FLORIDA STATE UNIVERSITY

COLLEGE OF EDUCATION

CRITICAL ISSUES IN SURVEY META-ANALYSIS

By

AHMET SERHAT GOZUTOK

A Dissertation submitted to the
Department of Educational Psychology and Learning Systems
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2018

Ahmet Serhat Gozutok defended this dissertation on November 9, 2018.

The members of the supervisory committee were:

Betsy J. Becker
Professor Directing Dissertation

Fred W. Huffer
University Representative

Yanyun Yang
Committee Member

Insu Paek
Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.

This dissertation is dedicated to my father,

Mehmet Muhterem Gozutok

and to the memory of my uncle,

Abdullah Emin Canpolat

ACKNOWLEDGMENTS

First, I would like to express my deepest gratitude to my academic advisor, Dr. Betsy J. Becker. She has been my mentor, my role model, my counselor, and my friend. She has always been there when I needed her. She has always supported me, motivated me to do better, to be better. She made me a better researcher. Words cannot express how grateful and thankful I am for her endless support, encouragement, inspiration and guidance. Thank you so much for everything you have done for me, Dr. Becker.

I would like to thank my dissertation committee members, Dr. Yanyun Yang, Dr. Insu Paek, and Dr. Fred W. Huffer for their feedback and suggestions.

My friends and colleagues have also provided me with invaluable support, and guidance. I would like to acknowledge Neslihan Canpolat Cig, Dr. Oguzcan Cig, Dr. Salih Binici, and Zubeyde Binici. They have always been helpful and supportive. I also would like to express my appreciation to Dr. Abdullah Alghamdi whom I consider to be my colleague, my brother, and my best friend throughout this venture.

Finally, I would like to express my sincere gratitude to my family, my mother, Ayse Gozutok, my father, Mehmet Muhterem Gozutok, my uncle, Mustafa Canpolat, my brother and his wife, Omer Serdar Gozutok and Aynur Gozutok, and my little niece, Inci Gozutok for their unconditional love, and emotional support. My work would not have been possible without the support from my family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In research synthesis, researchers may aim at summarizing peoples' attitudes and perceptions of phenomena that have been assessed using different measures. Self-report rating scales are among the most commonly used measurement tools to quantify such latent constructs in education and psychology. However, self-report rating-scale questions measuring the same construct may differ from each other in many ways. Scale format, number of response options, wording of questions, and labeling of response option categories may vary across questions. Consequently, variations across the measures of the same construct bring about the issue of comparability of the results across the studies in meta-analytic investigations.

In this study, I examine the complexities of summarizing the results of different survey questions about the same construct in the meta-analytic fashion. More specifically, this study focuses on the practical problems that arise when combining survey items that differ from one another in the wording of question stems, numbers of response option categories, scale direction (i.e., unipolar and bipolar scales), response scale labeling (i.e., fully-labeled scales and endpoints-labeled scales), and response-option labeling (e.g., "extremely happy" - "completely happy" - "most happy", "pretty happy", "quite happy"- "moderately happy", and "not at all happy" - "least happy" - "most unhappy"). In addition, I propose practical solutions to handle the issues that arise due to such variations when conducting a meta-analysis. I discuss the implications of the proposed solutions from the perspective of meta-analysis. Examples are obtained from the collection of studies in the World Happiness Database (Veenhoven, 2006), which includes various single-item happiness measures.

CHAPTER 1

INTRODUCTION

In almost all scientific disciplines, researchers use survey-research techniques in investigating their research interests. To emphasize the common use of survey research, Saris and Gallhofer (2007) examined the uses of different data-collection methods (e.g., survey, experimental, observational designs) in five disciplines (economics, sociology, political science, psychology, and public opinion). They reviewed studies published in prestigious journals of each discipline in the 1994-1995 period. They concluded that data collection based on survey designs was the method used most by the studies in the fields of sociology, psychology, and public opinion. They also compared their findings with the findings by Presser (1984), who had reviewed studies published in the same journals between 1949 and 1980 for the same purpose. They noted that the applications of survey research increased over time most markedly in the area of social psychology.

1.1 How Survey Items Vary

The intense use of survey-based research in the studies of the social sciences brings about diversity in applications of surveys. Consequently, survey items on the same research topic differ from each other in various aspects such as survey design, data-collection procedures, characteristics of survey items and response scales. Saris and Gallhofer (2007) presented a comprehensive work that shows various ways of designing survey items. They pointed out that researchers have many choices in formulating requests for answers, given the same concept.

1.1.1 Wording of the Question Stem

For example, the wording of an item may prompt respondents to provide a direct answer to a question: “How satisfied are you with...?” Alternatively, the wording of an item may be

formulated by including a statement, and respondents are asked to judge the appropriateness of the statement: “Please indicate the extent to which you agree with the following statement: I am satisfied with…”

1.1.2 Response-Option Diversity

Variations are also seen in the design of response-option alternatives. Items may be formed as open-ended questions by asking respondents to provide an answer using their own words, or as closed-ended questions by asking respondents to choose an answer from a set of response choices (Krosnick & Presser, 2009). Response options of closed-ended questions come with corresponding response scales. Researchers choose among many response-scale types such as ranking scales, rating scales, graphical scales, and semantic-differential scales. The characteristics of those response scales can differ in many ways. Saris and Gallhofer (2007) listed several points that should be considered in examining response scales of survey items. Some of these points are the response-category format (e.g., yes/no scales, frequency scales, and Likert-type scales), number of response categories, labeling of category options, correspondence between the labels and numbers of the category options, symmetry of the labels, bipolar and unipolar scales, agreement between the concept and the scale, the use of a neutral or middle category, the use of “don’t know” options, the use of vague quantifiers (e.g., very, fairly, rarely) or numeric categories, the use of reference points, the use of fixed reference points, and the measurement level (i.e., nominal, ordinal, and continuous scales).

1.2 Integrating Survey Items

Each individual survey study on a particular research phenomenon with its own survey characteristics makes a unique contribution to that research area. Diversity across the different survey questions about the same construct can cause results of studies to differ from each other.

However, this heterogeneity provides researchers opportunities to understand possible reasons why results appear inconsistent across studies. Researchers may well be interested in integrating findings across such different survey-based studies of the same research concept. In order to summarize the findings of different survey items on the same topic, meta-analysts should ensure that the items measure the same construct, and that outcome measures of the items across the different survey studies are comparable to one another. Comparability across the items may be hampered by the fact that items tapping the same construct may differ from each other on many of the aspects mentioned above. Meta-analysis can be an appropriate statistical method for dealing with heterogeneity across findings of different studies, summarizing them together.

In this study, I examine how one can handle diversity across survey items in meta-analytic investigations. In particular, I focus on diversity across single item-survey questions based on self-report rating scales about the same construct: This diversity is due to the wording of question stems, numbers of response-option categories, scale direction, response-scale labeling, and response-option labeling. I further elaborate on the variation in survey items in the next section. Below I briefly introduce the idea of meta-analysis.

1.2.1 Meta-Analysis

As defined by Glass (1976), meta-analysis is "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (p. 3). Meta-analysis includes statistical methods to synthesize results across empirical-quantitative research studies, and to summarize these results for the purpose of reaching an overall understanding of a particular body of research. Rosenthal (1995) defined the concept more specifically: "Meta-analytic reviews are quantitative summaries of research domains that describe the typical strength of the effect of phenomenon, its variability, its statistical

significance, and the nature of the moderator variables from which one can predict the relative strength of the effect or phenomenon" (p. 183).

Lipsey and Wilson (2001) pointed out situations in which meta-analysis can be applied. They stated that meta-analysis is used to summarize findings of empirical research studies based on experimental, observational, and survey designs. Empirical research studies that are meta-analyzed should be based on quantitative data that can be summarized by descriptive and inferential statistics. They also stated that the studies being meta-analyzed should focus on the same constructs and should report findings in similar statistical forms.

1.2.2 Cooper's Steps of Research Synthesis

Cooper (2016) outlined a framework for conducting meta-analysis. His framework consists of seven steps. The steps are as follows: Formulating the problem, searching the literature, gathering information from studies, evaluating the quality of studies, analyzing and integrating the outcomes of studies, interpreting the evidence, and presenting the results. These steps are not mutually exclusive. In many circumstances, the decisions made in one step may affect the actions researchers would take in the other steps. Below I briefly touch on these steps. However, given the purpose of this study, I mostly focus on step 3, gathering information from studies.

