MEASURING SUBJECTIVE WELL-BEING:

Does survey item wording shape life satisfaction reports?

Erasmus University, Rotterdam

ERASMUS SCHOOL OF ECONOMICS

MSc Economics and Business

Behavioural Economics

Master Thesis

Stella Kristin Schön, 467547

Supervisor: Dr. Martijn Hendriks

Second Reader: Dr. Jan Heufer

Date Submitted: 29. August, 2018

ACKNOWLEDGMENTS

I would like to thank my supervisor, Dr. Martijn Hendriks, who shared not only his ideas and expertise with me but also his survey and data. Without his aid, I could not have considered as many different aspects of the complex topic that survey item wording turned out to be. I am especially grateful for the always timely feedback and useful advice, which allowed me to proceed through difficulties and complete a more sophisticated analysis than would have been possible without guidance. My second assessor, Dr. Jan Heufer, I thank for reading and evaluating my work. I also appreciate my family and friends, who supported me in all kinds of ways during the writing of this thesis.

ABSTRACT

Measures of subjective well-being are becoming increasingly popular in the field of economics. Most commonly used are single-item life satisfaction scales (questions akin to "how satisfied are you with your life overall?" that respondents answer on a given response scale, often from 0=extreme dissatisfaction to 10=extreme satisfaction). Despite the potential benefits of such a direct welfare measure, the enthusiasm is hampered by concerns about data quality. A major issue is sensitivity of responses to situational factors, which has been shown in several studies. This thesis investigates survey item wording: the phrasing of the question and response scale. The possibility of wording effects is particularly problematic because they could systematically bias results and limit the comparability of data from surveys with dissimilar items. By means of an experiment, differently phrased items were compared, with treatments representing wording variations found among major surveys. In total, 328 people completed a survey with one of the treatment items, where assignment to conditions was random. Life satisfaction averages, dispersion and correlates were compared between items that were equivalent except for one aspect (question tone, specification of a time period or labelling of the response scale). No significant differences were found regarding life satisfaction average and dispersion between the matched versions. Some differences in correlates were found. Taken together, the results suggest that differently phrased survey items lead to similar reports in terms of the overall distribution but that the analysis of well-being determinants could be biased by wording effects. No conclusive evidence was found regarding mechanisms that might explain the divergence in correlate patterns.

TABLE OF CONTENTS

1. INTRODUCTION

Striving for happiness is a universal aspect of human life. Every person is unique and yet we all share the desire to be happy. This statement sums up a conflict inherent to the study of happiness: on the one hand, happiness is something that every human aspires, which makes it a highly relevant topic to study; on the other hand, interpersonal differences in nature and nurture render happiness a highly subjective experience, which complicates scientific investigation. The scientific term *subjective* well-being (SWB) captures this characteristic. SWB encompasses three broad concepts of well-being: cognitive evaluations, emotional experience, and eudaimonia – a sense of life's worthwhileness (Helliwell, Layard & Sachs, 2013). A majority of SWB research is based on life evaluations, particularly life satisfaction self-reports. These are most commonly elicited with a single item on a survey. The exact question wording and response scale format varies from survey to survey, but the typical form is similar to the item on the influential World Values Survey. There, respondents encounter the question: "all things considered, how satisfied are you with your life as a whole these days?" and answer on a scale from "completely dissatisfied" (1) to "completely satisfied" (10).

Subjective well-being measures are established in the domain of psychology but until recently have been rarely used by economists. Economists care about well-being but have been hesitant to rely on self-report measures. Traditional welfare metrics in economics include income on the individual level and GDP on the societal level. Even alternative indicators with a broader definition of welfare, such as the Human Development Index, are based on objective measures like life expectancy or education. Economists have not ignored that the same things can have different values to different people. The concept of utility recognizes that people vary in the amount of satisfaction they derive from consuming certain goods. Still, neoclassical economists seemed to have a solution to the measurement problem: if people make choices to maximize their well-being, purchasing behaviour is informative about their preferences. In other words, there seemed to be an objective way to infer the subjective value of an option. However, findings from economics and psychology, especially the emerging behavioural economics field, show that people do not always make decisions in a way that maximizes their well-being (Kahneman & Kruger, 2008).

In fact, a use of subjective well-being measures in economics is to test competing theories of behaviour, for example, whether high levels cigarette consumption reflect deliberate decision-making or self-control problems (Gruber & Mullainathan, 2005). Happiness research also suggests that the relationship between income and welfare is less straightforward than assumed in standard economics. There is, for example, no consistent link between GDP growth and average national happiness, which may be explained by the findings that people care about relative income and adapt to changes in income over time (Di Tella & MacCulloch, 2006). There are several possible applications of SWB measures in the public policy domain: to track societal welfare; to value public goods and "bads" (e.g. pollution); and to evaluate fundamental aspects of political systems, like democratic rights (Stutzer & Frey, 2010). Research using subjective well-being measures promises to improve economic theory and, by application, the lives of many people. Despite the potential benefits of using SWB measures and the increased interest in doing so, many economists still reject the use of self-reports that reflect subjective

assessments. Many have expressed doubts about the validity of SWB measures, first and foremost the commonly used life satisfaction survey questions.

Economists may be particularly critical but most of the best-known studies cited to illustrate such claims were produced by psychologists, notably by Norbert Schwarz and colleagues. According to Schwarz and Strack (1999), people have no stable sense of life satisfaction and only form a judgment when prompted. The spontaneous evaluations are shaped by the information most available to the respondent, which generally includes life aspects they think about frequently and also momentarily available information (like current mood). As a result, global happiness reports are susceptible to the influence of irrelevant situational factors, so Schwarz and Strack. Using experiments, Schwarz and Clore (1983) found that higher SWB is reported on sunny days than on rainy days, but making respondents aware of the weather (by asking about it) reduced this effect. Seeing the national soccer team win before completing a survey lead to higher reported life satisfaction, as did filling out the survey in a pleasant versus an unpleasant room (Schwarz, Strack, Kommer & Wagner, 1987). Such findings are attention-grabbing, but SWB measures should not be completely rejected on the basis of single studies. Comprehensive and balanced reviews suggest that these scales capture meaningful differences in well-being even though situational factors play a role (Diener, Inglehart & Tay, 2013; OECD, 2013). Reports of subjective well-being are thus informative to a degree and deserve further attention. Our understanding of the measures is still incomplete and more work needs to be done to improve the quality and comparability of SWB data.

There is not yet a standardized way of asking people about their satisfaction with life. Several aspects of the question and response scale vary from survey to survey. Item format effects (distortions in life satisfaction reports due to the specific form of the question and response scale) are a concern. How framing shapes responding has been a topic in the general methodology, psychology and behavioural economics literatures – but little systematic evidence for such effects exists pertaining to life satisfaction questions in particular. This study will contribute to closing this gap in the literature on measuring SWB by adding experimental evidence on item wording, which has been identified as an area for future research in several reviews (Diener et. al, 2013; OECD, 2013). It is important, for reliability, validity and comparability reasons, to use questions that are understood in the same way across respondents and capture the same type of information. Standardization of wording appears as an obvious step to take in maximizing the comparability of SWB data across groups and time periods. However, since many versions are already in use, making changes represents a trade-off between improved comparability across studies and decreased comparability of future with past data. So far, it is largely unclear how significant the impact of particular dissimilarities in wording is, and which framing might be the most useful. This leads to the following main research question:

"Does survey item wording influence life satisfaction reports?"

Table 1 shows a selection of life satisfaction items from major surveys. The examples illustrate some of the ways in which wording can vary across surveys. For example, the question tone can be positive ("satisfied") or more balanced ("dissatisfied or satisfied"). A reference period, a certain time frame respondents are supposed to consider, may be mentioned ("these days") or

not. The verbal labels of the scale end-points, the scale anchors, might be labelled differently (e.g. "completely" versus "totally"). Other types of variation in wording exist, but to consider all is beyond the scope of this article. The following sub-questions emerge:

- Does question tone affect the average, dispersion and correlates of life satisfaction?
- Does reference period use affect the average, dispersion and correlates of life satisfaction?
- Does scale anchor labelling affect the average, dispersion and correlates of life satisfaction?

Table 1 – Life satisfaction items from major surveys

Survey	**Question**	**Response Scale**
World Values Survey (WVS)	All things considered, how **satisfied** are you with your life as a whole **these days**?	**Completely** dissatisfied (1) – **Completely** satisfied (10)
British Household Panel (BHP)	How **dissatisfied or satisfied** are you with your life overall?	Not at all satisfied (1) – Completely satisfied (7)
German Socio – Economic Panel (SOEP)	How satisfied are you with your life, all things considered?	Completely dissatisfied (0) – Completely satisfied (10)
Household, Income and Labour Dynamics in Australia Survey (HILDA)	How satisfied are you with your life, all things considered?	**Totally** dissatisfied (0) – **Totally** satisfied (10)

Note: Bold type has been added for emphasis of between-scale differences in wording

The remainder of this paper is structured as follows: chapter two expands on the relevant literature, chapter three introduces the methodology and data, chapter four contains the main analysis, and chapter five ends with the discussion and conclusion.

2. THEORETICAL FRAMEWORK

2.1. Subjective well-being: concepts & measures

The terms "subjective well-being" (SWB) and "happiness" are often used interchangeably in the literature. There is some agreement that subjective well-being includes, but is not limited to happiness. Both terms cover multiple concepts of well-being. According to the 2013 World Happiness Report, the term "happiness" has at least two connotations: if people say they are happy, it might mean that they are satisfied with how their lives are going, or it might mean that

they are experiencing a positive emotional state (Helliwell et. al, 2013). SWB can be defined in broader terms, to include "eudaimonia" (a sense of purpose or life's worthwhileness) in addition to evaluative and emotional well-being (National Research Council, 2013). Accordingly, there are several types of subjective well-being measures: life evaluations, emotional reports and worthwhileness accounts.

Most SWB research, in general and in economics, is based on life evaluation measures. One advantage is that respondents can choose to weigh in any factors they find relevant (Tay, Chan & Diener, 2014). A more pragmatic advantage is the ease of administration – especially compared to some affective measures (e.g. Kahneman, Krueger, Schwarz & Stone, 2004; Csikszentmihalyi & Larson, 2014). Life evaluations are generally elicited via surveys, often using only one question. The three main types of single items are the Cantril ladder (respondents position themselves on a ladder from worst to best possible life), happiness with life as a whole, and life satisfaction (Helliwell et al., 2013). Questions about life satisfaction are the most commonly used measure and therefore the focus of this article.

2. 2. How do people answer life satisfaction questions?

People do not usually think about how satisfied they are with life on a scale from zero to ten (for example), until they encounter such a question on a survey. According to Schwarz and Strack (1999), life satisfaction reports are highly sensitive to context because people have no stable sense of overall satisfaction that can be readily communicated. Rather, they evaluate life in the moment they are asked, and situational factors enter the judgment. Multiple models of the general survey response process exist, but there is wide agreement on the basic mental tasks respondents need to complete: understand the question, recall relevant information, determine the answer, fit it to the given response format and, possibly, edit it to be more appropriate in some way before reporting (Schwarz, 2007; Podsakoff et al., 2003). Sensitivity to context is an issue throughout all stages (Schwarz, 2007). People will generally give accurate responses if they have the ability and motivation to do so, which may be influenced by characteristics of the construct of interest and the measurement method (OECD, 2013).

An important construct factor is complexity. Life satisfaction is a rather broad concept, making it challenging to respond to questions about it: individuals are asked to consider everything that may affect how happy they are with their lives and express their conclusions by picking a number from the response scale. People might have difficulties understanding what exactly is asked of them, they may fail to remember certain events, and they may only be willing to expend a limited amount of time and energy on responding (OECD, 2013). In fact, it would take immense cognitive effort to really identify each relevant factor, recall all pertaining information and integrate the many pieces into one overall numerical judgment to report. Rather than expending that effort, respondents tend to "satisfice" – conserve mental resources by only attempting to give an adequate answer rather than the best possible one (Krosnick, 1991). Satisficing generally increases with task difficulty, mediated by individual differences in underlying ability and motivation (Krosnick, 1999). Thus, not all individuals satisfice equally, but complexity increases the level of ability and motivation that is required on the part of respondents to give the best possible answer. As a result, response accuracy is generally higher for easier tasks than for more complex ones.

According to Schwarz and Strack (1999), people only consider the most accessible information when making life evaluations. (Memory failures make some information inaccessible, while satisficing explains a lack of effort in searching for more information.) Schwarz and Strack distinguish between chronically and temporarily accessible information: the former tends to reflect aspects of life that consistently impact the person's happiness with how their life is going. Therefore, it is mostly a source of valid variation. Information that is available only in the moment, on the other hand, leads respondents to give different answers to live satisfaction questions across measurements, even if the person's life circumstances and experiences have not changed. An example for chronically accessible information would be financial worries in people of low socio-economic status. For these individuals, such worries are likely a daily experience, for a prolonged period of time or even their whole lives. An example for temporarily accessible information would be if an otherwise happily married woman experiences frustration with her husband in the moment of taking the survey because she just discovered he forgot to take out the trash.

Answers to life satisfaction questions do not only depend on which information is available but also on how that information is used. Firstly, people will give little weight to information they recognize is irrelevant, even when it is available. In support, Schimmack and Oishi (2005) show that prompting respondents to think about their satisfaction with traffic or the weather does not lead them to weigh these unimportant factors more heavily in a subsequent life satisfaction evaluation. Secondly, judgments of life satisfaction are made by comparing one's reality to some reference point, for example one's past, one's ideal life or the societal standard (Diener et al, 2013). Salience of specific standards can influence which one is used: for example, people that filled out a survey with a handicapped confederate in plain sight reported higher satisfaction than a control group (Strack, Schwarz, Chassein, Kern & Wagner, 1990, as cited in Schwarz & Strack, 1999). In addition to being a reference point, social norms are sometimes used by respondents to determine whether their honest judgment is acceptable to report: evidence suggests that the less anonymous the measurement setting, the more respondents edit their responses to match socially approved levels of well-being (Schwarz, Strack, Hippler & Bishop, 1991). The desire to conform with society can thus compete with the motivation to give an accurate response.

Three conclusions can be drawn for the risk of context effects. Firstly, life evaluations are likely susceptible to these effects due to the nature of the construct and respondents, such as the inherent complexity of making a judgement about life as a whole and the tendency of people to conserve mental energy. Secondly, measurement conditions may influence total risk of bias by affecting respondent ability and motivation to give a good answer (for example, by increasing task difficulty or perceived pressure to give certain responses). Thirdly, situational factors can shape which information is available and how it is used (for example, by making a certain reference point salient).

2.3. The validity and reliability of life satisfaction measures

How problematic are context effects really? And what portion of the variation in reported life satisfaction actually reflects changes in well-being? Two streams of research assess these types of questions: studies investigating how changes in measurement conditions affect results, and

studies focussing on the stability of scores across measurements and the variation of scores with more persistent life circumstances (Diener et al., 2013).

According to the *OECD Guidelines on the Measurement of Subjective Well-Being* (2013), there is a large literature about how reports of subjective well-being are affected by the conditions of measurement, possibly because the use of such measures is so contested. Chapter two of the OECD report provides an overview of the literature concerning any effects the measurement conditions have on SWB reporting. Factors from general context factors to details of survey design have been found to affect subjective well-being reports, for example time of measurement, survey mode, and number of response options.

