoven, R. (1991). 'Cross-National Differences in Happiness: Cultural Bias or Societal Quality?' In: Bleichrodt, N. and Drenth, P.J. (eds.), Contemporary issues in cross-cultural psychology: selected papers from a regional conference of the International Association for Cro. 168-184.

Podsakoff, P.M., MacKenzie, S.B., Lee, J.Y., & Podsakoff, N.P. (2003). Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. *Journal of Applied Psychology, 88(5*, 879-903.

Rasinski, K.A. (1989). The Effect of Question Wording on Public Support for Government Spending. *Public Opinion Quarterly, 53(3)*, 388-394.

Rehdanz, K., & Maddison, D. (2005). Climate and Happiness. *Ecological Economics*, 111-125.

Rodgers, W.L., Herzog, A.R., & Andrews, F.M. (1988). Interviewing Older Adults: Validity of Self-Reports of Satisfaction. *Psychology and Aging, 3(5)*, 264-272.

Schuman, H., & Presser, S. (1977). Question Wording as an Independent Variable in Survey Analysis. *Sociological Methods & Research*, 151-170.

Scollon, C.N., Diener, E., Oishi, S., & Biswas-Diener, R. (2004). Emotions across Cultures and Methods. *Journal of Cross-Cultural Psychology*, 304-326.

Seligman, M. E., Steen, T. A., Park, N., & Peterson, C. (2005). Positive psychology progress: empirical validation of interventions. *American psychologist, 60(5)*, 410.

Sgroi, D, Proto, E., Oswald, A.J., & Dobson, A. (2010). Priming and the Reliability of Subjective Well-being Measures. *Coventry, United Kingdom: University of Warwick*.

Sirgy, M. (2012). The Psychology of Quality of Life: Hedonic Well-Being, Life Satisfaction, and Eudaimonia (2nd Edition). *Dordrecht; New York: Springer*.

Smith, T. W. (1987). That Which We Would Call Welfare by any Other Name Would Smell Sweeter: An Analysis of the Impact of Question Wording on Response Patterns. *Public Opinion Quarterly, 51(1)*, 75-83.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The Psychology of Survey Response. *Cambridge, England: Cambridge University Press*.

Treasury, H. M. (2010). *Spending review.*

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Journal of Science, 185(4)*, 1124-1131.

Voicu, B. (2015). Priming Effects in Measuring Life Satisfaction. *Social Indicators Research, 124(3)*, 993-1013.

Watson, D., Clark, L.A., & Tellegen, A. (1988). Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *Journal of Personality and Social Psychology, 54(6)*, 1063-1070.