1.2.2.1 Formulating the problem. Formulating the problem in meta-analysis is a core step, as the decisions made in this step serve as the basis for the further steps. In this step, researchers determine the concepts related to their research interests. Then, variables representing that particular concept are determined. Meta-analysts provide conceptual and operational definitions of these variables so that the relevant and irrelevant studies can be identified appropriately in the next stages of meta-analysis (e.g., searching the literature). In the

end, researchers want to clarify a specific research question(s) they will address, and the relevant research evidence that they need to search for in the primary studies.

For example, assume that one wants to integrate the findings of studies on happiness. First, the conceptual definitions of happiness should be delineated. What aspect of happiness is the focus of the research: hedonic level of affect, contentment, or overall evaluation of life? (Veenhoven, Ehrardt, Ho, & de Vries, 1993). Reviewers should ask themselves whether the conceptual definition of happiness in their meta-analysis covers similar concepts such as life satisfaction, and subjective wellbeing, and must assess whether the concepts are used interchangeably. Also, researchers should specify whether they are interested in simply describing the variable of interest (e.g., expressed happiness levels across nations) or examining the variable's relationship(s) with other variables (e.g., relationship between happiness and income). Clear specifications of these questions help the meta-analyst determine what kind(s) of measurement instruments and outcome measures they should seek in the primary studies to be included in meta-analysis (e.g., correlational outcomes, descriptive summary statistics).

1.2.2.2 Searching the literature. Once a research question is formulated, the next step is to locate the studies relevant to the specific research interest. Ultimately, researchers want to ensure that the study collection for their meta-analysis includes all previous studies that are relevant to the topic of interest. Meta-analysts have a variety of sources to use when searching for relevant studies. Lipsey and Wilson (2010) listed some of these sources as prior meta-analyses and review studies, reference lists in previous related studies, computerized bibliographic databases, bibliographic reference volumes, relevant journals, conference programs and proceedings, authors or experts in the area of interest, and government agency document collections. A good search is comprehensive (including all relevant searching sources), unbiased

(including both published and grey literature), and representative (including all relevant key words in electronic searches).

1.2.2.3 Gathering information from studies. The next step is to gather information from studies. In this step, meta-analysts construct a coding protocol that provides coders with what kind(s) of study features should be coded for the studies. Cooper (2016) presented an outline that lists typical types of information meta-analysts may want to gather from primary studies. The information is related to various aspects of research, including study report characteristics (e.g., author list, publication year, publication type), experimental conditions (e.g., intervention characteristics in detail such as intensity and duration of intervention, independent variables used, and control group conditions, if any), study settings (e.g., geographic and institutional characteristics of studies), sample characteristics (e.g., gender, ethnicity, age, socio-economic status, and education level), outcome variables (e.g., types of outcome, measurement scales of outcome, and validity and reliability evidences for outcome measures), types of research design (e.g., experimental randomized control trials, quasi-experimental designs, cross-sectional designs, and longitudinal designs), statistical outcomes and effect sizes (e.g., statistical tests used, type of effect size reported), and coder and coding characteristics (e.g., coder name, duration of coding time). The ultimate purpose of coding is to ensure that information gathered from studies is expressed in a systematic, organized, unbiased, and standardized way.

The coding process involves subjective judgements of coders. In order to ensure that the coders complete the coding process in the intended way, meta-analysts want to train coders about the coding protocol prior to the actual coding task. The purpose of training is to have coders comprehend the use of the coding protocol and meanings of the variables that are being coded. Trainings may include group activities such as practice with the coding protocol by working on a

small but representative sample of studies, and discussing the issues and discrepancies among coders that are identified during and after the practice session. In addition, to minimize risks in the coding process, meta-analysts may want to select some coders who are at least at the level of doctoral students, and who have strong familiarity with meta-analysis, social science methodology, and the research field that is meta-analyzed. (Lipsey &Wilson, 2001).

1.2.2.4 Evaluating the quality of studies. After the studies are collected, the next task is to evaluate the quality of the studies to determine which studies are kept and which ones are excluded from meta-analysis. In this stage, the studies that have passed through the literature search and coding process are taken under quality-check investigation to determine whether they meet the inclusion criteria. Each study is carefully examined to ensure that the research methods applied in the primary study are appropriate to the particular research purposes of the meta-analysis. Meta-analysts check whether the studies have included targeted study characteristics such as the intended population, appropriate research design, and sufficient statistics to extract effect sizes. Also, meta-analysts scrutinize whether unintended events occurred during the research implementation of the primary studies. In the end, the collection of studies is finalized to be analyzed in the further steps of meta-analysis.

1.2.2.5 Analyzing and integrating the outcomes of studies. The next step is to determine what statistical procedures meta-analysts should use to integrate the findings across studies. Meta-analysts determine the appropriate effect-size metric based on the research questions that need to be addressed. For example, if the purpose is to summarize the findings of studies that compare the effect of a particular intervention between comparison and control groups, typical effect-size metrics may be the index for the standardized mean difference (i.e., d, or g index). If the purpose is to describe a relationship between continuous independent variables

and continuous outcomes, the most typical, appropriate metric would be the Pearson's product-moment correlation coefficient (r-index). When both the independent variable and the outcome are dichotomous, then typical effect-size metrics would be the odds ratio and risk ratio. For detailed information about the effect-size metrics, see Borenstein, Hedges, Higgins, and Rothstein (2009). Regardless of what effect-size metric is used, meta-analysts should provide substantive interpretations of magnitude of effect sizes so that readers can have a clear understanding of the meaning of the findings.

In addition, meta-analysts should examine whether the magnitudes of effect sizes vary across studies (e.g., using the forest plot, *Q* statistic, or I^2-index). In case heterogeneity exists, potential moderator variables should be tested to see if any degree of heterogeneity in effect sizes across studies can be accounted for by these variables. Depending on the targets of analyses, meta-analysts conduct their analyses under particular statistical models such as the random-effects model, mixed-effects model, and fixed-effects model. For detailed information about the statistical models in meta-analysis, see Borenstein, Hedges, Higgins, and Rothstein (2009).

1.2.2.6 Interpreting the evidence. When it comes to interpreting the evidence, meta-analysts should discuss the results of meta-analysis in light of their generalizability and the limitations of the review. It is almost impossible not to face missing-data issues in meta-analysis. Some studies may meet all inclusion criteria; however, sufficient information to extract an effect size may not be found in the reports of those studies. There may be other studies that have not been published in any of the journals searched, so could not be included in the study collection. These problems may result in inaccurate estimations of the true effect size in the population. Meta-analysts want to address this potential missingness and publication bias. Cooper (2016) listed some of the techniques to assess these issues: the Rank Correlation Test (Begg &

Mazumdar, 1994), the Egger Test (Egger, Davey, Smith & Minder, 1997), the Funnel Plot Method (Macaskill, Walter, & Irwig, 2001), and the Trim-and-Fill Method (Duval & Tweedie, 2000). Each has its strengths and weaknesses. For those who are interested in details of the techniques, refer to Rothstein, Sutton, and Borenstein (2005).

Another point meta-analysts should discuss in interpreting evidence is whether the results would have been different if the analyses had been conducted using different statistical methods or under different data assumptions. This may be assessed by conducting a set of sensitivity analyses or by analyzing the data under different data assumptions. Meta-analysts want to ensure that the results stay consistent under different statistical conditions. If not, they should inform the readers and provide alternative interpretations about why inconsistency occurs.

1.2.2.7 Presenting the results. The final step of meta-analysis is to present the results in meta-analytic reports. Meta-analysts should have clear and organized documentation that contains the necessary information obtained from earlier steps. A few guidelines for reporting standards for meta-analysis exist in the literature, including the Meta-Analysis Reporting Standards, called MARS (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008). The MARS guidelines describe what information should be reported in each section of a meta-analytic report. Mainly, the report should clearly document problem statements and researcher questions, inclusion and exclusion criteria, search and coding procedures, results of statistical analyses and interpretations of effect sizes, discussions of generalizability of results, and limitations of the study.