Question-order effects are one of the most often considered issues in this literature. One of the best-known studies regarding question order is Strack, Martin and Schwarz (1988). They found that dating frequency was uncorrelated to overall life satisfaction if the dating question preceded the life satisfaction item, but that the correlation was significant when the question order was reversed. Schwarz and Strack (1999) explain that question order effects occur when information that is not already chronically available is made salient by a preceding question and is regarded as relevant enough to enter judgment. This is an example of a priming effect, which is when some cue in the situation makes certain information or emotions more available and thereby alters the response pattern (Janiszewsky & Wyer, 2014). Question order is also thought to influence how respondents interpret what the question is asking for. Schwarz and Strack (1999) hold, that conversational norms influence how an item is understood. For example, information about one's romantic life, if elicited by a previous question, may not be used in a following global satisfaction evaluation. Due to a conversational norm of non-redundancy, respondents might think they are asked to provide new information only (i.e. they understand the question as asking: how satisfied with life are you, *other than your romantic life*?).

A second, frequently discussed issue is the role of current mood. As previously stated, people generally disregard information that they perceive to be irrelevant. How, then, do factors like finding a dime, watching the national team win or having good weather, influence life satisfaction judgments? The researchers behind these findings, Schwarz and colleagues, have suggested that mood is the link. One explanation is that answering life satisfaction questions based on current mood is a sort of mental short-cut. The rule-of-thumb can lead to satisfactory answers with little effort, as people who feel happy emotions often tend to be satisfied with life (Sandvik, Diener & Seidlitz, 1993; Robinson, 2000). But using this heuristic also introduces error, because mood during survey-taking is not necessarily representative of how respondents generally feel about life. A priming effect of current mood, leading to the recall of memories that match in terms of emotional valence, has also been suggested (Bower, 1981). This would cause bias by making some types of relevant memories more available than others.

Considerable research has been done to show context effects, but there is also much evidence in support of the reliability and validity of life satisfaction scales. The latter has recently been reviewed by Diener, Inglehart and Tay (2013). In support of reliability, the authors present evidence of similar satisfaction scores across measurements at different points in time. They note that single-item scales show less stability than multiple-item ones and that scores become more dissimilar with longer time between measurements, as is expected given that life conditions change over time. As an indicator, Fujita and Diener (2005) found that

retest-correlations for a single item declined from .56 after one year to .24 after 16 years. This is broadly representative of the extant literature: using meta-analysis, Schimmack and Oishi (2005) estimate only slightly lower retest-correlations over short horizons (around .5), with estimations further approximating the Fujita and Diener findings for longer horizons. Stability across measurements indicates that respondents construct life evaluations in a consistent manner.

Even a perfectly reliable measure is essentially useless, if it does not capture the intended construct. Diener et al. (2013) present evidence for validity from various sources: life satisfaction scores have been found predictive of health, career and relationship outcomes (e.g. suicide and separation); self-rated satisfaction correlates with related measures (e.g. reports by others and physiological variables) and with life circumstances (e.g. political freedom on the national level and marriage on the individual level). A notable study in this vein is Oswald and Wu (2010), who compared an objective and a subjective account of well-being in the USA. They found that an objective quality of life indicator and a single-item life satisfaction scale give matching results regarding well-being on the state level.

How do the two streams of literature – findings of context effects with evidence for reliability and validity – fit together? The presented studies suggest as much as: various situational factors can influence results, but reliable and valid measurement is not completely precluded by context sensitivity. Stutzer and Frey (2010) suggest using a measurement error framework to judge the importance of findings regarding circumstantial influences. Measurement error refers to variance in the data that does not reflect the intended construct. A general distinction is made between random error, which affects data points arbitrarily and systematic error, which affects groups of observations with some regularity. According to Stutzer and Frey (2010), random error is merely "white noise" and causes no major problems besides reducing statistical fit; systematic errors, on the hand, are a serious threat: they distort results and can lead researchers to draw faulty conclusions.

Lucas and Donnellan (2007) used panel data from two different surveys with single-item life satisfaction scales to investigate what influences the stability of scores. They found that 34-36% of variance is due to unchanging traits (e.g. genetic make-up), 29-34% is due to a moderately stable autoregressive component (e.g. longer-term but changeable life circumstances) and the remainder is due to the combined influence of situational factors and random measurement error. For single-item scales, it is impossible to separate the situation-specific systematic component from the random one (Schimmack & Oishi, 2005). Testing a multiple-item indicator, Eid and Diener (2004) estimated that 74% of variance in life satisfaction is due to chronically accessible information, 16% to temporary sources, and 10% to random measurement error. Based on a combined meta-analysis of studies using multiple and single items, Schimmack and Oishi (2005) estimated the proportions to be 80%, 10% and 10%, respectively.

Schimmack and Oishi note that they might have underestimated the impact of temporarily accessible information on judgements, since the analysis was based on studies using repeated testing of identical questionnaires. Thus, factors like question order or wording could have influenced scores on all measurement occasions in the same way and would not have been accounted for. This is problematic, given that method effects (variance that is attributable to the way in which the measure was obtained) are a major source of systematic

error (P. Podsakoff, MacKenzie, Lee & N. Podsakoff, 2003). In response to this concern, Schimmack and Oishi (2005) assessed how much of an impact question order may have. A meta-analysis indicates that order effects are small but significant. Taking into account that there were few underlying studies with varying findings, the authors conclude that temporarily accessible information could have a strong impact in some situations but not in others. Further, Schimmack and Oishi conducted five experiments and found no notable order effects, regardless of whether an irrelevant or relevant domain satisfaction question preceded the life satisfaction scale. These findings may seem to suggest that information made accessible by the questionnaire has only a small impact on responses. However, item wording was identical within each study and could have had an impact which, given that design, was not distinguishable.

2.4. The influence of item wording on life satisfaction scores

Item wording refers to phrasing of the question and labelling of the response scale. Item format is considered to be an aspect of methodology and can systematically affect results (Podsakoff et al., 2003; Schwarz & Strack, 1999). Considering the Stutzer and Frey (2010) measurement error framework, wording effects should thus be seen as a serious threat, that can lead unaware researchers to draw faulty conclusions. Recall that, in general, the conditions of measurement can increase the risk of error by undermining respondent ability and motivation to give good answers. Situational factors can also influence which information is available and how it is used. As per Schwarz (2007), sensitivity to context is an issue throughout all stages of the response process, and method effects are often significant.

Podsakoff et al. (2003) identify which types of method bias pose the greatest risk at each stage, some of which are relevant in the context of item wording and some of which relate to the elements of the Schwarz and Strack (1999) judgment model. In the comprehension stage, difficult or unclear wording complicates the task of understanding what a question is asking for (and wording may influence which interpretation respondents arrive at). Complexity might also influence carefulness of consideration and response time. In the retrieval stage, respondents recall memories, where cues in the item could influence the retrieval (what information is searched for and the ease of recalling certain types of information). Next, respondents must integrate the information into an overall judgement and item phrasing may influence how information is used (what appears most relevant and what the standard of comparison should be). Fourth, respondents must communicate their answers using the given response format, where labelling of the response scale influences the interpretation of options. Finally, respondents might engage in editing. Most relevant, they might communicate a response that is appears more appropriate even if untrue, where phrasing might be suggestive of what is expected. Respondents do not deliberately follow these steps in answering a question. Rather, the whole process tends to happen quickly and largely automatically and several tasks may be worked on simultaneously (Podsakoff et al., 2003). As a result, wording effects may not neatly fit into one of the categories suggested above.

2.4.1. Priming, framing, anchoring

In standard economic theory, the way in which information is presented is not expected to have an effect. However, this assumption is inconsistent with findings in the behavioural economics and related psychology literature. Three concepts appear particularly relevant to the study of wording effects: priming (which has been briefly mentioned), framing and anchoring.

Priming effects are said to occur when the presence of a particular stimulus affects subsequent task performance, decision-making, judgments or behaviour (Janiszewsky & Wyer, 2014). Associative memory is understood to be the brain basis of such effects: knowledge, goals, and emotions are contained in this memory network and activation spreads from one point to others along the strongest connections (Morewedge & Kahneman, 2010). In other words, priming with one thing causes related things to become more accessible. Many types of priming effects have been found in various domains (Janiszewsky & Wyer, 2014). A famous example is Bargh, Chen and Burrows (1996), who primed participants with words related to elderly stereotypes (old, grey) and found that they subsequently walked at a lower speed than a control group (which was primed with neutral words). Priming also works via metaphorical associations, as evidenced by Williams and Bargh (2008), who found that holding a hot drink led participants to rate a stranger as more warm. Morewedge and Kahneman (2010) claim that judgment biases, in general, are due to overweighing strongly activated information and neglecting other factors. The spreading activation along interconnected points underlies both anchoring and framing effects, so the authors.

Framing was introduced to economics by Tversky and Kahneman (1981). They found that respondents preferred a risky over a certain option when a choice problem was framed in terms of losses, but that preferences were reversed when the problem was framed in terms of gains. Significant and robust framing effects have been found in a variety of domains and are not generally remedied by more careful consideration of the given problem (LeBoeuf & Shafir, 2003). An explanation is that frames evoke certain emotions, the influence of which can only be avoided if the respondents are able to reframe the problem for themselves (Morewedge and Kahneman, 2010). According to Levin, Schneider and Gaeth (1998), there are three distinct types or framing. Most relevant in the context of life evaluations is attribute framing, where a target is judged more or less favourably depending on whether it is presented in a positive or a negative frame. The resulting distortion is usually valence-consistent, meaning a positive frame leads to more positive evaluations and vice versa. For example, Levin, Johnson, Russo and Deldin (1985) found that evaluations of student performance were higher when exams scores were expressed as % correct compared to % incorrect answers. Levin, Schneider and Gaeth (1998) conclude that attribute frames work like primes, leading the subjects to access valence-consistent knowledge. They propose that the effect is so reliable because the prime is part of the target description rather than an unrelated stimulus.

Anchoring can also be traced back to Tversky and Kahneman (1974). In a now famous study, they showed that people's estimate of the percentage of African countries in the UN was biased towards a randomly generated number. Anchoring is a robust phenomenon: effects – judgments being biased towards an initially presented value – have been demonstrated in a variety of situations and in subjects with different levels of motivation and ability (Furnham &

Boo, 2011). Multiple explanatory mechanisms have been suggested, but the currently predominant view is that people consider whether the anchor provides a good answer, whereby mostly confirmatory evidence comes to mind (Furnham & Boo, 2011). This mechanism is also referred to as "confirmatory hypothesis testing" (Mussweiler and Strack, 1999). Chapman and Johnson (1999) show that anchoring is reduced when participants are prompted to think of contradictory information. If asked to think of supporting information, the results are equal to when no prompt to think of anything in particular is given, evidencing that anchor-consistent information automatically comes to mind. Mussweiler and Strack (1999) conclude that anchoring is mediated by a process like priming. They suggest that anchoring effects is a more robust phenomenon than typical priming effects, because the accessible information is perceived as relevant. Therefore, the influence of the anchor is not being corrected for.

The most commonly used life satisfaction measures differ in wording in three respects: question tone, reference period, and scale anchors. The following sections will discuss the available evidence and relevant theoretical considerations, including the role of priming, framing and anchoring, for each of these.

2.4.2. Question tone

Question tone, here, refers to how the construct of interest appears in the question in terms of valence. As mentioned, the tone can be positive (how *satisfied* are you?) or more balanced (how *dissatisfied or satisfied* are you?*)*. Tone could also be negative (how *dissatisfied* are you?) or neutral (*how do you feel* about your life?). The priming, framing and anchoring literature suggests that positive tone could lead to higher, and negative tone to lower life satisfaction reports, compared to balanced or neutral formulations.

Although anchoring studies typically use numerical anchors, the "confirmatory hypothesis-testing" mechanism could also apply if the anchor is a word. If that is the case, respondents would test whether the answer implied by the question wording is a good answer, and mostly confirmatory evidence would come to mind. For example, if the question asks how "satisfied" the respondent is, they would test whether "I am satisfied" is the correct answer to give. In doing so, they would think mostly about life events and circumstances that are consistent with that conclusion. This is also true for negative wording, except that respondents would then test whether "I am dissatisfied" is a fitting answer, and relatively more information consistent with being dissatisfied with life would come to mind. The effect of balanced phrasing is somewhat more difficult to predict. Being "dissatisfied or satisfied" might not be registered as a useful response to give at all. Alternatively, such wording might suggest testing the middle of the scale as the possible response (which might be understood as meaning something like being both dissatisfied and satisfied). Positive and negative experiences would likely both come to mind equally, when considering such a hypothesis. In contrast to the other three tones, neutral phrasing provides no anchor at all. Thus, such wording is not expected to systematically bias results in any direction.

Priming explains how exposure to certain words influences recall and subsequent judgments. For example, using the word "satisfied" can lead to an upward bias in life satisfaction reports by making is easier to recall positive than negative things. During question comprehension, the mental representation of the term is activated in associative memory;

activation then spreads along the most closely connected points, which would be emotions, knowledge, goals and other points related to being satisfied in some way. The strongly activated information would be overweighed, leading to an upward bias in judgment. Negative wording would ease the recall of negative experiences and lead to downward biased life satisfaction reports. Balanced wording would make both positive and negative information accessible, while neutral wording would make neither one more available. Framing, similarly, suggests that approaching the question from a positive angle would lead to a valence-consistent shift in responses because frame-consistent knowledge is accessed. Again, the reverse would be true for a negative frame. Balanced and neutral framing, on the other hand, do not push respondents to consider just one side. Neutral framing, at the least, does not hinder respondents from considering both positive and negative factors. Balanced framing could even be seen as actively suggesting to take both sides into account.

Based on priming, framing and anchoring, potential differences in average life satisfaction scores between different tones would be due to the consideration of different types of information (in terms of valence). Using the same logic, the correlates of life satisfaction should also vary between the tones. If relatively more positive material comes to mind and enters the judgment when tone is positive, the global satisfaction scores should strongly correlate with areas of life that the respondent is happy with. Again, the opposite is expected for negative tone. If balanced tone brings positive and negative experiences to mind to an equal degree, the correlates of life satisfaction should include areas of life the respondent is happy with as much as the aspect he or she dislikes. Finally, neutral tone is not expected to make either positive or negative factors relatively more salient. Thus, life satisfaction likely correlates most strongly with whichever aspects are most available and important to the respondent, regardless of valence. It is not straightforward, to make predictions about the exact effects of tone on life satisfaction correlates. In theory, individual differences could completely cancel each other out. Even if respondents consider more positive experiences when tone is positive, one person might be happy with their finances and unhappy with their health, while the opposite may be true for the next person. Then, no notable effect would be found in the pattern of correlates on the group level. Yet, it seems likely that the individual effects do not completely neutralize, and some differences in correlate patterns depending on tone will be observable.

Although the relationship between question tone and life satisfaction correlates has not been previously studied, some limited evidence exists regarding the effect of positive and negative tone on life satisfaction score means and standard deviations. In an experimental study, Davern and Cummins (2006) compared questions framed in terms of satisfaction with questions asking about dissatisfaction. When using a response scale that only covers dissatisfaction, the responses to the dissatisfaction questions were opposite to responses to satisfaction question, i.e. around 30% of the scale maximum for dissatisfaction compared to around 70% of the scale maximum for satisfaction. This scale type is call *unipolar* because the end points mark minimum and maximum of the *same* thing. On the other hand, the positive tone led to higher mean ratings than the negative tone, when both were measured on the typical "completely dissatisfied" to "completely satisfied" scale. This scale type is called *bipolar* because the end points are *opposites*. The authors' interpretation was that respondents are biased towards expressing satisfaction due to an inherent positivity bias, even when they are asked about their level of dissatisfaction.