CHAPTER 2

LITERATURE REVIEW

Latent constructs are often measured by self-report survey questions. However, such questions differ from each other in many ways, across different surveys. This hampers the comparability of responses obtained from different survey items even though they intend to target the same construct. In this chapter, I introduce certain issues that bring about heterogeneity in the survey questions on the same construct. I also provide a few examples from the literature that examine diversity in the survey questions caused by the number of response-option categories, scale direction, response-scale labeling, and response-option labeling.

Lietz (2010) conducted a review in the marketing area to investigate how different survey questionnaire designs impact responses. She provided researchers with a set of general recommendations for best practice in developing questionnaires. She stated that the number of response categories for Likert-type scales should be between of 5 and 8. In designing scales with numeric response categories, she suggested having unipolar scales with numerical labels at each point of the scale, and verbal labels at the endpoints only. Additionally, she pointed out that the labels “extremely” and “not at all” appeared to be the most effective verbal intensifiers. As for the verbal quantifiers, she recommended using specific labels such as “the number of times per a period” rather than vague ones such as “frequently” and “usually”. Despite the recommendations, it is apparent that the survey literature includes various types of questions tapping the same construct but differing in methodological factors. Consequently, one may not be able to directly compare the results of the survey questions across different studies.

2.1 Question Stem: Variations in Wording of the Question

To make questions from different surveys of the same construct comparable with one another, one should make sure that the questions target the same phenomenon of interest. However, questions are often worded in different ways. How do we determine if differences in wording across the questions alter the interpretation and understanding of respondents of the construct being targeted? If the differences are trivial and ignorable, then responses to non-identical but similar questions may be summarized together in a research synthesis. For example, in assessing overall life satisfaction, with "life as a whole", can we assume that the question in survey A "Taken all together, how satisfied are you with your life as a whole?" taps the same construct as the question in survey B "Considering all aspects of your life, how satisfied are you?" or as the question in survey C "Are you satisfied or dissatisfied with the general state of your life?" Given that the questions sound 'similar', can we treat them as the same question? In other cases, wordings of question stems may differ considerably across the questions. For example, in attempt to assess overall happiness in life, survey X asks "Generally speaking, how do you personally feel about your life?" while survey Y asks "Taking everything into consideration, do you consider yourself happy or unhappy?"

In concluding whether or not to summarize the results of such non-identical but similar questions together in research synthesis, one critical step is to evaluate the similarities in meaning and interpretation of question stems across surveys. In this process, researchers may consult with experts from research methodology in survey research, meta-analysts, and content experts. For example, "strength" level of the question statements may be rated by multiple coders (e.g., content experts, researchers, small groups of randomly chosen individuals, etc.) to see

whether or not the question statements sound the same in terms of difficulty. Differences, if any, may be modeled as a moderator/grouping variable in further analyses.

Even when the question stem is the same across questions, the formats of the response category options may vary in different ways. They can be numerical or verbal. Scale formats also may vary in terms of number of response-category options, scale direction (e.g., symmetry, polarity, existence of middle category or a "don't know" option), and response-option labeling (e.g., intensifiers anchored on response-category options). Below, each of these features is discussed.

2.2 Number of Response-Option Categories

Based on a comprehensive review, Eutsler and Land (2015) showed how frequently different numbers of response categories had been used. The review included 480 studies that had implemented various rating scales. The studies were published in seven top journals in the accounting literature between 2000 and 2014. Their report revealed that most of the studies used 7 point scales (34%); others used 11 point scales (22%), 5 point scales (13%), and 9 point scales (7%). They also stated that variance was maximized when the scales used had 7 points, and all response categories were labeled. This is consistent with earlier findings by Krosnick and Fabrigar (1997) who had suggested that seven is the optimal number of response categories for Likert scales. Krosnick and Fabrigar (1997) also added that using more than seven categories does not benefit much in terms of obtaining more information.

Kieruj and Moors (2010) examined the relationship between the number of response categories and respondent tendencies of choosing extreme-response and middle-response categories. Their findings indicated that the likelihood of choosing the extreme-response categories does not change across scales with different numbers of response categories (5 to 11

points). On the other hand, they also found that the middle-response categories were more likely to be chosen when the questionnaires were presented with relatively large numbers of response categories (i.e., 9, and 10 point scales).

Batz, Parrigon, and Tay (2016) examined the comparability of linearly transformed scores of single-item happiness and life-satisfaction rating scales (i.e., the 3, 4, 5, 7, 10, and 11-point scales obtained from the World Database of Happiness collection) after controlling for nation and year effects. They applied a linear transformation to rescale the rating scales with varying numbers of response categories to a reference 10-point continuum scale. They found that the mean scores transformed to the 10-point reference scale were significantly different from the original mean of the reference scale (i.e., grand mean). This result implied that the linear transformation method failed to fully solve the comparability issue. Transforming scale scores with more response categories on to scales with fewer response categories (i.e., moving an 11-point scale on to a 10-point scale) resulted in deflated scale means. In contrast, scale means were inflated when transforming scores from scales with fewer response categories on to scales with more response categories (i.e., 3, 4, 5, and 7- point scales on to a 10-point scale) (Batz et al., 2016).

To synthesize the findings from individual studies in a meta-analysis, one should ensure that outcome measures from the individual studies are on a common metric, so they are comparable with one another. Due to the fact that one questionnaire item may be presented with different numbers of response options across studies, the original metrics of individual studies are often different from each other. Thus, summary statistics obtained from such metrics may not be directly compared with one another (e.g., in terms of means, proportions, etc.). To enable comparisons, scale conversions may be applied to rescale individual outcome measures, and to

locate them on a common metric. Once the scale points across studies are on a common scale, then meta-analysts can consider summarizing the results from the scales across studies. Some scale transformation methods available in the literature are discussed later.

2.3 Scale Direction: Unipolar vs. Bipolar Scales

Polarity can be conceptualized in the context of two types of rating scales: Polarity in verbal scales and polarity in numerical scales. In verbal scales, polarity (i.e., unipolar and bipolar) takes place in verbal labeling of the response options of scales. A verbal scale is called bipolar if the response options of a 5-point scale are labeled symmetrically around the neutral midpoint. For example, the two sides of the midpoint represent two contrasting poles such as "Unhappiness" and "Happiness". The same intensifiers are used on two halves of a scale. An example of 5-point bipolar verbal scale response categories would be: "Extremely unhappy", "Moderately unhappy", "Neither unhappy nor happy", "Moderately happy" and "Extremely happy". On the other hand, unipolar verbal scales do not possess symmetry in response labeling (e.g., lack of unhappiness). The response option labels of unipolar verbal scales follow one directionality. For example, the options may run from "Least happy" to "Most Happy" or "Not at all happy to "Extremely happy".

In numerical scales (with no label anchoring), bipolarity exists if the number "0" is attached to midpoint of the scales and scale points on two opposite poles have the same numerical values with opposite signs (i.e., negative and positive). Thus, bipolar numerical scales hold symmetry and bi-directionality (assuming equidistance intervals between response options). For example, response option numbering for a 5-point bipolar numerical scale would be -2, -1, 0 +1, and +2. On the other hand, in unipolar numerical scales, the number "0" doesn't necessarily have to be attached to midpoint of the scale. Thus, symmetry can't take place unless verbal labels

are attached to numerical values. Typically, response option numbering for unipolar numerical scales runs from positive to positive (e.g., 1, 2, 3, 4, and 5 for a 5-point unipolar numerical scale).

Meta-analysts would be interested in combining the response options by their relative positions on the scales regardless of differences in numbering of the numerical scales and labeling of the verbal scales. The implications of such kinds of summarization are discussed below.

Schwarz, Knäuper, Hipple, Noelle-Neumann, and Clark (1991) indicated that alongside the use of appropriate category labels, values assigned to the category labels also play a crucial role in respondents' choice of a response-category option. Schwarz et al. (1991) found that respondents were more likely to choose lower point categories in an 11-category rating scale with points labeled 0-10 than in an 11-category scale with labels -5 to 5. Thirty-four percent of respondents chose categories between 0 and 5 in the 0-10 scale. When the -5 to 5 scale was presented, only 13% of the respondents chose the categories between -5 and 0.