An alternative explanation for the findings is that respondents interpret the dissatisfaction question differently, depending on the scale. It seems likely that respondents understand the question as asking strictly about dissatisfaction if the scale covers only dissatisfaction, but not if it also covers satisfaction. Rather, it appears that with the bipolar scale, most respondents understand the satisfaction and dissatisfaction questions as asking about the same thing: how they feel about life in the range of extreme dissatisfaction to extreme satisfaction. From this perspective, the finding of lower mean scores in response to the dissatisfaction question (compared to the satisfaction question) indicates that positive and negative wording introduce a valence-consistent shift in responses. This finding of Davern and Cummins is therefore consistent with predictions based on the priming, framing and anchoring literature.

Miscommunication alone could also explain the lower mean in response to bipolar dissatisfaction item, compared to the bipolar satisfaction item. It appears that the authors expected respondents to interpret the negatively framed question to ask strictly about dissatisfaction, regardless of the response scale. As reasoned above, it is possible that respondents have interpreted the bipolar dissatisfaction item as asking about the same thing as the bipolar satisfaction item. If the majority of respondents understood the bipolar dissatisfaction item as asking about their life evaluations from extreme dissatisfaction to extreme satisfaction, while a minority of respondents understood the question as asking about dissatisfaction only, this would also result in a somewhat lower mean score in response to bipolar dissatisfaction compared to satisfaction.

If it is true that some people understand a question one way and some another, it seems likely that many people are unsure about what the correct interpretation is. As previously mentioned, comprehension difficulty promotes error. In line with this reasoning, the survey methodology literature suggests that respondents have difficulties with processing negatively worded questions and using them can increase error (Lietz, 2010; van Sonderen, Sanderman & Coyne, 2013). Using negative wording is likely to result in a more dispersed distribution of scores. This is consistent with both miscommunication – some people interpreting the question one way and some another – as well as with difficulties in making sense of the question and using one of many response strategies, including random responding. In line with this, Davern and Cummins (2006) found that the standard deviation of life satisfaction scores was higher for the negative bipolar item compared to the positive bipolar item. Taken together, the considerations regarding possible question tone effects lead to the following hypotheses:

H1: Question tone has an effect on life satisfaction average, dispersion, and correlates.

H1a: Compared to neutral and balanced question tone, average life satisfaction reports are higher when tone is positive and lower when tone is negative.

H1b: Compared to positive, neutral and balanced question tone, life satisfaction reports are more dispersed when tone is negative.

2.4.3. Reference period

The reference period is the time frame that respondents are asked to consider. Life satisfaction questions may specify no bound at all, or they may include a very vague term for present times (like "these days" or "nowadays"). Including such a term suggests that the question is about recent times rather than the whole life, and the recent past is expected to have the strongest impact on the evaluation (Diener et al., 2013). If a reference period is included, the exact term influences which time periods people think the question asks about. Some, but little, evidence exists in regards to the effect of reference period choice. Ralph, Palmer and Jayne (2011) tested the terms "nowadays" and "these days". They found that the latter is perceived to cover a shorter time frame than the former. The respondents, taken together, considered time periods from six years in the past – often thinking back to a significant event – to a few months in the future. It was not investigated which time frames people consider when no reference period is mentioned.

In terms of the response process model, it appears that choice of reference period influences how people understand the question. Specific terms likely act as a recall cue that leads respondents to search for memories matching the suggested time frame. Moreover, in the judgment stage, information that does not seems to match (too long ago, too far in the future) would be left out of the evaluation. Alternatively, information about the past might serve as standard to which recent times can be compared. Diener et al. (2013) suggest that questions with a reference period produce less stable scores because they reflect a shorter time period than the whole life. They cite Krueger and Schkade (2008), who administered a question using "these days" twice, 14-days apart, and found a correlation of .59 between scores. However, this does not seem particularly low in light of the meta-analysis by Schimmack and Oishi (2005), which suggests re-test correlations of around .50 over a similar period, based on questions with varying reference periods. This cannot be tested in the current study due to the design, which is explained in chapter 3.

Correlates pattern could differ depending on wording, if people really consider much shorter periods of their lives when a term like "these days" is used, compared when it is not. If only the recent past is considered, the current situation and daily experiences would likely receive a higher weight compared to when life as a whole (or at least a significant part of it) is considered. Then, "big picture" factors and the overall direction life is developing in might be of more importance. It is however difficult to predict, which factors are the most important to different people in the short-term and long-term and whether individual preferences add up or cancel out. Evidence on this topic, as well as on the effect of using a reference period on life satisfaction averages is lacking.

The dispersion of life satisfaction scores might be affected by reference period phrasing due to task difficulty: a too demanding (too long) reference period makes misremembering more likely, increases sensitivity to context, and leads respondents to rely on simple response rules or give otherwise biased reports (OECD, 2013). This would suggest more dispersion in life satisfaction scores when no reference period is mentioned, compared to when one is specified. While there is evidence that a short reference period is critical for emotional reports (e.g. Thomas & Diener, 1990), no evidence is available showing exactly how choice of reference period affects life satisfaction reports. Given that life satisfaction items of any type prompt respondents to consider a wide range of domains and a significantly longer period than

is typical for emotional reports, it is unclear whether the lack of a vague reference period really further increases task difficulty. Due to a lack of evidence and difficulties in making predictions about the effect of including a reference period on life satisfaction, the following hypothesis stated in the null form:

H2: Including a reference period has no effect on life satisfaction average, distribution, and correlates.

2.4.4. Scale anchors

Most of the large surveys with single life satisfaction questions provide a numerical response scale, where only the end points are verbally labelled. However, there is some variation in the wording of these labels. Evidence about possible wording effects of this type is scarce. One exception is Davern and Cummins (2006), whose experimental study included differently labelled response scales. They found that a positively worded life satisfaction item produced similar scores in terms of mean and standard deviation, regardless of whether the low anchor was "completely dissatisfied" or "not at all satisfied" (with the high anchor being "completely satisfied" in both cases.) While the authors take this as a sign of good comprehension and accurate responding, that might not be an appropriate conclusion. The similarity of satisfaction scores between scales could in fact suggest that respondents do not pay attention to scale labels, because the same satisfaction level should technically translate to a different score on these two scale types. For example, the middle of the bipolar scale could indicate being neither satisfied nor dissatisfied, while the absence of satisfaction should be a zero ("not at all satisfied") on the unipolar scale.

The present study will not be concerned with unipolar versus bipolar scales, as the majority of life satisfaction survey items employs bipolar scales. Instead, the focus is on differences in the exact wording of scale anchors in bipolar scales. An example among the well-known surveys (see Table 1) is between the German Socio-Economics Panel and the Household, Income and Labour Dynamics in Australia surveys. The former uses a response scale with "*completely* dissatisfied" and "*completely* satisfied" as the scale anchors, while the latter uses "*totally*" (dis)satisfied. Since "totally" and "completely" essentially mean the same, it seems likely that respondents interpret the response scales in the same way if either word is used. The comprehension of the question, as well, should be equal between the two versions. It is not expected that such labels – differently worded but in essence equivalent – should produce significantly different scores in any way. This is reflected in the following hypothesis:

H3: Scale anchor labelling has no effect on life satisfaction average, distribution, and correlates.

3. DATA & METHODOLOGY

3.1. Subjects

Data was collected with a self-administered online survey that could be taken on a mobile phone or computer. Subjects were told that the survey takes about 10 minutes and that responses were anonymous and treated confidentially. The high level of anonymity minimizes social desirability concerns. Each respondent took the survey only once (which means analysis of life satisfaction stability over time is not possible). The final sample contains 328 people that completed one of the survey versions relevant to this study (the survey contained some questions for use in other research).[1] The group of respondents is best described as a convenience sample, and the full sample is comprised of three sub-groups. One part of the respondents were participants in a Massive Open Online Course (MOOC) on scientific literacy. The MOOC is called "Deception Detox" and was developed by Erasmus University. Secondly, the survey was administered to students participating in the university's minor "Quality of Life and Happiness Economics" and attendees of guest lectures that Erasmus University Professor Dr. Martijn Hendriks gave. Lastly, a third group of participants were recruited via social media: the survey was accessible in survey exchange groups with students from other institutions, in a Facebook group with students from my degree and to my personal contacts. Other than good English skills, there was no a priori limitation as to who could participate. The survey distributed to each group was identical in terms of content and design, except that the social media group received an additional question about time period consideration.

3.2. Design

This study follows the experimental tradition of studying context effects. Experimental testing is the most rigorous way to disentangle the effect of one specific factor from others and to claim causality. To test for the impact of the most common wording variations, six different versions of a typical single item life satisfaction measure were used, shown in Table 2.

Table 2 – Survey versions (the six treatments)

#	Version	Question	Response Scale
1	***Positive*** (Tone)/ ***Reference***	All things considered, how **satisfied** are you with your life as a whole **these days**?	completely dissatisfied (0) – completely satisfied (10)

[1] Some additional observations had been registered but were deleted due to concerns of being accidentally registered test versions or from persons participating more than once. Specifically, data from any IP-address that appeared five or more times was deleted. Among the remaining observations, three IP-addresses are linked to more than one finished version. This might be explained by two different people using the same device. As a robustness check, the full analysis was repeated with a sample where only the first finished observation from each IP-address was included. The results are qualitatively similar between the two specifications. Additionally, one observation where an age of 11 years was indicated was excluded from the analysis.

2	***Balanced*** (Tone)	All things considered, how **satisfied or dissatisfied** are you with your life as a whole these days?	completely dissatisfied (0) – completely satisfied (10)
3	***Neutral*** (Tone)	All things considered, **how do you feel about** your life as a whole these days?	completely dissatisfied (0) – completely satisfied (10)
4	***Negative*** (Tone)	All things considered, how **dissatisfied** are you with your life as a whole these days?	completely dissatisfied (0) – completely satisfied (10)
5	***Completely*** (Label)/ ***No reference***	All things considered, how satisfied are you with your life?	**completely** dissatisfied (0) – **completely** satisfied (10)
6	***Totally*** (Label)	All things considered, how satisfied are you with your life?	**totally** dissatisfied (0) – **totally** satisfied (10)

Note: Bold type has been added for emphasis of between-scale differences in wording

The experimental design is between-person, meaning each participant was presented with only one of the six survey versions. The life satisfaction question was the first item of the survey to avoid order effects. Response time was tracked as an indicator of carefulness of consideration, although long response times can also indicate difficulties with answering the question rather than just more deliberate thinking (Yan & Tourangeau, 2008). Multiple scales are equivalent except for the wording aspect to be tested for. For example, items 1 and 2 are identical except that "satisfied" (positive tone) in the first, is replaced by "satisfied or dissatisfied" (balanced tone) in the second version. Any differences in life satisfaction mean, distribution and correlates can then be attributed to this particular wording variation. At least, this is true if the treatment groups are otherwise equal (or rather, sufficiently similar). To neutralize the effects of individual differences, assignment to treatments was random. Four versions are used to test question tone options (1, 2, 3, 4). Two versions each are used to investigate choice of reference period (1, 5) and variation in scale label wording (5, 6).

The choice of questions reflects wording variations between some of the major surveys with single-item life satisfaction scales (compare to Table 1). Not only are many studies based on data collected with these large surveys, but smaller surveys also tend to adopt the best-known items or at least use them as guideline in question construction. The item in version 1 (positive tone; reference period) is taken from the influential World Values Survey (WVS). One minor difference is that the numerical response scale in the WVS ranges from one to ten, but the lowest number on the response scale is zero in the survey used for this study. This was done for reasons of experimental control. Zero as lowest choice is used in other major surveys, for example the German Socio-Economic Panel (SOEP) and the Household, Income and Labour Dynamics in Australia (HILDA) ones. The survey versions to test scale labelling are inspired by the SOEP (version 5) and HILDA (version 6) items. Since version 5 is also used to test for the effect of including a reference period, by comparison to version 1, there is a need for consistency in response scales.

The four "tone" versions are all similar to the WVS item. Like in the WVS, tone is positive in SOEP and HILDA but the British Household Panel (BHP) employs balanced

phrasing. Again, for reasons of experimental control, the BHP question was not compared to the WVS one to test for the effect of tone because these two items differ in more than one respect. Instead, version 2 is exactly like version 1, other than using "satisfied or dissatisfied" instead of "satisfied". Negative and neutral tone items (versions 3 and 4, respectively) were created for completeness reasons.

Neither the BHP nor the SOEP surveys specify a reference period. Version 5, which is used to test the effect of including a reference period or not, is based on the SOEP life satisfaction item, which is more similar to the WVS item than the BHP one. To further the likeness, "all things considered" is put at the beginning of the sentence (rather than at the end, as it is in the SOEP). This way, the effect of reference period inclusion is not confounded with possible order effects. To investigate scale label wording, version 5 is compared to version 6. The latter, as previously mentioned, is based on the HILDA life satisfaction item, which differs from the SOEP only by using "totally" rather than "completely" in the wording of scale anchors.

The next block of questions elicits data on difficulty of understanding; social norm perception; domain salience and importance (in the social media subgroup also time period salience and importance); and scale interpretation. Recall, that these are possible mechanisms to explain why wording could have an effect. Specifically, respondents were asked how difficult they found the life satisfaction item they were presented with to understand; which domains and time periods they thought about when answering; what they thought the average answer given to the question by people in their country of residence was; and where there would place "a bit satisfied" and "a bit dissatisfied" on the response scale. The order of the questions was determined by a randomizer to prevent any systematic bias from possible order effects. Domains include: financial situation, health, achievements in life, personal relationships, work or study, feeling of safety and quality of the local environment. Table 3 shows a (slightly abbreviated version) of the relevant survey questions.

Table 3 – Mechanisms: survey questions

Mechanism	**Question**	**Response Scale**
Difficulty of Understanding	How difficult was it for you to understand the question and the response scale?	Easy (1) Neither easy nor difficult (2) Difficult (3)
Social Norm Perception	What would you estimate to be the average score given by people in your country of residence to this question?	completely (totally) dissatisfied (0) – completely (totally) satisfied (10)
Scale Interpretation	What number on the scale would correspond to being "a bit satisfied/ a bit dissatisfied" with life?	completely (totally) dissatisfied (0) – completely (totally) satisfied (10)

Domain Salience	Indicate to what extent the following domains have affected your answer: a) my ***financial*** situation b) my ***health*** c) my ***achievements*** in life d) my ***personal relationships*** e) my ***work***/***study*** f) my feeling of ***safety*** g) the quality of my local ***environment***	I ***didn't think about it*** when answering the question. (1) I thought about it but decided it was ***not important*** for choosing my answer. (2) I thought about it and decided it was ***somewhat important*** for choosing my answer. (3) I thought about it and decided it was ***very important*** for choosing my answer. (4)
Relevance of time periods	Indicate to what extent the following time periods have affected your answer: a) ***> 1 year ahead*** b) ≤ ***1 year*** *ahead* c) *the* ***past 7 days*** *(incl. today)* d) ***> 1 week ago – 6 months*** *ago* e) ***> 6 months*** *ago* ***– 5 years*** *ago* f) ***> 5 years*** *ago*	I ***didn't think about it*** when answering the question. (1) I thought about it but decided it was ***not important*** for choosing my answer. (2) I thought about it and decided it was ***somewhat important*** for choosing my answer. (3) I thought about it and decided it was ***very important*** for choosing my answer. (4)

Note: Respondents were asked to reconsider the specific question they had initially received. The response scales, as well, correspond to the respondent's treatment item (1-6). The exact phrasing of the question can be found in the appendix, section 7.1. Bold, italic type has been added for emphasis of essential information.