In order to investigate the impact of directionality of response-category options on the responses to 11-point endpoints-labeled unipolar rating scales, Yan and Keusch (2015) conducted an experimental study with 496 participants who were randomly assigned to either of the two versions of a telephone survey differing in response-scale direction (i.e., 0-10 point, and 10-0 point rating scales). The respondents were asked to rate ten countries and locate them on the two versions of the 10 point development scales. Their overall findings indicated that for some countries, the average ratings appeared to be significantly lower in the 0-10 point scale as compared to the average ratings obtained from the 10-0 point scale.

2.4 Response Scale Labeling: Fully-Labeled vs. Endpoints-Labeled Response Categories

Fully-labeled scales are ones in which all response categories are verbally anchored with labels. In endpoints-labeled scales, however, verbal labels are attached to only the two endpoints. Fully-labeled 5-point scales, running from 1 to 5, may be endorsed by verbal labels such as "Extremely unhappy", "Unhappy", "Neither unhappy nor happy", "Happy" and "Extremely happy", respectively. In endpoints-labeled 5-point scales, only the end response categories 1 and 5 are labeled "Extremely unhappy", and "Extremely happy" respectively, while no labels are attached to the middle-response categories 2, 3, and 4.

Eutsler and Land (2015) reported the frequencies of types of scale labeling. They stated that 72% of the studies used the scales labeled at endpoints only, while only 5% of the studies had scales in which all response categories were labeled. They also conducted two experimental studies with a total of 767 college students. They examined the impacts of number of scale response categories and scale labeling on the distributional and statistical characteristics of the responses such as response patterns, variance, normality, power, and error. Their findings indicated that labeling all response categories resulted in increased power, minimized error, and maximized variance.

Weijters, Cabooter, and Schillewaert (2010) investigated the impacts of response-category labels on the responses to rating scales with different numbers of response categories. They administered eight different scale formats in an online survey of 1207 participants. The participants were randomly assigned to one of eight conditions differing in number of response categories (i.e., 4, 5, 6, 7 points) and response-category label style (i.e., fully labeled vs. endpoints labeled). They found that a tendency to agree, as compared to disagree, was more likely to occur regardless of content when scale-response categories were fully labeled. They

also reported that the participants were less likely to choose extreme-response categories in the fully-labeled scales than in the endpoints-labeled scales. They recommended meta-analysts to consider the scale-format characteristics as potential covariates.

Hamry and Levine (2016) also examined differences in response patterns to Likert-type scales differing in number of response categories (i.e., 4, 5, and 7 response categories) and in response-category labeling (i.e., only endpoints-labeled and fully-labeled scales). Their 882 participants were randomly assigned to answer survey questions with endpoints labeled or all points labeled. In both scale conditions, each question was presented in three different ways: with four, five, and seven point response categories. Hamry and Levine's findings were in agreement with the findings of Weijters et al. (2010).

Cabooter, Weijters, Geuens, and Vermeir (2016) examined the impacts of scale polarity and numbering of scale response categories on the respondent's choice on a 7-point agreement Likert-type scale labeled at the endpoints only. In their two-way design, 337 participants were assigned to one of four scale-format conditions: a unipolar scale with positive numbering, a unipolar scale with positive and negative numbering, a bipolar scale with positive numbering, or a bipolar scale with positive and negative numbering. Cabooter et al. (2016) found that bipolar scale formats elicited more responses toward the direction of "agree" as compared to the unipolar scale formats. The lowest ratings were given by the respondents in the unipolar positive numbering group. Their findings also indicated that extreme responses were chosen more frequently in the bipolar positive numbering group than the bipolar positive and negative numbering group. Cabooter et al. (2016) also recommended that researchers avoid comparing the results of rating scales with different scale formats, and they suggested examining the effects of the scale-format types as moderating variables or covariates in meta-analytic studies.

Kieruj and Moors (2014) further investigated if a tendency to choose endpoint response categories was induced by response-scale labeling and numbering (i.e., fully-labeled vs. endpoints-labeled response categories). Two scales, each including four items, were administered to 3266 respondents. The respondents were randomly assigned to one of the five scale-format conditions: Fully-labeled scales with/without unipolar numerical values (i.e., 1 to 7), endpoints-labeled scales with/without unipolar numerical values, and an endpoints-labeled scale with bipolar numerical values (i.e., -3 to 3). Their findings revealed that the respondents were more likely to choose the endpoint categories in endpoints-labeled scales as compared to fully-labeled scales. Also, they found that choosing endpoint categories appeared to be more frequent in bipolar numerical scales (-3 to 3) than in unipolar numerical scales (1 to 7).

Meta-analysts would be interested in combining response categories across the scales differing from each other in labeling format. Assuming that all other scale characteristics across the scales are kept constant, one may consider summarizing the results of these scales in a meta-analysis if the distribution of responses on response categories is not affected by scale-labeling format. The effect of labeling format with regards to summarizing results across scales in meta-analysis is discussed in this dissertation.

2.5 Response-Option Labeling: Slightly Different Response-Option Labels

Response-option labels may vary in many different ways across scales measuring the same concept. In one case (study (1)), a 7-point rating scale with responses running from 1 to 7 may be presented with response options labeled “Extremely unhappy”, “Very unhappy”, “Unhappy”, “Neither unhappy nor happy”, “Happy”, “Very happy”, and “Extremely happy”. In another study (2), the response option labels of the same question may be slightly different: “Completely unhappy”, “Pretty unhappy”, “Unhappy”, “Neither unhappy nor happy”, “Happy”,

"Pretty happy", and "Completely happy". The same question may also be presented in a way that some of the labels are in common, but their positions differ: "Very unhappy", "Unhappy", "Slightly unhappy", "Neither unhappy nor happy", "Slightly Happy", "Happy", and "Very happy", study (3). Many other variations of the combinations of response labeling and positions are also possible.

Jurges, Avendano and Mackenbach (2008) examined whether versions of a 5-point self-report heath survey differing in response-category labels were comparable to one another. They compared the World Health Organization version (WHO; 1 – 5 point scale, labeled as very good, good, fair, poor, very poor) and the US version (US; 1 – 5 point scale, labeled as excellent, very good, good, fair, poor). Based on the analyses of the survey data from 11643 elderly Europeans who responded to both versions, they found that the number of respondents who chose the most positive opinion in the WHO version (i.e., very good) was about twice the number who selected the most positive opinion in the US version (i.e., excellent). The response frequency for the negative endpoint of the US version (i.e., poor) was about four times larger than that of the WHO version of the scale (i.e., very poor). This study also showed that respondents were more likely to choose identically labeled categories across the versions (e.g., choosing 'very good' on both versions regardless of its relative position on scales) rather than identically positioned categories on the scales across the versions (e.g., choosing the 4^{th} category on both versions even though the labels of 4^{th} category are not the same across the scales).

When it comes to combining data from the two scales, Jurges et al. (2008) pointed out two options. One option is to obtain comparable scales consisting of four response categories by collapsing responses in the top two categories of the US version and the responses in the bottom two categories of the WHO version. The other option is to transform one scale to the other scale

(reference scale) by using the conditional probabilities of selecting each category of the reference scale.

To investigate labeling and positioning effects on response choices, Friedman, Cohen, and Amoo (2003) administered two versions of a 7-point response-category questionnaire to 175 college students who were randomly assigned to take one of the two versions. The versions differed from each other in some of the response-category labels and their relative positions on the scales: On a rating scale from 1 to 7, one had the categories "Completely Dissatisfied", "Very Dissatisfied", "Dissatisfied", "Neither Satisfied Nor Dissatisfied", "Satisfied", "Very Satisfied" and "Completely Satisfied" while the other consisted of "Completely Dissatisfied", "Dissatisfied", "Somewhat Dissatisfied", "Neither Satisfied Nor Dissatisfied", "Somewhat Satisfied", "Satisfied" and "Completely Satisfied", respectively. Testing the differences in the response-category frequency distributions and response patterns between the groups, they found that the response patterns were different between the scales, but the overall means of the distributions were the same, suggesting that the participants made their choices based on the response-category labels, not the relative position of the response categories. They also noted that the category "Satisfied" was more likely to be chosen in both versions of the questionnaires even though the position of the category was different on the two versions.