Later, respondents were also asked about their satisfaction with the same domains (financial situation, health, achievements in life, personal relationships, work or study, feeling of safety and quality of the local environment). This block was separated from the earlier questions by several other subjective well-being items, which were not directly relevant to this study. The separation should prevent any artificial correlation between life satisfaction and domain satisfaction due to a consistency motive or other reasons.

The survey also contained several other items that measure possible correlates of life satisfaction. This includes measures of personality, trust, optimism, perceived societal standing and materialism. (The specific scales are the *Ten Item Personality Inventory* (TIPI), the trust measure from the WVS, the *Life Orientation Test-Revised* (LOT-R), the *McArthur Scale of Subjective Social Status* and the *Material Values Scale* (MVS), respectively.) Moreover, the last part contained questions about demographics, which are also expected correlates of subjective well-being. Questions were about: perceived health; having a partner and/or children; education and employment; country of origin and residence; gender, age and income. To be able to control for possible translatability and comprehension issues, respondents were also asked to indicate whether English was their native language. The exact survey questions used in this study can be found in the appendix, section 7.1.

3.3. Analysis

In order to answer the research questions, the analysis will be centred around comparing the averages, dispersion and correlates of life satisfaction between the six treatments. In particular, scale versions that are equivalent except for the aspect of interest will be compared. To investigate tone, data from survey versions 1 through 4 are compared; to test for use of reference period, data from version 1 versus 5; and to test scale label wording, 5 versus 6. The results of these comparisons answer the sub-questions: "Does question tone (reference period use; scale anchor labelling) affect the average, dispersion and correlates of life satisfaction?" The conclusions regarding the three sub-elements, taken together, help to answer the main research question: "Does survey item wording influence life satisfaction reports?"

First, differences in reported life satisfaction between treatments in terms of averages will be tested. Although parametric tests (e.g. ANOVA) are more powerful than non-parametric alternatives, they might not be appropriate. A potential problem is that life satisfaction data tends to be skewed rather than normally distributed (Cummins, 2003). That would violate the assumptions of parametric mean comparison tests. The Shapiro-Wilks test is used to test the normality assumption. If indeed violated, rank-based methods like the Kruskal-Wallis or Mann Whitney U tests will be used. These tests are usually interpreted as comparing the medians of independent samples. They are appropriate for a categorical independent variable and a continuous dependent variable. Although life satisfaction is not strictly continuous, the ordinal data is often treated as cardinal (Dolan et. al, 2008). It has been found that treating data as cardinal instead of ordinal leads to similar results in the analysis of well-being determinants (Ferrer-i-Carbonell & Frijters, 2004). More direct evidence in favour of treating data cardinally comes from Kristoffersen (2017), who found a strong correlation between a mental health and a life satisfaction measure and (based on assumed linearity of the mental health response function) concludes that life satisfaction scores can be interpreted as monotonically increasing, with approximately equal distances between points on the response scale.

The Kruskal-Wallis test can compare several groups at once, for example to test whether any of the six versions is significantly different from the others. The Mann-Whitney U test can be used to test for differences between two groups. For example, if the Kruskal-Wallis shows significant differences, as is hypothesised for tone, the Mann-Whitney U test can be used to pin-point which scales exactly differ. An assumption that needs to be met in order for the results to be meaningful, is that variables besides life satisfaction do not significantly differ between the groups (or the results could be due to those factors rather than wording). Due to the randomization into treatments, it is expected that groups are similar in make-up (i.e. in terms of demographics and personality). If this assumption is not met, such factors would need to be controlled for.

Secondly, the six versions are compared in terms of dispersion. Specifically, the variance of life satisfaction scores is compared between the versions. The appropriate formal test for this is the Levene's test. If life satisfaction data is in fact skewed, an alternative specification of the Levene's test, as in Brown and Forthyte (1974), will be used. This variation of the tests uses more robust measures of centrality than the mean and has been found to perform well under non-normality. The method can compare several treatments at once, under the null hypothesis of homoscedasticity (equal variances). If all versions are tested simultaneously with

the result that at least one group is significantly different, as is hypothesized for tone, a pair-wise comparison can be used to find exactly which groups differ from each other. Again, the assumption should hold that groups are similar except for the treatment. Otherwise, possible differences in dispersion, or a lack thereof, could be due to differences between groups in terms of other factors than wording.

Thirdly, the pattern of correlates is compared between the versions. In a first step, bivariate correlation coefficients are considered, which show the association of two variables if no other factors are controlled for. If life satisfaction scores are in fact not normally distributed, spearman's correlation are computed, which are rank-based and thus distribution free. If the normality assumption does hold, Pearson correlations can be used instead. The coefficients are compared on a pairwise basis using z-tests (Fisher's z-transformation). In a second step, conditional correlations are considered. This involves controlling for other factors when investigating whether possible correlates have a different relationship to life satisfaction depending on wording. The purpose of conditional correlation analysis, in addition to comparing bivariate correlation coefficients, is to reduce the likelihood of omitted variable bias.

Studies about the determinants of subjective well-being generally model SWB as an additive function of social, economic, and environmental factors with individual differences in reporting captures in the error term (Dolan et al., 2008). Since treating life satisfaction data as cardinal is advantageous and has been show appropriate, ordinary least square (OLS) regression models can be used for the conditional correlation analysis. The calculated coefficients indicate the relationship between a variable and life satisfaction when other factors are accounted for. To determine whether this relationship depends on wording, an interaction term with the variable of interest and a survey version dummy is included. The procedure is explained in detail in the appendix, section 7.2.3.

Finally, explanatory mechanisms will only be tested, if any of the scales for tone, reference period or scale labelling produce significantly different results from each other in terms of life satisfaction average, distribution or correlates. As stated in the hypotheses, variations in tone are expected to cause such differences. In a first step, differences in the averages of explanatory variables (response time, ease of understanding, social norm perception, scale interpretation, and domain/time period consideration) can be tested in a similar fashion to testing the differences in reported life satisfaction. The appropriate test, again, depends on which assumptions are met. Further, if differences exist, the mechanism variables could be added into the regression analyses to see whether the differences between version, that had been found, disappear. More formal mediation analysis is also an option, depending on the results of the analysis thus far.

4. RESULTS

4.1. Descriptive statistics

The participants are not a homogeneous group or representative of any particular population. People coming from over 50 different countries participated. For the analysis, country of residence, rather than country of origin is considered. The participants resided in 40 different

countries, with over 70% living in European countries. Most often represented are people living in the Netherlands (30%) and in Germany (19%). Regarding non-Europeans, 17% of the full sample are from the Americas and 11% from other continents (Asia, Oceania and Africa). For 27% of them, English is their first language. The age ranged from 17 to 77, with an average of 27 years – although only one fifth were older than that; median age was 24. Most people were students (40% were students only and 16% were students that also work). 33% of respondents were non-student workers and 9% were unemployed or retired. Over 75% of people in the sample were highly educated (70% went to university or similar and 5% underwent vocational training after high school). The remainder had completed secondary education (24%) or less (<1%). The female to male ratio was around 6:4.

The descriptive statistics for all demographic and personality variables used in the analysis are presented in Table 4. Some categorical variables were transformed into binary dummies for the analysis for reasons of group size. Specifically, the transformations include: education (higher education or not), residence country (European or not) and employment (student or not, with student-workers counting as student). Moreover, no variable for having children or not was used in any part of the analysis because less than 10 people in each treatment were parents. All statistics are provided of the sample as a whole and per survey version. There seem to be notable differences between the six groups, probably due to the limited sample size. This is problematic because differences in life satisfaction between the treatment groups may be due to differences in group composition, rather than being due to wording effects. The reverse is also true: wording effects might be masked by group peculiarities, so that similar life satisfaction scores across treatments could be the result of other factors cancelling out wording effects.

The variation in income across treatments is particularly worrisome. For example, group 5 has a much higher average income than group 1 (over 60,000€/year compared to less than 25,000€/year), which could bias the analysis of reference period effects. In fact, of all variables, income has the most dispersed scores. Average household income is almost 40,000 Euros per year, but many declared zero income. Others have extremely high income: up to 1.7 million Euros per year. To reduce the effect of outliers, the logarithm of income is used in the analysis. However, this does not altogether solve the problem of group dissimilarity. Many other variables show notable between-group differences, for example: higher education (especially between balanced and negative tone) and student (higher ratio of students in groups 4 and 5 compared to 1); moreover, group three has particularly high proportions of females and people in relationships.

All demographic and control variables were formally tested to see whether between-group differences are significant. Age, (log) income, subjective health and the five personality dimensions were treated as continuous; group differences were assessed using the Kruskal-Wallis test, because the assumptions of ANOVA were not generally met. The significance of group dissimilarities in categorical variables (the dummies for gender, higher education, being a student, having a partner, European residence and being a native English speaker) are assessed with Pearson's chi-squared tests. None of the tests statistics regarding any of the demographic or personality variables were statistically significant at 5%.

Table 4 – Descriptive statistics: demographic and personality variables

Variable	Version / Statistic	All (n=328)	(1) *Positive* (Tone)/ *Ref* (n=52)	(2) *Balanced* (Tone) (n=57)	(3) *Neutral* (Tone) (n=50)	(4) *Negative* (Tone) (n=57)	(5) *Complet.* (Label)/ *No Ref.* (n=58)	(6) *Totally* (Label) (n=54)
Demographics								
Age	Mean (SD) *min; max*	26.55 (9.41) *17;77*	27.85 (12.07) *17;77*	26.60 (9.75) *18;62*	27.74 (10.80) *18;69*	25.39 (5.88) *18;43*	25.71 (8.39) *17;53*	26.31 (8.94) *18;62*
Female	%	60.98	59.62	54.39	72.00	57.89	60.34	62.96
Higher education	%	75.30	73.08	68.42	72.00	82.46	81.03	74.07
Student	%	56.40	50.00	49.12	54.00	61.40	65.52	57.41
Household Income (€/yr)	Mean (SD) *min; max*	38,743 (133,957) *0;1.7M*	23,717 (33,792) *0;136K*	32,127 (112,605) *0;850K*	23,110 (29,476) *0;130K*	34.74 (121.41) *0;900K*	61.159 (154.22) *0;1.02M*	54.82 (230.07) *0;1.7M*
Partner	%	42.26	42.31	45.61	60.00	47.37	41.38	48.15
Health (self-rated)	Mean (SD) *min; max*	3.95 (0.81) *1;5*	3.88 (0.83) *2;5*	3.81 (0.90) *1;5*	4.10 (0.79) *2;5*	3.88 (0.80) *2;5*	4.03 (0.75) *2;5*	3.98 (0.77) *2;5*
Europe	%	71.34	76.92	64.91	74.00	75.44	67.24	70.37
English Native	%	27.13	25.00	28.07	26.00	31.58	24.14	27.78
Personality								
Extraversion	Mean (SD) *min; max*	4.36 (1.39) *1;7*	4.38 (1.25) *1.5;6,5*	4.71 (1.15) *1.5;7*	4.32 (1.50) *1;7*	4.21 (1.48) *1;7*	4.16 (1.51) *1;6.5*	4.36 (1.39) *1.5;7*
Agreeableness	Mean (SD) *min; max*	4.19 (1.03) *1.5;7*	4.31 (1.09) *1.5;6.5*	4.10 (0.98) *2;6*	4.16 (0.92) *2;6.5*	3.99 (1.08) *1.5;7*	4.20 (1.06) *1.5;7*	4.39 (1.01) *2;7*
Conscientious-ness	Mean (SD) *min; max*	5.23 (1.27) *1;7*	5.35 (1.17) *3;7*	5.18 (1.26) *2;7*	5.37 (1.10) *2.5;7*	5.38 (1.22) *1.5;7*	5.13 (1.50) *1;7*	5.00 (1.33) *2;7*
Openness	Mean (SD) *min; max*	5.39 (0.99) *1.5;7*	5.46 (0.98) *2.5;7*	5.40 (1.17) *1.5;7*	5.41 (0.96) *2.5;7*	5.47 (0.98) *2.5;7*	5.25 (0.80) *3.5;6.5*	5.35 (1.03) *2;7*
Emotional Stability	Mean (SD) *min; max*	4.49 (1.32) *1.5;7*	4.53 (1.30) *1.5;7*	4.67 (1.34) *1.5;7*	4.50 (1.31) *2;7*	4.26 (1,34) *2;7*	4.48 (1.31) *2;7*	4.48 (1.32) *2;7*

4.2. Average

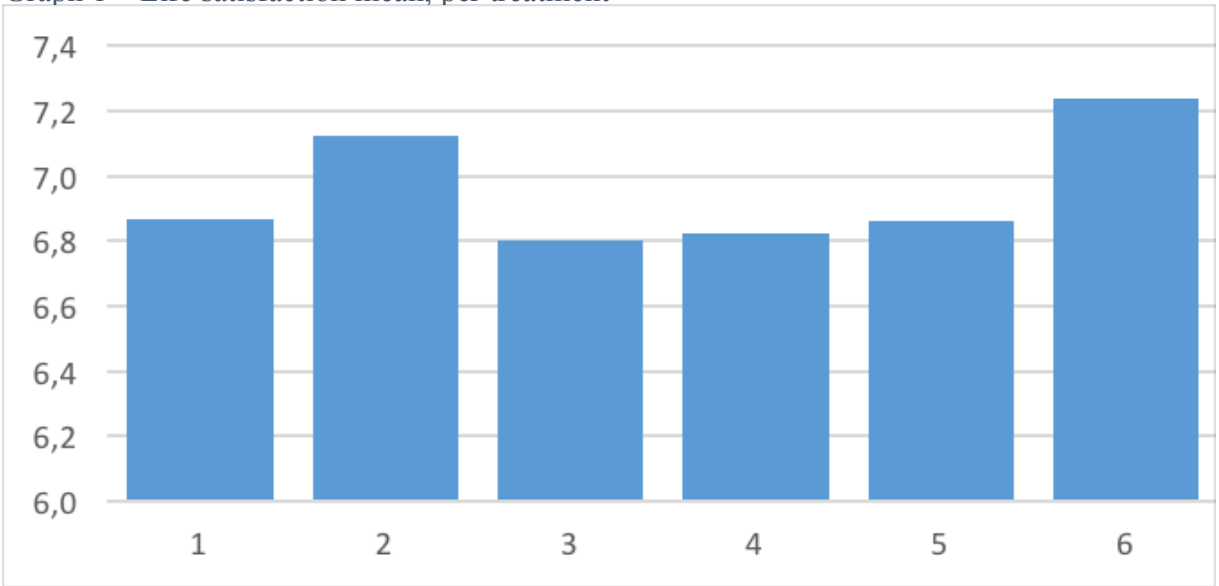
Table 5 shows descriptive and inferential statistics regarding the central tendency of life satisfaction scores, for the sample as a whole and for each of the treatments. The mean and median are around 7 for all groups. In fact, the median is 7 in all groups but group 6, where it is 8. Mean life satisfaction is 6.95 for the full sample. The group with the lowest mean life

satisfaction is group 3 (6.80), and the group with the highest average is group 6 (7.24). Graph 1 visualizes the mean scores of the six treatment groups. Among the tones, the balanced tone group has a slightly higher average score than the other three groups. Both reference period treatments (groups 1 and 5) have similar mean life satisfaction. Finally, group 6 (scale label with "totally") has a somewhat higher mean level than group 5 (scale label with "completely").

Table 5 – Life satisfaction: measures of central tendency

Variable	Version / Statistic	All	(1) *Positive* (Tone)/ *Ref*	(2) *Balanced* (Tone)	(3) *Neutral* (Tone)	(4) *Negative* (Tone)	(5) *Complet.* (Label)/ *No Ref.*	(6) *Totally* (Label)
Life satisfaction	Median	7	7	7	7	7	7	8
	Mean	6.95	6.87	7.12	6.80	6.83	6.86	7.24

Graph 1 – Life satisfaction mean, per treatment



To determine which method is most appropriate to compare averages, the normality and equal variance assumptions are tested. The Shapiro-Wilks test is used to assess normality. Both, for the sample as a whole and for each subgroup, the test is significant at the 1% level. This confirms that life satisfaction scores are in fact not normally distributed. The robust Levene's test specification, after Brown and Forthyte (1974), indicates no significant differences in the variance of life satisfaction scores between the different survey versions at the 5% level ($p=0.47$). The analysis so far suggests that life satisfaction is not normally distributed, that variances are equal between groups and that control variables show no significant differences between treatments. Given these results, the Kruskal-Wallis and the Mann-Whitney U tests are appropriate to compare groups. Results can be interpreted as comparing medians, given the equality in variances. No control variables are included in these analyses. The test results indicate no significant differences in medians between any of the survey versions at the 5% level. This holds when comparing tone (Kruskal-Wallis with versions 1,2, 3 and 4), reference period (Mann-Whitney U with versions 1 and 5) and scale labelling (Mann-Whitney U with versions 5 and 6) separately.

Even though no significant between-group differences in demographics and personality were found, the possibility of omitted variable bias is not eliminated. Therefore, OLS regressions with life satisfaction as dependent variable and a survey version dummy among the independent variables were performed. No differences between versions were found regardless of whether all demographic and personality variables (as listed in Table 4), a part (age, female, student, Europe and the five personality dimensions) or no controls were included. The exact procedure is described in the appendix, section 7.2.1. Taken together, the results suggest that item wording has no effect on average reported life satisfaction. This is in line with H2 and H3 (which predicted no differences in average life satisfaction depending on reference period use and scale labelling, respectively). The results are not in line with H1 and H1a (which predicted differences in average life satisfaction depending on tone; and higher (lower) average life satisfaction when tone is positive (negative) compared to balanced and neutral).

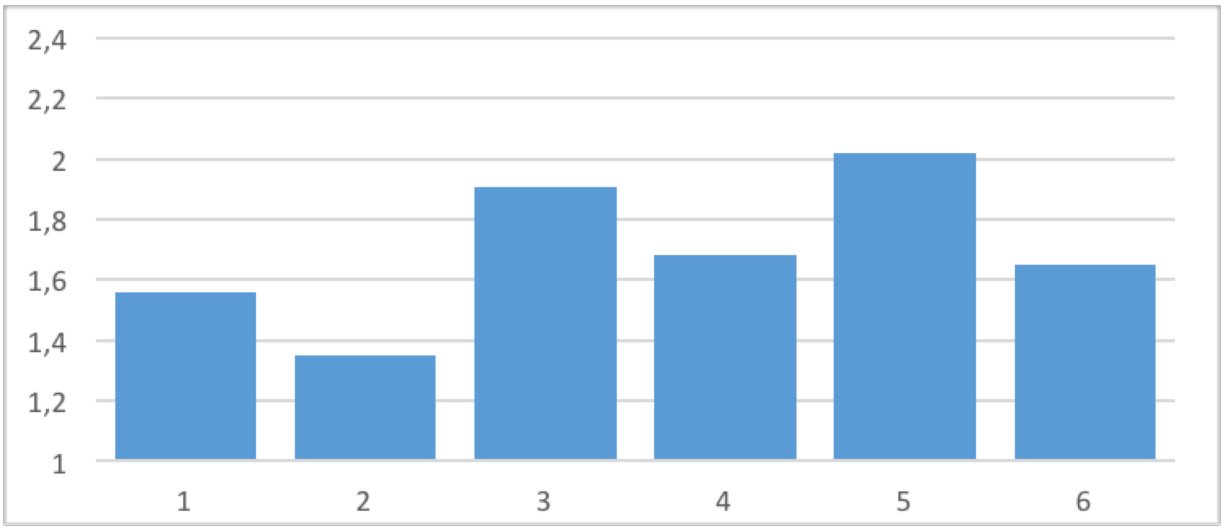
4.3. Dispersion

Table 6 includes measures of dispersion for the life satisfaction scores in the full sample and per treatment group. Graph 2 visualizes the standard deviation of life satisfaction scores, per treatment. Life satisfaction scores seem to be most dispersed in treatment group 5 and least dispersed in group 2.

Table 6 – Life satisfaction: measures of dispersion

	Version		**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Variable	Statistic	All	*Positive* (Tone)/ *Ref*	*Balanced* (Tone)	*Neutral* (Tone)	*Negative* (Tone)	*Complet.* (Label)/ *No Ref.*	*Totally* (Label)
Life satisfaction	Variance	2.90	2.43	1.82	3.63	2.83	4.09	2.72
	(SD)	(1.70)	(1.56)	(1.35)	(1.91)	(1.68)	(2.02)	(1.65)
	min; max	*0;10*	*3;10*	*3;10*	*0;10*	*2;10*	*0;10*	*3;10*
	IQ Range	*6 – 8*	*6 – 8*	*7 – 8*	*6 – 8*	*6 – 8*	*6 – 8*	*7 – 8*

Graph 2 – Life satisfaction standard deviation, per treatment



As a formal test, between-group equality of variance is investigated using the robust Levene's test specification after Brown and Forsythe (1974). No control variables are included in these

analyses. The results indicate no significant differences, regardless of whether all versions are compared simultaneously (p=0.47), or whether the versions for tone, reference period use and scale labelling are tested separately. Thus, variance is not significantly different depending on tone (p=0.46), reference period use (p=0.32) or scale anchor labelling (p=0.36).

Again, to reduce the risk of omitted variable bias and as a robustness check, controls were added to the analysis. Regressions were performed with a life satisfaction residual as dependent variable and a treatment dummy among the explanatory variables. (The residual is computed by taking the absolute distance of an individual's life satisfaction score from the mean life satisfaction in their treatment group.) No differences between question tone, reference period and scale label versions were found regardless of whether all demographic and personality variables (as listed in Table 4), a part (age, female, student, Europe and the five personality dimensions) or no controls were included. The exact procedure is illustrated in the appendix, section 7.2.2. Taken together, the results suggest that item wording has no effect on the dispersions of life satisfaction scores.[2] These results are in line with H2 and H3 (which predicted no differences in the dispersion of life satisfaction scores depending on reference period use and scale labelling, respectively). The results are not in line with H1 and H1a (which predicted differences life satisfaction score dispersion depending on tone; and higher dispersion when negative compared to positive, balanced or neutral).

4.4. Correlates

4.4.1. Bivariate correlations

As a first step in determining whether the pattern of life satisfaction correlates differs depending on item wording, bivariate correlations are considered. In other words, correlations between life satisfaction scores and other variables are considered one by one, without the inclusion of any control variables. Besides the demographic and personality variables (as in Table 4), domain satisfaction, as well as trust, materialism, social standing and optimism enter the analysis. Descriptive statistics for these additional variables can be found in the appendix, section 7.3. Correlations between life satisfaction and the other variables are computed using the spearman's rank procedure. (Pearson correlations are not appropriate, given the non-normal distribution of life satisfaction reports.) Fisher's z-transformation is applied in testing whether correlation coefficients of two groups are significantly different. Table 7 displays the correlation coefficients of each variable with life satisfaction per survey version and notes significant between-group differences.

For tone, any significant differences among the demographic variables are between balanced tone and another version. Being a student is more positively related to life satisfaction when tone is positive compared to balanced. Higher education is more positively correlated with life satisfaction when tone is balanced compared to all other tones. Lastly, having children

[2] Although not relevant for the analysis of tone, reference period or scale labelling, the difference in variance between groups 2 and 5 is significant in pair-wise comparison according to the robust specification of the Levene's test (p=0.04). The residual is also significantly higher in group 5 than 1, based on the regressions with no controls or the limited set of controls. When all demographic and personality variables are included, the difference between the groups is no longer significant.

is more positively related to life satisfaction when tone is positive compared to balanced. There are no significant differences regarding the personality variables. Regarding domain satisfaction, financial satisfaction is more positively correlated with life satisfaction when tone is positive compared to neutral. Achievement satisfaction is more positively correlated with life satisfaction when tone is negative, again, compared to neutral tone. Among the other variables, materialism more positively associated with life satisfaction when tone is balanced, in comparison to positive and negative tone.

Between version 1 (with reference period) and version 5 (without reference period), none of the coefficients of demographic or personality variables are significantly different from each other. Among the domain satisfaction and other variables, financial satisfaction and perceived social status are both more strongly correlated with life satisfaction when a reference period is mentioned compared to when it is not. Between the scale label treatments, extraversion is more strongly correlated with life satisfaction when "totally" (version 6) rather than "completely" (version 5) is used. Likewise, satisfaction with one's financial situation, safety and the local environment are all more strongly related to global satisfaction in version 6 compared to 5. Finally, social status also shows a significantly more positive coefficient in group 6 than in group 5.

Table 7 – Differences in bivariate correlations

Version / Variable	signif. differ.	(1) Positive (Tone)/ Ref.	(2) Balanced (Tone)	(3) Neutral (Tone)	(4) Negative (Tone)	(5) Completel. (Label)/ No Ref.	(6) Totally (Label)
Demographics							
Age		-0.25	0.00	-0.01	0.08	0.10	-0.07
Female		-0.24	-0.02	-0.04	-0.10	-0.05	0.09
Europe		0.10	-0.17	-0.03	-0.12	0.23	0.17
Student	(1)(2)	**0.23**	**-0.29***	-0.09	0.03	-0.01	0.07
Higher education	(1)(2) **(2)(3)** (2)(4)	-0.09	**0.33**	**-0.20**	-0.09	0.24	-0.02
Partner		0.04	0.23	-0.08	0.05	0.23	0.27
English Native		-0.08	-0.00	0.14	-0.05	-0.18	-0.25
Subj. Health		0.19	0.32*	0.21	0.12	0.19	0.34*
Ln Household Income (€/yr)		0.05	0.08	-0.09	0.22	-0.01	0.18
Personality							
Extraversion	(5)(6)	0.13	0.06	0.07	0.34*	0.12	0.49**
Agreeableness		-0.06	0.28*	0.08	0.00	0.05	-0.29*
Conscientious-ness		0.24	0.34**	0.31*	0.24	0.11	0.37**
Openness		0.10	0.16	0.15	0.32*	0.06	0.10
Emotional stability		0.27	0.16	0.37**	0.44**	0.40**	0.47**
Domain Salience							
Financial	(1)(3) (1)(5) (5)(6)	0.59**	0.33*	0.26	0.41**	0.27*	0.65**
Health		0.11	0.46**	0.21	0.33*	0.41**	0.58**

Achievement	(3)(4)	0.60**	0.43**	0.37**	0.67**	0.68**	0.59**
Relationships		0.49**	0.52**	0.28*	0.43**	0.61**	0.58**
Work/Study		0.60**	0.49**	0.52**	0.61**	0.58**	0.46**
Safety	(5)(6)	0.36**	0.44**	0.24	0.36**	0.14	0.50**
Environment	**(5)(6)**	0.49**	0.45**	0.44**	0.30*	**0.22**	**0.63****
Other Variables							
Trust		0.31*	0.35**	0.24	0.42**	0.38**	0.48**
Materialism	**(1)(2)** (2)(4)	**-0.27**	**0.24**	0.05	-0.18	-0.08	-0.18
Social Status	**(1)(5)** (5)(6)	**0.57****	0.38**	0.27	0.29*	**0.09**	0.44**
Optimism		0.32*	0.43**	0.16	0.46**	0.16	0.20

Note: Significant differences between pairs of coefficients are indicated by orange for the lower coefficient and green for the higher coefficient. Bold (or not) type indicates the significance level of the difference between two coefficients $p < 0.05$, **$p < 0.01$**. Significant differences of a coefficient from zero are indicated by asterisks. *$p < 0.05$, **$p < 0.01$.

4.4.2. Conditional correlations

To test whether omitted variable bias might drive the results of the bivariate correlation analysis, conditional correlations are examined. Specifically, this involves controlling for the influence of other variables when questioning, whether the relationship between life satisfaction and a given other variable depends on wording. This part of the analysis is done by regressing life satisfaction and three different sets of explanatory variables. The effect of wording on life satisfaction correlates is investigated by adding interaction terms of survey version and explanatory variables (one per regression). A detailed illustration of the analytic procedure can be found in the appendix, section 7.2.3.

One of the variable sets was satisfaction in the seven domains (finances, health, achievements, personal relationships, work/study, safety and environment) with control variables (age, gender, European residence, being a student and the five personality dimensions). Table 8 reports the coefficients of the domain satisfaction variables per survey version and notes where the association between domain and global satisfaction differs significantly depending on question tone, reference period or scale anchor wording. More significant differences are found than in the bivariate correlation analysis. Satisfaction in the financial, health, work/study, safety and environment domains has a different association with global satisfaction depending on question tone. Financial satisfaction is more positively related to life satisfaction when tone is positive compared to negative. Health and global satisfaction are more positive related when tone is balanced compared to neutral. Satisfaction with work or study has a more positive relationship to life satisfaction when tone is neutral compared to negative and safety satisfaction is more positively associated when tone is positive or neutral compared to negative. Finally, satisfaction with one's local environment shows a more negative association with life satisfaction when tone is negative compared all other tone versions. None of the above differences had been found significant in the bivariate correlation analysis and the differences in raw correlation coefficients that had been found regarding tone, are no longer significant.

Regarding the choice of reference period, versions 1 and 5 show significant differences in three domains. Satisfaction with one's financial situation, safety and environment are all more positively related with global satisfaction when a reference period is specified than when

no reference period is mentioned in the question. The difference in the financial domain had also been found when comparing the bivariate correlation coefficients, while no significant differences had been detected in the other two domains. Lastly, a similar pattern is found when comparing version 5 and 6 (as when comparing 5 and 1). Satisfaction with one's financial situation, safety and local environment has a more positive relation to global satisfaction when "totally" is used than when "completely" is used instead. Differences between versions 5 and 6 in the three domains had already been detected in the comparison of bivariate correlation coefficients.