A few considerations should be made when one attempts to combine responses from diverse scales. Considering studies (1) and (2), slightly different intensifiers are used across the scales. If one assumes that the meanings of the intensifiers are the same to the respondents (e.g., "extremely" and "completely" are treated as the same), then the response categories across the scales could be summarized together. Now consider studies (1) and (3). There are two possible ways one may combine the response options. The first possibility is to combine the identical

labels across scales regardless of their positions (e.g., “Very Unhappy”, associated with the number 2 in (1), and “Very Unhappy” positioned at 1 in (3)). The second possibility is to combine identical positions regardless of their labels (e.g., “Extremely Unhappy”, positioned at 1 in (1), and “Very Unhappy” positioned at 1 in (3)). The implications of applying these options in summarizing the results of the scales across the studies in meta-analysis are discussed in this study.

2.6 The World Database of Happiness

In this study, I use examples of rating scales that are intended to measure the concept of “happiness”. At this point, I would like to introduce Veenhoven’s “The World Database of Happiness”, a focused archive that stores the descriptive and correlational summaries of research findings on subjective enjoyment of life. The database can be found at http://worlddatabaseofhappiness.eur.nl (Veenhoven, n.d.).

Veenhoven (2014) stated that overviews of up-to-date research findings in social science have become more challenging due to the dramatic increase in the number of publications in happiness research in recent years. This issue necessitates more focused techniques to search and gather available research findings in the early steps of meta-analysis. He stated that The World Database of Happiness archive is one such tool. It stores not only bibliographic references but also descriptive information on the findings of the studies of happiness across various nations and languages. The specific concept of happiness for this database is defined. Thus: “Overall happiness is the degree to which an individual judges the overall quality of his/her own life-as-a-whole favorably. In other words: how much one likes the life one leads.” (Veenhoven et al., 1993, p. 17).

The database consists of four main sources. The first source is a bibliography that contains detailed information about both published and unpublished research on happiness. Information includes but is not limited to subject matter, authors, keywords, place, and time. A directory of the authors is also provided. It serves as a tool to overview the happiness literature. The next sources are the collections of distributional and correlational findings of the studies. For particular nations, regions, or publics, the distributional findings present information about the survey characteristics, measurement instruments, sampling, frequency distributions of responses, transformed means, and standard deviations on a 0-10 common scale with confidence intervals. The correlational findings include information about the results of studies that focus on relationships between happiness and other factors that are hypothesized covariables such as income and age. The information relates to population, study design, scale measures, assessment, correlational statistics, and effect sizes.

The final source is the collection of measures. In this collection, the full texts of all eligible happiness measures are stored. They are classified based on some methodological characteristics such as conceptual focus of happiness, timeframe, survey mode, type of scale, and wording. (For details about measure eligibility procedures, see Veenhoven, 1984). For example, take a single item: 'Taken all together, how would you say things are these days? Would you say that you are…?' with three response categories: 3 Very Happy, 2 Pretty Happy, 1 Not too happy. The classification of this question is O-HL/c/sq/v/3/aa. O-HL corresponds to Overall-Happiness in Life as the focus of happiness, c corresponds to currently (presently, today) as the timeframe, sq corresponds to single question as the survey mode, v corresponds to verbal: each response option labeled as the scale type, 3 corresponds to three response categories as the scale range,

and aa is the corresponding item wording code for the phrasing of the question stem. All items in the database are classified based on the provided coding schema.

As of September 2018, the archive contained 1184 happiness measures varying in question wording and response-scale characteristics. The database of the measures can be very handy when it comes to refining rating scales in which information is standardized in a comparable way. By using search options in the database, researchers may be able to gather the studies in which the same survey items were used. Also, the database allows one to gather cases of rating scales that measure the same concept, but differ slightly from one another in question wording, response option labeling, and number of response option categories. Further, these items may be subject to being summarized together in a meta-analysis. For detailed information about the other aspects of the database, see Veenhoven (2014).

2.7 Scale Transformations

2.7.1 The Linear-Transformation Method: Linear Stretching

When it comes to comparing the responses to the same item with different response-option characteristics, one common issue is the differences in the number of response-category options across rating scales. A common method to handle this issue is to convert the scale points of the primary scales (individual scales) into a common secondary scale by applying the linear transformation. If the number of response options of a primary scale is smaller than that of the common scale, transformation is done by linearly stretching the scale points of the smaller scale on to the larger scale (e.g., moving a 5 point scale to 10 point scale). If the number of primary scale response options is larger than the number of the common scale-response options, the primary scale is linearly compressed into the boundaries of the common scale (10 point scale to 5 point scale). In order for verbal response scales to be linearly transformed on to a common scale,

the first operation is to assign numerical values to the verbal response categories (e.g., 1 for unhappy, 2 for neutral, 3 for happy). In this sense, linear transformations may be more appropriate to apply to numerical scales than verbal scales (DeJonge, Kalmijn, Veenhoven & Arends, 2015).

Linear transformations provide a way to transform scales with varying numbers of response categories to a common metric. The Linear Stretching method is commonly used to transform 5 or 7-point scales on a common scale with wider scale length, such as 0 to 10-point scale (DeJonge, Veenhoven & Arends 2015). Linear stretching projects the endpoints of the primary scale (i.e., 1 and 5 for 5 point scale) onto the endpoints of the common scale (i.e., 0 and 10 for a 10 point scale). The middle category points of a primary scale (2, 3, and 4, with interval distance of 1) are projected onto the middle points of the common scale in a way that the distances between categories stay equivalent to each other (2.5, 5, and 7.5, with interval distance of 2.5). Consequently, the outputs of original scales differing in the number of response categories are brought to a comparable level.

However, the linear transformation method is problematic for at least two reasons. The first is the assumption of equidistance between response categories; for a scale, the interval distances between successive response option categories are assumed to be equal to one another. It implies, for example, that the difference between the "very happy" and "extremely happy" categories is the same as the difference between the "neither happy nor unhappy" and "happy" categories. This assumption rarely holds, especially for fully-labeled scales with more response options. The second reason is that both stretching and compressing of primary scales ignore the existence of verbal response-option labels. Assume that a 3-point scale with categories '1-unhappy, 2-neither happy nor unhappy, 3- happy' is being stretched to 5-point scale '1-

extremely unhappy, 2- unhappy, 3-neither happy nor unhappy, 4- happy, and 5-extremely happy'. When stretched, the values 1, 2, and 3 would convert to 1, 3, and 5, respectively. Assume a respondent who expressed his/her feeling as happy had chosen 3 on the old scale. The response option 3 corresponds to response option 5, which is extremely happy on the new scale. Consequently, linear transformations may result in incorrect categorizations of respondents due to differences in the labeling of response options across scales.

2.7.2 The Nonlinear-Transformation Method: Semantic Judgement of Fixed Word Value

To get around the issues caused by the equidistance assumption and ignorance of response-option labels, the Semantic Judgement of Fixed Word Value approach has been proposed. The approach is based on using subjective judgments of individuals to determine corresponding values for possible response-option labels on a secondary numerical scale. A group of individuals (e.g., experts on the research context, representative samples from the population of interest, or random respondents) are asked to place each label on a common metric that is bounded by predetermined values (e.g., the lowest and highest response options are numbered with 0 and 10, respectively). One of the early applications of this approach was done by Jones and Thurstone (1955). They asked 905 respondents to rate 51 phrases that are used to indicate like-dislike status for food (e.g., strongly like, tasty, bad). Each phrase was presented independently with a 9-point equidistant numerical scale running from -4 to 4, on which endpoints were labeled with 'Greatest Dislike' and 'Greatest Like' and the midpoint (0) was labeled with 'Neither Like Nor Dislike'. The averages of the individual ratings for each phrase determined the values of the phrases on the common scale. Consequently, all the phrases were placed on a common 9-point interval scale.

Veenhoven et al. (1993) made use of the Jones and Thurstone (1955) approach in his initial version of the World Database of Happiness. Ten content experts were asked to rate 29 various response-option labels that appeared in nine happiness items (e.g., quite happy, fairly happy, and very happy). The question stems of the items were almost identical, but items differed from one another in the number of response options and response-option labeling. The experts judged the response-option labels and assigned a value for each of them on a secondary scale. The averages of the expert ratings were assigned as the values for the response labels on the common scale, a continuous scale running from 0 to 10. Unlike Jones and Thurstone (1955), in which integer values were assigned to the labels (i.e., -4 to 4), the judges in Veenhoven's study were allowed to assign any integer or non-integer number between 0 and 10 to the response-option labels.