Table 8 – Differences in conditional correlations: domain satisfaction

Versions:		(1)	(2)	(3)	(4)	(5)	(6)
Variable	signif. differ.	Positive (Tone)/ Ref	Balanced (Tone)	Neutral (Tone)	Negative (Tone)	Complet. (Label)/ No Ref.	Totally (Label)
Financial	**(1)(4)** **(1)(5)** (5)(6)	**0.27**** **(0.07)**	0.18* (0.08)	0.09 (0.09)	**-0.01** **(0.07)**	-0.01 (0.07)	0.21** (0.07)
Health	(2)(3)	0.05 (0.11)	0.15 (0.10)	-0.14 (0.10)	0.01 (0.09)	-0.06 (0.09)	0.15 (0.11)
Achievement		0.19 (0.10)	0.23 (0.12)	0.32** (0.10)	0.18 (0.10)	0.24** (0.09)	0.15 (0.09)
Relationships		0.06 (0.07)	0.21* (0.09)	0.17 (0.10)	-0.01 (0.08)	0.15* (0.07)	0.11 (0.08)
Work/Study	(3)(4)	0.12 (0.10)	0.20 (0.12)	0.26* (0.10)	0.00 (0.09)	0.05 (0.09)	0.01 (0.09)
Safety	(1)(4) (2)(4) (1)(5) (5)(6)	0.13 (0.09)	0.14 (0.11)	-0.06 (0.12)	-0.19 (0.02)	-0.16 (0.12)	0.18 (0.10)
Environment	**(1)(4)** (2)(4) **(3)(4)** (1)(5) (5)(6)	**0.24*** **(0.97)**	0.18 (0.11)	0.29* (0.11)	**-0.12** **(0.09)**	-0.01 (0.09)	0.21* (0.08)
+ Controls							

Note: The coefficients show the (conditional) association of the listed variables with life satisfaction, per survey version. Significant differences between versions in a category (tone, reference, label) are indicated by orange for the lower coefficient and green for the higher coefficient. Bold (or not) type indicates the significance level of the difference $p < 0.05$, **$p < 0.01$**. Significant differences of a coefficient from zero are indicated by asterisks. *$p < 0.05$, **$p < 0.01$. Standard error is given in parentheses. Control variables included are: age, female, Europe, student and personality (all five).

For the other two of the three variable sets, no significant differences between tone, reference period and scale label wording treatments were found. One of these sets of independent variables were all demographic and personality variables, as listed in Table 4. The lack of significant differences in the conditional correlation analysis is inconsistent with the results of the bivariate correlation analysis, where a few differences regarding the demographic and personality variables were found between question tone and scale label versions. The other set of explanatory factors that showed no significant differences in the conditional correlation analysis were trust, materialism, subjective social status and optimism (age, gender, European residence, being a student and the five personality dimensions were included to control for

individual differences). Compared to the bivariate correlation analysis, the differences in the association of materialism and life satisfaction (depending on tone) and of subjective social status and life satisfaction (depending on reference period choice and scale label wording) disappear. Coefficients for these two variable sets are reported in the appendix, section 7.2.

The differences between the results of the bivariate and the conditional correlation analyses indicate that omitted variable bias is likely a problem. Further, the conditional correlation analysis suggests that the most significant wording effects on life satisfaction correlates are in regards to the relationship of domain satisfaction and global satisfaction. Taken together, the results of the correlation analyses suggest that the pattern of correlates differs depending on tone, reference period use and scale labelling. This is in line with H1 (predicted differences in correlates depending on tone), but not in line with H2 and H3 (no predicted differences in correlates depending on reference period use and scale labelling, respectively). Overall, the evidence indicates that wording has an effect on life satisfaction correlates.

4.5. Mechanisms

In this sample, no significant differences in life satisfaction averages and dispersion were found between wording treatments. However, some significant differences were found in regards to correlate patterns. In particular, the association of satisfaction in some domains with overall life satisfaction has been found to differ depending on question tone, choice of reference period and scale label phrasing. In this section, possible explanations for the discrepancies, which were discussed in the theoretical framework, are investigated.

The variables collected for testing mechanisms were compared among the groups. Considered mediator variables include response time and ease of understanding, which indicate how carefully respondents considered the life satisfaction items and how much difficulty they had with understanding the questions. Moreover, differences in scale interpretation would be captured with variables indicating where respondents fit "a bit satisfied" and "a bit dissatisfied" on the response scale. Finally, the considered mediators also include variables indicating how much different time periods and domains influenced the respondent's evaluation, which is indicative of time period/domain salience and assigned weight. Descriptive statistics for these possible mediator variables can be found in Table 9.

Table 9 – Descriptive statistics of variables of possible mediator variables

	Version		**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Variable	Statistic	All	*Positive* (Tone)/ *Ref*	*Balanced* (Tone)	*Neutral* (Tone)	*Negative* (Tone)	*Complet.* (Label)/ *No Ref.*	*Totally* (Label)
Norm, Complexity, Interpretation								
Perceived Social Norm	Mean (SD) *min; max*	6.21 (1.52) *0;10*	6.06 (1.59) *2;8*	6.16 (1.44) *2;8*	6.14 (1.71) *0;10*	6.19 (1.58) *2;9*	6.41 (1.36) *2;9*	6.26 (1.49) *2;8*
Ease of Understanding	easy (%)	54.88	53.85	57.89	44.00	49.12	60.34	62.96
Response Time (seconds)	Mean (SD) *min; max*	18.87 (40.73)	17.44 (14.51)	20.26 (25.79)	14.11 (9.96)	30.90 (86.78)	15.56 (25.92)	13.61 (10.36)

		0.02;636	*0.02;79*	*0.02;166*	*4.77;53*	*3.86;636*	*0.02;170*	*0.05;43*
Scale Interpretation								
"a bit… …***satisfied***"	Mean (SD) *min; max*	5.63 (1.41) *0;9*	5.75 (1.24) *3;8*	5.60 (1.34) *1;8*	5.44 (1.62) *0;9*	5.91 (1.24) *2;9*	5.40 (1.65) *0;8*	5.67 (1.24) *2;9*
…***dissatisfied***"	Mean (SD) *min; max*	4.05 (1.50) *0;9*	4.02 (1.63) *0;8*	4.00 (1.32) *0;9*	3.86 (1.65) *0;9*	4.26 (1.38) *2;8*	3.98 (1.67) *0;9*	4.17 (1.38) *2;8*
Domain Consideration								
Financial	yes	70.43	65.38	68.42	66.00	68.42	81.03	72.22
Health	yes	70.12	69.23	75.44	66.00	63.16	74.14	72.22
Achievement	yes	83.23	88.46	78.95	76.00	80.70	93.10	81.48
Relationships	yes	87.20	76.92	82.46	88.00	87.72	96.55	90.74
Work/Study	yes	91.77	90.38	89.47	92.00	94.74	91.38	92.59
Safety	yes	45.12	50.00	42.11	36.00	42.11	50.00	50.00
Environment	yes	65.85	75.00	57.89	62.00	63.16	67.24	70.37
Time Period Consideration		n=158	n=23	n=26	n=20	n=30	n=31	n=28
Past 7 days (incl. today)	yes	70.89	78.26	76.92	70.00	66.67	61.29	75.00
More than 1 week ago – 6 months ago	yes	79.11	86.96	73.08	80.00	76.67	74.19	85.71
More than 6 months ago – 5 years ago	yes	53.80	65.22	53.85	50.00	46.67	58.06	50.00
More than 5 years ago	yes	22.78	21.74	19.23	30.00	20.00	22.58	25.00
The near future (up to 1 year ahead)	yes	73.42	69.57	76.92	80.00	70.00	70.97	75.00
The far future (more than 1 year ahead)	yes	43.67	34.78	45.00	45.00	40.00	54.84	50.00

Note: Colours indicate significant differences between groups with orange indicating a lower level of the mechanism variable and green a higher level. Sample size is smaller for time period consideration because the questions were only included in the survey for one of the three sub-samples (the group recruited via social media).

To begin with, all possible mediators (as listed in Table 9) were compared between tone, reference period and scale label versions to see whether there are any differences in terms of level (e.g. whether perceived social norm is higher in one group than the others or whether one of the versions was rated as more difficult to understand compared to the rest). Response time, social norm and scale interpretation were treated as continuous and versions were compared using the Kruskal-Wallis test, as the assumptions of ANOVA were generally not met. Ease of understanding, as well as domain and time period consideration, are coded as binary dummies for reasons of group size (respondent found the question easy or not; a domain/time period entered the life satisfaction evaluation or not). The dummy variables were assessed with Pearson's chi-squared test. In each case, the independent variable is a treatment version dummy (thus, the six treatments are compared simultaneously).

The only significant difference ($p<0.05$) is found in regards to consideration of the personal relationships domain. To pin-point which versions differ, the chi-squared test was repeated three more times, by category (tone, reference, label). The treatments that differ are versions 5 and 1. Thus, the degree to which people think of personal relationships and find them "somewhat" or "very" important appears to depend on wording. Specifically, more people consider personal relationships in their life evaluations when no reference period is mentioned. This fits with the conditional correlation analysis, where the association between relationship and global satisfaction was found to be stronger when no reference period is specified than when "these days" is used. The result is also consistent with the theoretical consideration that people consider a shorter period, when reading a term like "these days" than when no time is specified. It is plausible that people consider personal relationships as very important in the big picture, but neglect them when thinking about recent developments (at least, if there were no big changes regarding their social connections).

Although the significant finding is consistent with theoretical considerations and the correlation analysis, it is only one single piece of evidence and not enough to draw a definitive conclusion. The fact that no other differences were detected suggests that, in this sample, the differences in correlation patterns between the life satisfaction items under investigation cannot generally be explained by the considered mechanisms. Therefore, no more formal mediation analysis was pursued.

5. DISCUSSION AND CONCLUSION

The goal of this study was to test how the wording of life satisfaction items influences responses. By means of an experiment, four question tone options, a version with and a version without reference period, as well as two scale labels variants were investigated. Specifically, life satisfaction average, dispersion and correlates were compared within each category (tone, reference and label). The analysis failed to show any significant differences regarding life satisfaction average and dispersion between the survey versions under investigation. Some differences in correlates were found. In particular, it appears that the association of satisfaction with life as a whole and in certain domains is influenced by item wording. The difference in the association of global and relationship satisfaction, depending on reference period use, may be explained by wording affecting the degree to which the domain is salient and appears relevant. Otherwise, no significant differences between treatment groups in the level of possible mediator variables were found. As a whole, the results suggest that differently phrased survey items lead to similar life satisfaction reports in terms of the overall distribution, but that the analysis of the well-being determinants could be biased by wording effects. Although domain consideration might be a mediator, the detected differences cannot generally be explained by the tested mechanisms (ease of understanding, carefulness of consideration, scale interpretation, consideration of domains and time periods).

There are a few limitations, which render the results less conclusive than desired. The most important caveats concern sample size and heterogeneity. Maybe most problematic was the limited sample size, in combination with some notable dissimilarities between individuals. For example, most people were young adults, had low income and no children; yet, some people were older adults, had very high income and/or children. All of these factors are known to affect

satisfaction with life. Due to the small size of some subgroups, it was not possible to control for the effect of certain factors (i.e. having children) or only to an extent (e.g. residence country being in Europe or not). An indication for the problem are the conditional correlation results, as presented in Table 8. The differences between groups 1 and 5, in terms of direction and significance, are very similar to the differences between groups 5 and 6. Based on theoretical considerations, differences between the reference period treatments are not entirely implausible, but no differences were expected between the label treatments. It seems possible that the disparities between group 5, on the one hand, and groups 1 and 6, on the other hand, are driven by peculiarities in the make-up of group 5, rather than by wording effects.

Furthermore, an inherent limitation of the experimental design in this study was that within-person comparisons over time were not possible. Yet, particular in exploring the effect of including are reference period or not, such data could provide valuable insights. Questions like whether scores are less stable over time if a reference period is mentioned than when it is not, require testing at multiple points in time. Moreover, a within-person design has advantages in terms of controlling for fixed effects (the effects of factors which are stable over time).

The contribution of this study to the literature is in providing a systematic investigation of some of the obvious differences in life satisfaction item wording between influential surveys. However, the findings are only a broad indication of the potential role of wording and have very restricted explanatory power because of the limitations. The results are not to be seen as proof that wording has no effect on life satisfaction averages and dispersion. The differences in correlate patterns found in this sample should also not be seen as generalizable. The study does not deliver conclusive evidence that mechanisms like difficulties with understanding a question play no role in the context of life satisfaction surveys. Finally, the study does not show that theories like priming, framing and anchoring do not apply to life evaluations. A cautious response to the research question of this study ("does survey item wording influence life satisfaction scores?") might be that wording influences life satisfaction scores in some ways, but not above and beyond other well-being determinants. This conclusion fits with the general literature on the validity, reliability and comparability of single-item life satisfaction scales and the role of context effects: although situational factors influence self-reported life satisfaction, the influences are not so large to completely preclude valid and reliable measurement.

Most of all, the limitations of this study should be seen as opportunities for future research. Although a large sample size and repeated testing of the same individuals were beyond the scope of this thesis, such designs are feasible for other researchers. Especially, the creators and distributors of major surveys could further study the question of wording effects, as they often have access to large groups of people over many years. This could include experiments with test-groups or a strategical implementation of new wording on the main survey. For example, by phasing-in an item with a changed aspect of wording over several years in a random order (e.g. by region), the effect of the wording variation could be isolated relatively well. Until more evidence regarding the exact effects of different wording variations and the mechanism driving them is available, researchers should exercise caution in using data from more than one source. Although it is not known whether a particular item is most useful, comparability concerns can be minimized by standardized wording. Thus, a final recommendation to creators of new surveys is to adopt one of the popular life satisfaction items rather than creating yet another version.

6. REFERENCES

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action. *Journal of Personality and Social Psychology, 71*(2), 230-244.

Bower, G. H. (1981). Mood and Memory. *American Psychologist, 36*(2), 129-148.

Brown, M. B., & Forsythe, A. B. (1974). Robust Tests for the Equality of Variances. *Journal of the American Statistical Association, 69*(436), 364-367.

Chapman, G. B., & Johnson, E. J. (1999). Anchoring, Activation, and the Construction of Values. *Organizational Behavior and Human Decision Processes, 79*(2), 115-153.

Csikszentmihalyi , M., & Larson, R. (2014). The Experience Sampling Method. In M. Csikszentmihalyi, *Flow and the Foundations of Positive Psychology* (pp. 21-34). Dordrecht: Springer.

Cummins, R. A. (2003). Normative Life Satisfaction: Measurement Issues and a Homeostatic Model. *Social Indicators Research, 64*, 225-256.

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

Di Tella, R., & MacCulloch, R. (2006). Some Uses of Happiness Data in Economics. *Journal of Economic Perspectives, 20*, 25-46.

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

Dolan, P., Peasgood, T., & White, M. (2008). Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being. *Journal of Economic Psychology*, 94-122.

Eid, M., & Diener, E. (2004). Global Judgments of Subjective Well-Being: Situational Variability and Long-Term Stability. *Social Indicators Research, 65*, 245-277.

Ferrer-i-Carbonell, A., & Frijters, P. (2004). How Important is Methodology for the Estimates of the Determinants of Happiness. *The Economic Journal, 114*, 641-659.

Fujita, F., & Diener, E. (2005). Life Satisfaction Set Point: Stability and Change. *Journal of Personality and Social Psychology, 88*(1), 158-164.