The main shortcoming of the Semantic Value of Fixed Word Value approach is that experts rate the response options irrespective of the other response-option labels of the original scale and its relative position on the common scale that contains all other response-option labels. For example, when the response label 'very happy' is placed as the endpoint of a scale, one would rate this label to be on the very end of the secondary scale, such as at 10. On the other hand, if 'very happy' is placed on a scale in which the endpoint category label is 'extraordinarily happy', one would be likely to assign 'very happy' a lower value than 10, considering that 10 should be assigned to a stronger response category of that scale (i.e., extraordinarily happy).

In addition to the issues above, this approach does not take into account the differences in the wordings of the question stem and the numbers of response options of the primary scales. The differences in question stems may impact the way that judges interpret the response-option labels being rated. Also, semantic intensity of the same response-option label may be interpreted

differently depending on whether it is presented in a long scale (e.g., with 7 categories) or a short scale (e.g., with 3 categories).

2.7.3 The Nonlinear-Transformation Method: Semantic Judgement of Word Value in Context

The Semantic Judgement of Word Value in Context approach has judges assign values to response-option labels by taking into account the other aspects of the original scale. These include the relative position of the label, other response-option labels, the number of response-option categories, and the wording of the question stem. The scale interval study by Veenhoven (2009) may be considered under this approach; it is described below.

2.7.3.1 Happiness scale interval study. To overcome some of the issues of earlier transformation methods described above, Veenhoven (2009) initiated the study called "Happiness Scale Interval Study". The study aimed to have judges evaluate the semantics of response-option labels by considering the wording of question stem and other response-option labels of the corresponding scale. The study intended to rate all possible response labels that have been used in happiness surveys across many nations. The judges were chosen from a college-student population. When asked to rate the degree of happiness denoted by the response labels, the original item and its scale with response options were also presented to the judges. In addition, this method differed from the earlier methods in that response options were treated as intervals on a 0-10 continuum scale, rather than discrete points, which had been the case in the other methods. Judges were asked to provide interval boundaries on a range of 0 to 10 for each label considering the labels and positions of other response categories on the original scale. The judges were instructed to divide a 0-10 scale into *k* intervals by assigning *k*-1 threshold values, where *k* is the number of response-category options. The threshold values become the cutoff

points where a response label transits from its adjacent response labels. The middle of the lower and upper subinterval boundaries (mid-interval value) represented the value for each label. Consequently, the means of the mid-interval values across judges for the various response labels appearing in different questions were placed on a 0-10 numerical scale.

To assign values of happiness based on verbal response options, Veenhoven and Hermus (2006) developed a web-based tool called the "Scale Interval Recorder". The recorder was used in the scale interval study. The screen of the recorder consists of two parts. On the left side, the question stem and its corresponding response-option labels are presented. On the right side, a vertical scale, running from 0 to 10 with an interval of 1 between numbers, is shown. The two ends of the scale are labeled as "best possible" and "worst possible". The numbers are located on the left side of the scale, and the response options for the question are located on the right side of the scale. For *k* response options, there are *k*-1 slide bars on the scale that can be moved up and down using the cursor. The slides are used to divide the scale into *k* intervals. The response options move simultaneously with the slides to the midpoint values of the intervals. Judges are instructed as follows: "Shift the separation lines until you feel that the intervals correspond with the degree of happiness denoted by the words on the right".

Below, I illustrate an example of this approach using the trial version of the happiness scale interval recorder, which can be found at the website of Veenhoven's database (https://worlddatabaseofhappiness.eur.nl/scalestudy/scale_fp.htm). In this example, the scale is partitioned into five intervals as there were five response options for the question "How happy do you feel as you live now?". Each label was judged to denote an interval. The mid-interval values are assigned to represent the labels. For example, 'very happy', which was the most positive label, is assigned to the interval from 9 to 10. So, its mid-interval value is 9.5. The

boundaries of ‘fairly happy’ are 6 and 9 with mid-interval value of 7.5. ‘Neither happy nor unhappy’ is within the interval of 4 to 6, its mid-interval value is 5. ‘Fairly unhappy’ is within the interval of 1 to 4, its mid-interval value is 2.5. ‘Very unhappy is within the interval of 0 to 1, its mid-interval value is 0.5.

DeJonge, Veenhoven, and Arends (2014) compared the three transformation methods described above. They obtained the data on two slightly different survey questions on happiness from the Happiness Scale Interval Study. The survey questions were administered in the same nation and the same year. The first question was “To what extent are you satisfied with the life you currently lead?” with five response options (i.e., extraordinarily satisfied, very satisfied, satisfied, fairly satisfied, not very satisfied). The other question was “On the whole, how satisfied are you with the life you lead?” with four response options (i.e., very satisfied, fairly satisfied, not very satisfied, not at all satisfied). DeJonge et al. (2014) compared transformed values, obtained from different transformation methods, on a common metric running from 0 to 10. Transforming the scales by the Linear-Stretching method resulted in the ‘fairly satisfied’ option of the first scale corresponding to 2.5, while the ‘fairly satisfied’ option of the second scale corresponded to 6.7. By the Jones-Thurstone transformation, in which the average of ratings by judges is assigned to each label irrespective of the question context, ‘fairly satisfied’ got a 6.3, and the value was fixed for both scales. When the Semantic Judgment of Word Value in Context method was applied, ‘fairly satisfied’ got 4.5 for the first scale, whereas this response option in the second scale transformed to 6.6. The results suggested that wording of question stem and labeling of response options impact judges’ ratings of the response options.

The purpose of this study is to shed light on how best to deal with diversity across survey questions on the same construct in meta-analysis. The diversity across questions may occur for

any number of reasons. I limit my study to a few of them: wording of question stem, number of response-option categories, scale direction (i.e., unipolar and bipolar scales), response-scale labeling (i.e., fully-labeled and endpoints-labeled scales), and response-option labeling. I apply three transformation methods (i.e., Linear Stretching, Semantic Judgement of Fixed Word Value, and Semantic Judgement of Word Value in Context) to scale different but similar questions of the same construct on to a common metric. Then I examine the results of some example meta-analyses conducted based on the three transformation methods.

CHAPTER 3

METHODOLOGY

Given the purpose of the study, my research questions are as follows:

1. How can meta-analysts deal with survey questions on the same construct that differ slightly in wording of question stems, number of response-option categories, scale direction (i.e., unipolar versus bipolar scales), response-scale labeling (i.e., fully-labeled and endpoints-labeled scales), and response-option labeling?
2. Do the results of meta-analyses differ from each other when transforming scales of primary studies to a common scale by applying three different scale transformation methods: The Linear-Stretching, the Semantic Judgement of Fixed Word Value, and the Semantic Judgement of Word Value in Context approaches?

I deal with two main issues that arise because of wording of survey questions and response options. First, I discuss diversity in question stems, then I follow this up by discussing diversity in scale-response options and proposing the applications of the transformation methods in the context of meta-analysis. Having said that, I do not have any intention to claim whether one transformation method performs better or worse than the other methods. Nor is my goal to meta-analyze research studies of happiness. My aim is to provide practical solutions for meta-analysts who are in situations where they want to summarize the findings of sets of survey items across studies that differ in the aforementioned characteristics. Surveys may differ in many other aspects of research methodology such as sample characteristics, survey sampling methods (e.g., random, stratified, and cluster), survey sample weighting, and psychometric properties of the entire survey (e.g., validity, and reliability). Although beyond the scope of this study, researchers are advised to consider these points when conducting a meta-analysis based on survey data.

3.1 Diversity across Question Stems due to Variations in Wording of Question Stems

In the context of a meta-analysis, the literature search may yield many studies that meet the meta-analyst's inclusion criteria, such as tapping the construct being measured, examining the target population, using desired sampling procedures, and being produced during a specific range of years. However, unless the same survey is administered across studies, it is likely that the survey questions will differ from each other across studies, even though the studies' instruments measure the same construct. These differences might be due to variation in wording of the question stems of the survey questions. In order to integrate the findings of these different survey questions in a meta-analysis, meta-analysts must decide whether the wordings of the survey questions tap the same, intended construct. This decision should be made very carefully, as mistakes occurring in this step may result in summarizing findings of survey questions that should not be combined in the first place (i.e., combining apples and oranges).