Furnham, A., & Boo, H. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics, 40*, 35-42.

Helliwell, J. F., Sachs, J. D., & Layard, R. (2013). *World Happiness Report.* Earth Institute, Columbia University.

Janiszewski, C., & Wyer Jr., R. S. (2014). Content and process priming: A review. *Journal of Consumer Psychology, 24*(1), 96-118.

Kahneman, D., & Krueger, A. B. (2006). Developments in the Measurement of Subjective Well-Being. *Journal of Economic Perspectives, 20*(1), 3-24.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *Science, 306*, 1776-1780.

Kristoffersen, I. (2017). The Metrics of Subjective Wellbeing Data: An Empirical Evaluation of the Ordinal and Cardinal Comparability of Life Satisfaction Scores. *Social Indicators Research, 130*, 845-865.

Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology, 5*, 213-236.

Krueger, A. B., & Schkade, D. A. (2008). The reliability of subjective well-being measures. *Journal of Public Economics , 92*, 1833-1845.

LeBoeuf, R. A., & Shafir, E. (2003). Deep Thoughts and Shallow Frames: On the Susceptibility to Framing Effects . *Journal of Behavioral Decision Making, 16*, 77-92.

Levin, I. P., Johnson, R. D., Russo, C. P., & Deldin, P. J. (1985). Framing Effects in Judgment Tasks with Varying Amounts of Information. *Organizational Behaviour and Human Decision Processes, 36*, 362-377.

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All Frames Are Not Created Equal: A Typology and Critical Analysis of Framing Effects. *Organizational Behavior and Human Decision Processes, 76*, 149-188.

Lietz, P. (2010). Research intro questionnaire design: A summary of the literature. *International Journal of Market Reserch, 52*(2), 249-273.

Lucas, R. E., & Donnellan, B. M. (2007). How stable is happiness? Using the STARTS model to estimate the stability of life satisfaction. *Journal of Research in Personality, 41*, 1091-1098.

Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in Cognitive Sciences, 14*(10), 435-440.

Mullainathan, S., & Gruber, J. (2005). Do Cigarette Taxes Make Smokers Happier? *Advances in Economic Analysis and Policy, 5*(1), 1-43.

Mussweiler, T., & Strack, F. (1999). Hypothesis-Consistent Testing and Semantic Priming in the Anchoring Paradigm: A Selective Accessibility Model . *Journal of Experimental Social Psychology, 35*, 136-164.

National Research Council. (2013). *Subjective Well-Being: Measuring Happiness, Suffering, and Other Dimensions of Experience.* (A. A. Stone, & C. Mackie, Eds.) Washington, DC: The National Academic Press.

OECD. (2013). *OECD Guidelines on Measuring Subjective Well-being.* Paris: OECD Publishing. http://dx.doi.org/10.1787/9789264191655-en

Oswald, A. J., & Wu, S. (2010). Objective Confirmation of Subjective Measures of Human Weil-Being: Evidence from the U.S.A. *Science, 327*(5965), 576-579.

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.

Ralph, K., Palmer, K., & Jayne, O. (2011). Subjective Well-being: a qualitative investigation of subjective well-being questions. webarchive.nationalarchives.gov.uk/20120713154642/http://www.ons.gov.uk/ons/guide-method/user-guidance/well-being/advisory-groups/well-being-technical-advisory-group/working-paper---subjective-well-being--a-qualitative-investigation-of-subjective-well-being-questions.pdf?format=contrast

Robinson, M. D. (2000). The Reactive and Prospective Functions of Mood: Its Role in Linking Daily Experiences and Cognitive Well-being. *Cognition & Emotion, 14*(2), 145-176.

Sandvik, E., Diener, E., & Seidlitz, L. (1993). Subjective well-being: The convergence and stability of self- report and non-self-report measures. *Journal of Personality, 61*(317-342).

Schimmack, U., & Oishi, S. (2005). The Influence of Chronically and Temporarily Accessible Information on Life Satisfaction Judgments. *Journal of Personality and Social Psychology, 89*(3), 395-406.

Schwarz, N. (2007). Cognitive Aspects of Survey Methodology . *Applied Cognitive Psychology, 21*, 227-287.

Schwarz, N., & Clore, G. L. (1983). Mood, Misattribution, and Judgments of Well-Being: Informative and Directive Functions of Affective States. *Jounal of Personality and Social Psychology, 45*(3), 513-523.

Schwarz, N., & Strack, F. (1999). Reports of subjective well-being: judgmental processes and their methodological implications. (D. Kahneman, E. Diener, & N. Schwarz, Eds.) *Well-being: The foundations of hedonic psychology*, 61-84.

Schwarz, N., Strack, F., Bishop, G., & Hippler, H. -J. (1991). The Impact of Administration Mode on Response Effects in Survey Measurement. *Applied Cognitive Psychology, 5*, 193-212.

Schwarz, N., Strack, F., Kommer, D., & Wagner, D. (1987). Soccer, rooms, and the quality of your life: Mood effects on judgments of satisfaction with life in general and with specific domains. *European Journal of Social Psychology, 17*, 69-79.

Strack, F., Martin, L. L., & Schwarz, N. (1988). Priming and communication: Social determinantsof information use in judgments of life satisfaction. *European Journal of Social Psychology, 18*, 429-442.

Strack, F., Schwarz, N., Chassein, B., Kern, D., & Wagner, D. (1985). The salience of comparison standards and the activation of social norms: Consequences for judgments of happiness and their communication. *British Journal of Social Psychology, 29*, 303-314.

Stutzer, A., & Frey, B. S. (2010). Recent Advances in the Economics of Individual Subjective Well-Being. *Social Research*, 679-714.

Tay, L., Chan, D., & Diener, E. (2014). The Metrics of Societal Happiness. *Social Indicators Research*, 577-600.

Thomas, D. L., & Diener, E. (1990). Memory Accuracy in the Recall of Emotions. *Journal of Personality and Social Psychology, 59*(2), 291-297.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124-1131.

Tversky, A., & Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science, 211*(4481), 453-458.

van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of Reverse Wording of Questionnaire Items: Let's Learn from Cows in the Rain. *PLoS ONE, 8*(7).

Williams, L. E., & Bargh, J. A. (2008). Experiencing Physical Warmth Promotes Interpersonal Warmth. *Science, 322*(5901), 606-607.

Yan, T., & Tourangeau, R. (2008). Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times. *Applied Cognitive Psychology, 22*, 51-68.

7. APPENDIX

7.1. Survey

This section illustrates the structure of the survey used to collect data and shows the exact phrasing of all survey questions used in the analysis.

Introductory text:

"Thank you for participating in this survey. This survey takes about 10 minutes to complete. Your answers remain anonymous and will be treated confidentially. Please be honest in your responses. If you have any questions about this questionnaire, please email hendriks@ese.eur.nl"

Table 10 – Survey block I: life satisfaction item

#	Version	Question	Response Scale
1	***Positive*** (Tone)/ ***Reference***	All things considered, how satisfied are you with your life as a whole these days?	completely dissatisfied (0) – completely satisfied (10)
2	***Balanced*** (Tone)	All things considered, how satisfied or dissatisfied are you with your life as a whole these days?	completely dissatisfied (0) – completely satisfied (10)
3	***Neutral*** (Tone)	All things considered, how do you feel about your life as a whole these days?	completely dissatisfied (0) – completely satisfied (10)
4	***Negative*** (Tone)	All things considered, how dissatisfied are you with your life as a whole these days?	completely dissatisfied (0) – completely satisfied (10)
5	***Completely*** (Label)/ ***No reference***	All things considered, how satisfied are you with your life?	completely dissatisfied (0) – completely satisfied (10)
6	***Totally*** (Label)	All things considered, how satisfied are you with your life?	totally dissatisfied (0) – totally satisfied (10)

Note: Every participant was randomized into one of the treatments (there were two more versions, which were not relevant to this study).

Table 11 – Survey block II: possible mechanisms

Mechanism	Question	Response Scale
Domain Salience	Indicate to what extent the following domains have affected your answer to the question (insert life satisfaction question) a) my financial situation b) my health c) my achievements in life d) my personal relationships e) my work/study f) my feeling of safety g) the quality of my local environment	I didn't think about it when answering the question. (1) I thought about it but decided it was not important for choosing my answer. (2) I thought about it and decided it was somewhat important for choosing my answer. (3) I thought about it and decided it was very important for choosing my answer. (4)
Social Norm Perception	Consider again the question (insert the initially presented life satisfaction question) What would you estimate to be the average score given by people in your country of residence to this question?	completely (totally) dissatisfied (0) – completely (totally) satisfied (10) (corresponding to the initially presented life satisfaction item)
Scale Interpretation	Consider again the question (insert the initially presented life satisfaction question) Only the lowest and highest number on the response scale were labelled. We could also assign labels to the other numbers on the scale. a) In your opinion, what number on the scale corresponds to being "a bit satisfied" with life? b) And what number on the scale would correspond to being “a bit dissatisfied” with life? (b)	completely (totally) dissatisfied (0) – completely (totally) satisfied (10) (corresponding to the initially presented life satisfaction item)
Difficulty of Understanding	Consider again the question (insert the initially presented life satisfaction question) How difficult was it for you to understand the question and the response scale?	Easy (1) Neither easy nor difficult (2) Difficult (3)
Relevance of time periods	Indicate to what extent the following time periods have affected your answer to the question (Insert matching life satisfaction question) a) More than 1 week ago – 6 months ago b) More than 6 months ago – 5 years ago c) More than 5 years ago d) The near future (up to 1 year ahead) e) The far future (more than 1 year ahead)	I didn't think about it when answering the question. (1) I thought about it but decided it was not important for choosing my answer. (2) I thought about it and decided it was somewhat important for

		choosing my answer. (3) I thought about it and decided it was very important for choosing my answer. (4)

Note: The order of the questions in this block is determined by a randomizer.

Table 12 – Survey block III: correlates and controls

	Question	**Response Scale**
Domain Satisfaction	How satisfied are you with your a) financial situation b) health c) achievements in life d) personal relationships e) work/study f) feeling of safety g) the quality of your local environment	Not at all satisfied (0) – completely satisfied (10)
Personality (TIPI)	Here are a number of personality traits that may or may not apply to you. Indicate the extent to which you agree or disagree with that statement. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other. a) Extraverted, enthusiastic b) Critical, disagreeable c) Reliable, self-disciplined d) Anxious, easily upset e) Open to new experiences, complex f) Reserved, quiet g) Sympathetic, quiet h) Disorganized, careless i) Calm, emotionally stable j) Conventional, uncreative	Strongly disagree (1) Disagree (2) Somewhat disagree (3) Neither agree nor disagree (4) Somewhat agree (5) Agree (6) Strongly agree (7)
Trust (WVS)	Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?	You can't be too careful (0) – Most people can be trusted (10)
Optimism (LOT-R)	Indicate the extent to which you agree or disagree with the following statements: a) In uncertain times, I usually expect the best. b) If something can go wrong for me, it will. c) I'm always optimistic about my future. d) I hardly ever expect things to go my way. e) I rarely count on good things happening f) to me. g) Overall, I expect more good things to happen to me than bad.	Strongly disagree (1) Disagree (2) Somewhat disagree (3) Neither agree nor disagree (4) Somewhat agree (5) Agree (6) Strongly agree (7)

Subjective Social Status (McArthur)	There are people who tend to be towards the top of our society and people who tend to be towards the bottom. Please imagine a ladder with steps numbered from zero at the bottom to ten at the top. Suppose we say that the top of the ladder represents the top of our society and the bottom of the ladder represents the bottom of our society. If the top step is 10 and the bottom step is 0, where would you place yourself on this scale nowadays?	0 – 10
Material Values Scale (MVS)	a) I admire people who own expensive homes, cars, and clothes. b) The things I own say a lot about how well I'm doing in life. c) Buying things gives me a lot of pleasure. I like a lot of luxury in my life. d) My life would be better if I owned certain things I don't have. e) I'd be happier if I could afford to buy more things.	Strongly disagree (1) Somewhat disagree (2) Neither agree nor disagree (3) Somewhat agree (4) Strongly agree (5)
Subjective Health	How would you describe your current health?	Bad (1) Poor (2) Satisfactory (3) Good (4) Very good (5)
EN Native	Is English your first language?	Yes (1) No (2)
Children	Do you have any children?	Yes (1) No (2)
Partner	Do you have a partner?	Yes (1) No (2)
Education	What is the highest educational level that you have attained?	Primary education or no formal education (1) Secondary education (2) Vocational training (3) Tertiary education (4)
Country of birth & residence	What is your country of birth? What is your country of residence?	Drop down menu: ▼ Afghanistan (1) ... Other (194)
Employment	Are you currently employed or not? (multiple answers possible)	Full time paid employee (30 hours or more a week) (1) Part time paid employee (less than 30 hours a week) (2) Self-employed (3) Unemployed (4) Retired/pensioned (5) Housekeeper (6) Full-time student (7) Part-time student (8)

Gender	Your gender:	Male (1), Female (2)
Age	Your age in numbers:	
Income	Finally, we are interested in knowing something about your income. Please indicate your household's total income, after tax and compulsory deductions, from all sources? If you don't know the exact figure, please give an estimate. Indicate first in what currency you earn your income. Amount in numbers (do not include dots or commas and round to a whole number):	Drop down menu (currency) ▼ Euro (1) ... Zimbabwe Dollar (161)

Note: Blocks II and III were separated by subjective well-being items not relevant to this study.

7.2. Adding controls: analytic procedure

7.2.1. Conditional average

This section illustrates how conditional averages were analysed. OLS regressions were performed, with life satisfaction as the dependent variable and a dummy variable for survey version among the explanatory variables. First, the full set of demographic and personality variables was included as controls (all the variables listed in Table 4); second, a subset (age, gender, being a student, European residence and the five personality dimensions); and third, no controls.

The survey version dummy indicates whether life satisfaction is significantly higher or lower in the five other treatments compared to a chosen base category. For each set of variables, the regression was repeated with different reference categories, so that each of the relevant comparisons could be made (between all versions in one of the three categories tone, reference and labelling).

Set I: **Version dummy** + all demographic and personality variables (as in Table 4)
Set II: **Version dummy** + age, female, student, Europe, personality (all five)
Set III: **Version dummy**

To exemplify, Table 13 shows the output of the first regression. Life satisfaction was regressed with the first variable set: the treatment dummy as well as all demographic and personality variables. Version 1 was specified as the base category. The coefficients of the version dummy can be interpreted as: the level of life satisfaction does not significantly differ between positive tone compared to balanced, neutral or negative tone (p-values of categories 2, 3 and 4 are above 0.05). Moreover, the effect of age is not significantly different when the reference period "these days" is mentioned compared to when no reference period is specified (p-value of category 5 is above 0.05).

Table 13 – Conditional averages: regression 1 output.