Let's assume that the meta-analyst ends up having ten studies that measure exactly the same construct after a comprehensive and unbiased literature search. The studies have very similar survey questions. However, the wordings of the questions differ slightly from one another. The meta-analyst has already decided that these questions are similar enough and passed the eligibility criteria for being included in the study collection. Therefore, he/she may ignore these slight differences in the wording of question stems. Alternately, the differences can be considered in further analyses. In this study, I propose that variations in the wording of question stems across survey studies be accounted for by having coders rate the strength of question wordings. The purpose of doing so is not to determine whether the questions measure the same construct; that has been decided at the study-selection phase. The purpose is more exploratory: it is to see whether any amount of heterogeneity across the findings of studies (if it exists) may be

explained by the differences in ratings of the strength of the question-stem wordings across questions. This information (and other wording differences) may be modeled as a moderator variable in a meta-regression or other analysis (e.g., via an ANOVA-like model). In my analyses I focused on strength of stem wording. The questions were obtained from the World Database of Happiness.

Fourteen self-report single-item survey questions that measure the concepts of "overall happiness in life" or "happy person" were obtained from the studies in the "World Database of Happiness" collection. The wordings of the questions in the database were very similar to each other. In order to have more variation in strength of the question stems, I made up four questions in which the wordings of the question stems were relatively different from the wordings of the fourteen questions from the database. More information about the questions and their characteristics is provided later in this Chapter.

To obtain ratings of the strength of the question stems, a set of four coders were presented with eighteen slightly different survey items, differing in question wording and response options, from survey studies that were selected to be meta-analyzed. All question stems were presented on a computer screen. The questions and response scales were developed on the Qualtrics online survey software platform. The Qualtrics software is used to develop, distribute, and analyze surveys electronically. The features of the software allow one to customize many aspects of surveys such as the format of the questions and response-scale options, the appearance of surveys and more. The construction of the coding mechanism in Qualtrics is very similar to the typical survey-development process.

In a typical meta-analysis, study coders should be selected (or trained) to have particular competencies given the context of the meta-analytic study. In general, coders would be expected

to be knowledgeable about the content of research that is meta-analyzed, and key literature on the construct of interest. They also need to have a clear understanding of the goals of meta-analysis so that they can identify what to extract and code from primary studies. It is typical practice to train coders before the coding process begins. For purpose of demonstrating how this works in the context of this study, I prepared instructions for coding (rating in this case) the particular aspects of the survey questions that are described next.

In this study, four individuals were recruited to serve as coders. To avoid possible misunderstandings of the terminology used in the question stems, the coders were proficient in the English Language, being native English speakers. Participation was on a voluntary basis. I used my personal contacts to reach out to individuals who might be interested in participating in the rating task. The coders included two individuals who had doctoral degrees; one was in education and the other was in music. Two others were doctoral students; both in the department of Educational Psychology and Learning Systems at Florida State University. Because the concept of happiness is a very familiar one, no special qualifications were required of the coders.

The coders were asked to rate the strength of wordings of the question stems based on their personal judgement. The meaning of "strength" will likely vary depending on the content of the research summarized in any given meta-analysis. In this case, the strength of question wording refers to the intensity of degree of happiness. The instructions about how to perform the rating task were as follows:

In this task, you will be presented with eighteen similar survey questions about happiness that differ slightly from each other in the wording of their question stems. The purpose of rating the wordings of the question stems across different questions is to determine whether the strengths of the question stems sound similar. If not, we want to quantify the differences.

Your task is to rate the strength of the wording of the question stem for each question, based on your personal judgment. The term "strength" refers to "intensity of degree of happiness". All question stems are listed on a computer screen together. A dropdown list is located next to each question stem.

Please rate the strength of each question stem by choosing the rating 1, 2, or 3 in the dropdown list: A 1 corresponds to a weak statement, 2 corresponds to a moderate statement, and 3 corresponds to a strong statement about happiness.

The screenshot of the Qualtrics survey that includes the rating task for determining strength of the question stems is presented in Figure 3.1.

3.2 Diversity across Response Scales

Response scales differ from each other in numerous ways across different survey questions. In this dissertation, I focus on the rating scales of different survey questions measuring the same construct that differ in their number of response option categories, scale direction (i.e., unipolar and bipolar scales), response scale labeling (i.e., fully-labeled and endpoints-labeled scales), and response-option labeling. I use three scale-transformation methods to demonstrate how meta-analysts can handle such diversity. Next, I describe how each transformation method is used to scale responses of different survey questions onto a common scale running from 0 to 10.

3.2.1 The Linear-Transformation Method

The scale ratings of primary scales with varying response-option categories can be linearly transformed to a common 0-10 interval scale. The lower and upper end-point ratings of primary scales are directly transformed to the values 0 and 10, respectively. The other ratings are proportionately stretched so that the distances between adjacent response categories are equal.

The verbal labels of the response categories of the primary scale are assigned to the transformed response-category values on the common 0-to-10 scale. In cases where response-option labels of the primary response scales are not denoted by numerical values, successive numbers are assigned to the ordered response-category options to serve as the original values. I used integers, starting with a 1 assigned to the negative end of the scale in terms of the meaning of the construct (e.g., least happy).

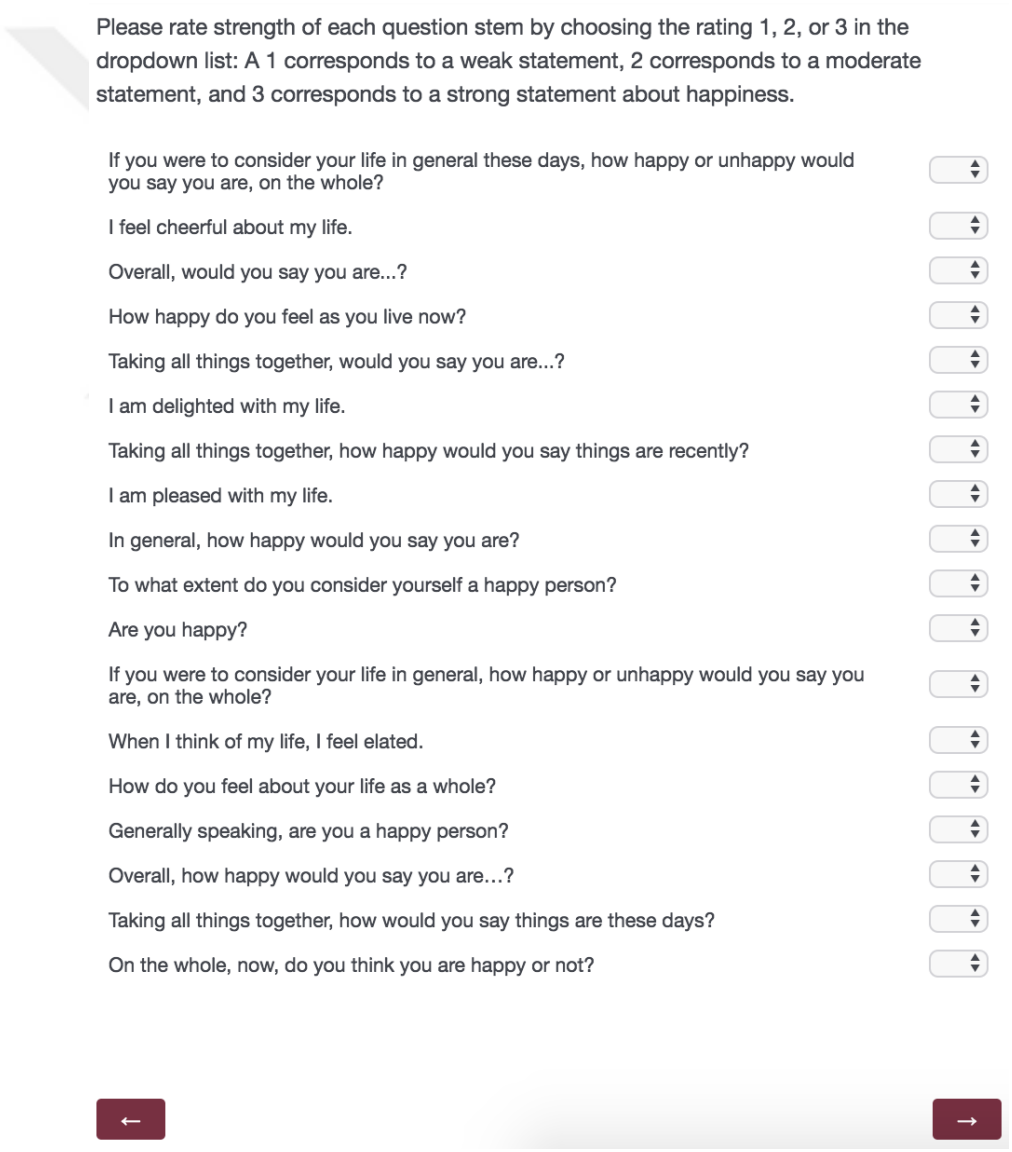


Figure 3.1. Screenshot of the rating task for determining strength of the question stems in Qualtrics.