Variable		**Coefficient**	**Standard Error**	**P-Value**
Version	**1**	**(base)**		
	2	**0.21**	**(0.31)**	**0.50**
	3	**-0.16**	**(0.32)**	**0.62**
	4	**0.06**	**(0.31)**	**0.85**
	5	**-0.07**	**(0.31)**	**0.82**
	6	**0.37**	**(0.31)**	**0.24**
Age		-0.02	(0.01)	0.07
Female		-0.01	(0.15)	0.96
Europe		-0.13	(0.24)	0.60
Student		-0.15	(0.21)	0.48
Higher Education		-0.01	(0.22)	0.98
Partner		0.21	(0.19)	0.26
English Native		-0.15	(0.23)	0.52
Subjective Health		0.29	(0.11)	0.01
Household Income		0.03	(0.03)	0.37
Extraversion		0.17	(0.08)	0.02
Agreeableness		0.14	(0.10)	0.14
Conscientiousness		0.20	(0.08)	0.01
Openness		0.09	(0.10)	0.33
Emotional Stability		0.33	(0.07)	0.00
Constant		1.70	(0.94)	0.07

7.2.2. Conditional dispersion

This section illustrates how conditional dispersion was analysed. First a life satisfaction residual was calculated for each individual (the absolute difference between treatment group mean and the individual score). OLS regressions were performed with the residual as dependent variable, and same three sets of independent variables as in the conditional average analysis. The version dummy indicates, whether the difference of individual scores from the mean is significantly larger or smaller depending on wording in the five other treatments compared to a chosen base category. Again, for each set of variables, the regression was repeated with different reference categories, so that each of the relevant comparisons could be made (between all tone, reference and label versions).

Set I: **Version dummy** + all demographic and personality variables (as in Table 4)
Set II: **Version dummy** + age, female, student, Europe, personality (all five)
Set III: **Version dummy**

To exemplify, Table 14 shows the output of the first regression. The life satisfaction residual was regressed with the first variable set: the treatment dummy as well as all demographic and personality variables. Version 1 was specified as the base. The coefficients of the version dummy can be interpreted as: the dispersion (absolute distance of individual scores from the mean) of life satisfaction does not significantly differ between positive tone compared to balanced, neutral or negative tone (p-values of categories 2, 3 and 4 are above 0.05). Moreover,

it is not significantly different when the reference period "these days" is mentioned compared to when no reference period is specified (p-value of category 5 is above 0.05).

In the next regression, version 2 is specified as the base. Then, the version dummy indicates whether dispersion is significantly different between balanced tone, on the one hand, and neutral or negative tone, on the other hand (differences to positive tone are also indicated, but these are already known from the first regression). The regression is repeated two more times with base groups 3 (to compare neutral and negative tone) and 5 (to compare the labels "totally" and "completely"). The whole process (four regressions) is repeated for the other two sets of variables.

Table 14 – Conditional dispersion: regression 1 output.

Variable		**Coefficient**	**Standard Error**	**P-Value**
Version	**1**	**(base)**		
	2	**-0.23**	**(0.23)**	**0.32**
	3	**0.23**	**(0.24)**	**0.33**
	4	**0.08**	**(0.23)**	**0.72**
	5	**0.43**	**(0.23)**	**0.06**
	6	**0.05**	**(0.23)**	**0.82**
Age		0.02	(0.01)	0.05
Female		-0.01	(0.15)	0.96
Europe		-0.28	(0.18)	0.11
Student		0.22	(0.15)	0.15
Higher Education		0.04	(0.16)	0.82
Partner		-0.08	(0.14)	0.57
English Native		-0.16	(0.17)	0.35
Subjective Health		-0.12	(0.09)	0.15
Household Income		-0.01	(0.02)	0.54
Extraversion		-0.02	(0.06)	0.73
Agreeableness		0.03	(0.07)	0.73
Conscientiousness		-0.02	(0.06)	0.68
Openness		-0.07	(0.07)	0.32
Emotional Stability		-0.05	(0.06)	0.39
Constant		2.14	(0.70)	0.00

7.2.3. Conditional correlates

This section illustrates how conditional correlations were analysed. OLS regressions were performed with life satisfaction as the dependent variable and three different sets of explanatory variables. To investigate how the relationship of life satisfaction and a given correlate depends on wording, an interaction term of that variable with a dummy for survey version is included (one per regression).

Set I: **Version*__: + demographic and personality** variables (as in Table 4)
Set II: **Version*__: trust, materialism, social standing, optimism** (+controls)
Set III: **Version*__: domain satisfaction** (+controls)

To illustrate, Table 15 shows output of the first regression in the conditional correlation analysis. Life satisfaction was regressed with the first set of explanatory variables: all demographic and personality variables (as listed in Table 4). Version 1 ("positive tone/reference") was specified as the reference category. This first regression is informative about the relationship between age and life satisfaction when tone is positive/a reference period is mentioned. The coefficient of age in Table 14 is not significantly different from zero (p=0.15). Thus, there is no significant relationship between age and life satisfaction in treatment 1. Even more importantly, the output is informative about how the life satisfaction and age relationship differs in the other versions compared to version one. The key variable for the conditional correlation analysis is the interaction term. The output in Table 14 regarding that variable can be interpreted as: the association of life satisfaction does not differ between positive tone compared to balanced, neutral or negative tone (p-values of categories 2, 3 and 4 are above 0.05). Moreover, the effect of age is not significantly different when the reference period "these days" is mentioned compared to when no reference period is specified (p-value of category 5 is above 0.05).

The second regression is equivalent except version 2 is specified as the base. The age coefficient then indicates the association of age and life satisfaction when tone is balanced. The interaction term indicates, whether the relationship is different compared to when tone is neutral or negative (whether it differs between positive and balanced tone is already known from the first regression). The regression is repeated again until each group has been the base. In each case, the age coefficient is informative about the relationship of age and life satisfaction in the base group. Additionally, the interaction term when group 3 is the base is informative about differences in the life satisfaction – age – association between neutral and negative tone, while differences between the scale label groups are first indicated by the interaction term when group 5 is the base.

Next, the gender dummy replaces age in the interaction term and another round (six regressions, once with each version as the base) is performed. Again, the female coefficient is informative about the relationship of gender and age in the base group and the interaction term indicates whether that relationship differs depending on question tone, reference period choice or scale labelling. The analysis of the first variable set is completed when each of the demographic and personality variables has been in the interaction term for a round of six regressions. Then, the whole process is repeated with the two other sets of variables – but interaction terms are only computed for the newly included variables (domain satisfaction in set II; trust, materialism, social standing and optimism in set III), not again for the controls.

Table 15 – Conditional correlations: regression 1 output.

Variable		**Coefficient**	**Standard Error**	**P-Value**
Version	1	(base)		
	2	0.24	(0.83)	0.77
	3	-0.31	(0.84)	0.72
	4	-0.18	(1.09)	0.87
	5	-0.78	(0.88)	0.38
	6	1.25	(0.87)	0.15
Age		-0.03	(0.02)	0.15

Version*Age	1 2 3 4 5 6	(base) -0.00 0.01 0.01 0.03 -0.03	 (0.03) (0.03) (0.04) (0.03) (0.03)	 0.94 0.78 0.79 0.28 0.29
Female		-0.22	(0.19)	0.26
Europe		0.07	(0.24)	0.78
Student		-0.25	(0.20)	0.22
Higher Education		-0.07	(0.21)	0.74
Partner		0.27	(0.19)	0.15
English Native		-0.06	(0.23)	0.79
Subjective Health		0.23	(0.11)	0.05
Household Income		0.01	(0.03)	0.80
Extraversion		0.18	(0.07)	0.01
Agreeableness		0.16	(0.10)	0.11
Conscientiousness		0.19	(0.08)	0.01
Openness		0.09	(0.09)	0.36
Emotional Stability		0.35	(0.07)	0.00
Constant		2.33	(0.98)	0.02

7.3. Results: additional output

Table 16 – Descriptive statistics: domain satisfaction and other variables (trust, etc.)

Variable	Version Statistic	All	**(1)** *Positive* (Tone)/ *Ref*	**(2)** *Balanced* (Tone)	**(3)** *Neutral* (Tone)	**(4)** *Negative* (Tone)	**(5)** *Complet.* (Label)/ *No Ref.*	**(6)** *Totally* (Label)
Domain Satisfaction								
Financial	Mean (SD) *min; max*	6.22 (2.34) *0;10*	5.92 (2.53) *0;10*	5.91 (2.10) *2;10*	6.40 (2.04) *1;10*	5.86 (2.45) *0;10*	6.85 (2.51) *0;10*	6.35 (2.31) *1;10*
Health	Mean (SD) *min; max*	7.42 (1.82) *2;10*	7.39 (1.67) *3;10*	7.30 (1.82) *2;10*	8.06 (1.74) *3;10*	7.07 (1.84) *2;10*	7.43 (2.00) *2;10*	7.37 (1.73) *3;10*
Achievement	Mean (SD) *min; max*	6.75 (1.98) *0;10*	6.56 (1.97) *2;10*	6.84 (1.49) *1;10*	7.06 (1.99) *0;10*	6.61 (1.88) *2;10*	6.69 (2.36) *0;10*	6.74 (2.11) *0;10*
Relationships	Mean (SD) *min; max*	6.97 (3.35) *0;10*	6.23 (2.53) *0;10*	7.37 (2.01) *1;10*	7.56 (1.85) *0;10*	6.86 (2.15) *1;10*	6.95 (2.78) *0;10*	6.87 (2.53) *0;10*
Work/Study	Mean (SD) *min; max*	6.67 (2.00) *0;10*	6.52 (1.93) *1;10*	6.86 (1.52) *3;10*	6.74 (1.87) *1;10*	6.61 (2.16) *0;10*	6.69 (2.37) *0;10*	6.57 (2.11) *0;10*
Safety	Mean (SD) *min; max*	8.11 (1.73) *0;10*	7.92 (2.11) *0;10*	8.04 (1.57) *4;10*	8.30 (1.58) *4;10*	7.95 (1.76) *2;10*	8.41 (1.50) *4;10*	8.06 (1.84) *3;10*
Environment	Mean (SD) *min; max*	7.49 (1.94) *0;10*	7.29 (1.91) *0;10*	7.68 (1.61) *3;10*	7.56 (1.63) *2;10*	7.26 (2.07) *0;10*	7.71 (2.11) *0;10*	7.41 (2.24) *0;10*

Other Variables								
Trust	Mean (SD) *min; max*	5.81 (2.25) 0;10	5.77 (1.99) *0;10*	6.07 (2.27) *0;10*	5.74 (2.26) 1;10	5.33 (2.60) 0;10	6.07 (2.12) 0;10	5.83 (2.22) 1;10
Optimism	Mean (SD) *min; max*	4.08 (0.52) *2.2;6*	4.13 (9.60) *2.2;5.7*	4.08 (0.50) *3;5.3*	3.95 (2.83) *2.8;5*	4.08 (0.54) *2.5;5*	4.03 (0.44) *3;5*	4.21 (0.51) *3.3;6*
Materialism	Mean (SD) *min; max*	2.81 (0.84) *1;4.8*	2.67 (0.67) *1.2;4.2*	2.99 (0.82) *1;4.7*	2.78 (0.88) *1;4.8*	2.78 (0.92) *1;4.7*	2.79 (0.82) *1;4.5*	2.82 (0.89) *1;4.8*
Social Standing	Mean (SD) *min; max*	6.85 (1.49) *0;10*	6.81 (1.12) *5;10*	6.74 (1.68) *0;10*	7.00 (1.21) *4;9*	6.61 (1.68) *1;10*	7.00 (1.60) *2;10*	6.98 (1.51) *0;9*

Table 17 – Differences in conditional correlations: demographics and personality

Version / Variable	signif. differ.	**(1)** Positive (Tone)/ Ref	**(2)** Balanced (Tone)	**(3)** Neutral (Tone)	**(4)** Negative (Tone)	**(5)** Complet. (Label)/ No Ref.	**(6)** Totally (Label)
Demographics							
Age		-0.03 (0.02)	-0.03 (0.02)	-0.02 (0.02)	-0.02 (0.04)	0.01 (0.03)	-0.06* (0.03)
Female		-0.66 (0.45)	-0.11 (0.43)	-0.53 (0.50)	0.06 (0.43)	-0.15 (0.43)	0.14 (0.45)
Europe		0.33 (0.54)	-0.32 (0.46)	0.11 (0.53)	-0.81 (0.50)	0.59 (0.45)	0.41 (0.50)
Student		0.51 (0.45)	-0.67 (0.43)	-0.53 (0.47)	-0.37 (0.43)	-0.45 (0.45)	0.08 (0.44)
Higher education		-0.08 (0.50)	0.37 (0.45)	-0.93 (0.50)	-0.44 (0.55)	0.53 (0.53)	-0,01 (0.49)
Partner		-0.03 (0.46)	0.45 (0.43)	0.04 (0.46)	0.02 (0.42)	0.77 (0.42)	0.36 (0.44)
English Native		-0.37 (0.51)	0.06 (0.47)	0.76 (0.53)	0.31 (0.45)	-0.47 (0.49)	-0.96 (0.50)
Subj. Health		0.34 (0.27)	0.36 (0.23)	0.16 (0.29)	0.03 (0.26)	-0.04 (0.29)	0.40 (0.29)
Ln Household Income (€/yr)		-0.06 (0.07)	0.01 (0.05)	-0.05 (0.08)	0.08 (0.06)	-0.02 (0.07)	0.03 (0.08)
Personality							
Extraversion		0.16 (0.18)	-0.14 (0.19)	0.10 (0.15)	0.28 (0.14)	0.12 (0.14)	0.43** (0.16)
Agreeableness		0.05 (0.21)	0.37 (0.23)	0.34 (0.25)	0.19 (0.20)	0.17 (0.20)	-0.24 (0.22)
Conscientious-ness		0.15 (0.19)	0.25 (0.17)	0.41* (0.21)	0.06 (0.18)	0.16 (0.14)	0.24 (0.16)
Openness		0.05 (0.23)	-0.05 (0.18)	0.25 (0.24)	0.26 (0.22)	0.06 (0.26)	-0.01 0.21
Emotional stability		0.34 (0.17)	0.15 (0.16)	0.46** (0.17)	0.30 (0.16)	0.53** (0.16)	0.38* (0.17)

Note: The coefficients show the (conditional) association of the listed variables with life satisfaction, per survey version. Significant differences of a coefficient from zero are indicated by asterisks. *p < 0.05, **p < 0.01. Standard error is given in parentheses.

Table 18 – Conditional Correlations: Other Variables (Trust, Materialism, Social Status, Optimism)

Version / Variable	signif. differ.	(1) Positive (Tone)/ Ref	(2) Balanced (Tone)	(3) Neutral (Tone)	(4) Negative (Tone)	(5) Complet. (Label)/ No Ref.	(6) Totally (Label)
Trust		0.11	0.10	0.15	0.12	0.30**	0.25**
		(0.11)	(0.09)	(0.10)	(0.08)	(0.10)	(0.09)
Materialism		-0.60	0.08	0.08	-0.38	-0.35	-0.03
		(0.31)	(0.25)	0.24	0.22	(0.24)	(0.23)
Social Status		0.45*	0.15	0.21	0.04	0.02	0.31*
		(0.19)	(0.12)	(0.18)	(0.12)	(0.13)	(0.14)
Optimism		0.62	0.66	-0.04	0.73	0.08	-0.29
		(0.35)	(0.40)	(0.42)	(0.38)	(0.46)	(0.40)
+ *Controls*							

Note: The coefficients show the (conditional) association of the listed variables with life satisfaction, per survey version. Significant differences of a coefficient from zero are indicated by asterisks. *p < 0.05, **p < 0.01. Standard error is given in parentheses. Control variables included are: age, female, Europe, student and personality (all five).