The generic formula for the linear-transformation method to transform ratings from one scale to another is given in Equation 1:

$$X_{transf} = \left[\left(X_{orig} - min_{orig}\right) * \frac{max_{transf} - min_{transf}}{max_{orig} - min_{orig}}\right] + min_{transf}, \quad (1)$$

where X_{orig} is the original rating of the primary scale, X_{transf} is the transformed value of the original rating of the primary scale on the secondary scale, min_{orig} and max_{orig} are the minimum and maximum possible ratings of the primary scale, respectively, min_{transf} and max_{transf} are the minimum and maximum possible ratings of the secondary scale, respectively (Card, 2011). Here, min_{transf} is 0 and max_{transf} is 10.

The linear-transformation method is applied to the response scales of each primary study that is selected to be meta-analyzed. By doing so, all response scales with varying numbers of response-option categories are scaled to a common metric running from 0 to 10. Below I illustrate a hypothetical example of how to apply the linear transformation.

Let's assume that our study collection for analysis includes three survey studies. The studies use similar survey items, but the items differ slightly in question wording, numbers of response-option categories, and response labeling. The questions are:

Q1-In general, how happy would you say you are these days?

Original ratings	1	2	3	4	5	6	7
Response-option labels	Not happy at all	Very unhappy	Somewhat unhappy	Neither happy nor unhappy	Pretty happy	Very happy	Extremely Happy
Linearly transformed ratings	0	1.67	3.33	5	6.67	8.33	10

Q2-How happy do you feel as you live now?

Original ratings	1	2	3	4
Response-option labels	Very unhappy	Not too happy	Pretty happy	Very happy
Linearly transformed ratings	0	3.33	6.67	10

Q3-In general, how do you feel in terms of happiness in your current life?

Original ratings	1	2	3	4	5
Response-option labels	Very unhappy	Somewhat unhappy	Neither happy nor unhappy	Somewhat happy	Very happy
Linearly transformed ratings	0	2.5	5	7.5	10

In order to combine the results of the response scales, we want to locate them on the same metric. Simply, I apply the Linear-Stretching method to transform the three response scales with varying numbers of response options to a common scale that runs from 0 to 10. After applying the linear-transformation method to each response scale, the transformed scores of the primary-scale ratings denote the response-option values on the common scale. For instance, on the common scale, the rating value for ‘Not happy at all’ is 0, it is 3.33 for ‘Not too happy’, and 10 for ‘Extremely happy’.

In some cases, the same response-option label may be used in multiple scales. If the verbal label is attached with different numerical labels across multiple scales, simple transformation can lead to inconsistencies in the numerical labels attached to the same verbal label on the transformed scale. In situations where the same response-option label appears in multiple scales, I average the transformed ratings of those response-option labels across the scales and use the average value to represent those labels on the common scale. For example, the label ‘Very unhappy’ appears on all three scales. Its corresponding value on the common scale is calculated by dividing the sum of the ratings on the transformed primary scales (1.67 + 0 + 0) by

the total number of primary scales in which the label appears (3), to yield 0.56. For the label "Very happy", the average transformed value is 9.4, and for "Somewhat unhappy", the average transformed value is 2.92.

Once the original ratings for each response-option label were linearly transformed to the 0-10 scale, the means ($\hat{\mu}$) and variances (s^2) were calculated for each transformed scale by entering the proportions of respondents in each of the response-option categories ($P(x)$) and the transformed ratings (x) in the following equations:

$$\hat{\mu} = \mathrm{E}(X) = \sum x * P(x), \tag{2}$$

$$s^2 = \sum (x - \hat{\mu})^2 * P(x). \tag{3}$$

Alternatively, one may calculate the linearly-transformed mean and standard deviation directly from the mean and standard deviation of the original scale (Kalmijn, 2010). Specifically,

$$\hat{\mu}_y = \left[(\hat{\mu}_x - min_x) * \left(\frac{max_y - min_y}{max_x - min_x}\right)\right] + min_y, \tag{4}$$

$$s_y = s_x * \left(\frac{\max(y) - \min(y)}{\max(x) - \min(x)}\right). \tag{5}$$

This type of linear transformation may be of use when the response-category labels of some scales are not reported. However, unlike the linear-stretching method described earlier, where the transformed ratings of the response-option labels were averaged across scales, Kalmijn's way of doing the linear transformation does not consider whether the same response-option labels are used across multiple scales.

In general, the Linear-Stretching method has a few drawbacks, especially when applied to transform verbal-response scales to a new scale. One of them is the assumption of equidistance between response categories of both the primary scale and the secondary scale. The other issue is

that the transformation does not take into account verbal anchoring of response options and differences in question stems.

3.2.2 Nonlinear Transformations

Along with the linear-transformation method, two nonlinear, verbal transformation methods are applied in this study. These are the Semantic Judgement of Fixed Word Value and Semantic Judgement of Word Value in Context methods. The main ideas of the methods have been introduced in the previous chapter. Both methods involve personal judgements made by judges. Coders are asked to judge the degree of happiness denoted by the response option labels used in these survey items. Next, I describe how these methods can be used in meta-analysis.

3.2.2.1 The Semantic Judgement of Fixed Word Value method. In this method, one generic happiness question and the response options from all other questions were presented together to the coders. For example, take the question "How happy are you in general?" as the generic question. The response options taken from all individual questions were listed under the generic question (e.g., extremely happy, very happy, pretty happy, not too happy, somewhat unhappy, fairly unhappy, not at all happy, and so on). A horizontal, continuous scale was located under each response option. The scale ran from 0 to 10. The scale had a slider bar that could be moved left or right using the cursor. The position of the slider bar on the scale determined the value (any value between 1 and 10 with up to 1 decimal place) assigned to the corresponding response-option label. Coders were asked to shift the slider bar for each label until they felt that the position of the slider bar represented the degree of happiness denoted by the label. Once the input from all coders was obtained, the average values across the set of coders were treated as the final values for the response options across all studies and items. Based on the final values for

the response options, the means and standard deviations on the 0-10 scale were calculated for each question.

Still, this method does not take into account differences in the wordings of question stems. Also, the response-option labels are judged independent of the context of the original question, other response-option labels of that question, and position of the label with respect to the other response options. The screenshot of the preview for the Semantic Judgement of Fixed Word Value method is presented below (Figure 3.2).

3.2.2.2 The Semantic Judgement of Word Value in Context method. In the Semantic Judgement of Word Value in Context method, each individual question and its corresponding response-option labels were presented to the coders on a separate screen page. The questions with endpoints-labeled scales were presented after all questions with fully-labeled scales were presented. A horizontal, continuous scale running from 0 to 10 was located on the bottom of each response option. For example, take the question "How would you say you are feeling these days?" with the following response options: not at all happy, not too happy, pretty happy, and extremely happy. For each particular response option, the coders assigned a position and the associated number that they considered the most appropriate for that response option on the 0-10 scale. Judge A might assign the position of 1 for not at all happy, the position of 3 for not too happy, the position of 6 for pretty happy, and the position of 9 for extremely happy. Another coder could rate the same labels differently. Also, the same coder might assign different positions to the same response-option label when it is presented in the contexts of different questions (e.g., with different question stems and response-option labels). For the endpoints-labeled scales, intermediate response categories were labeled as "NLA", for "no label attached".

Coders were asked to assign positions to the intermediate response categories as they did for the labeled endpoints.

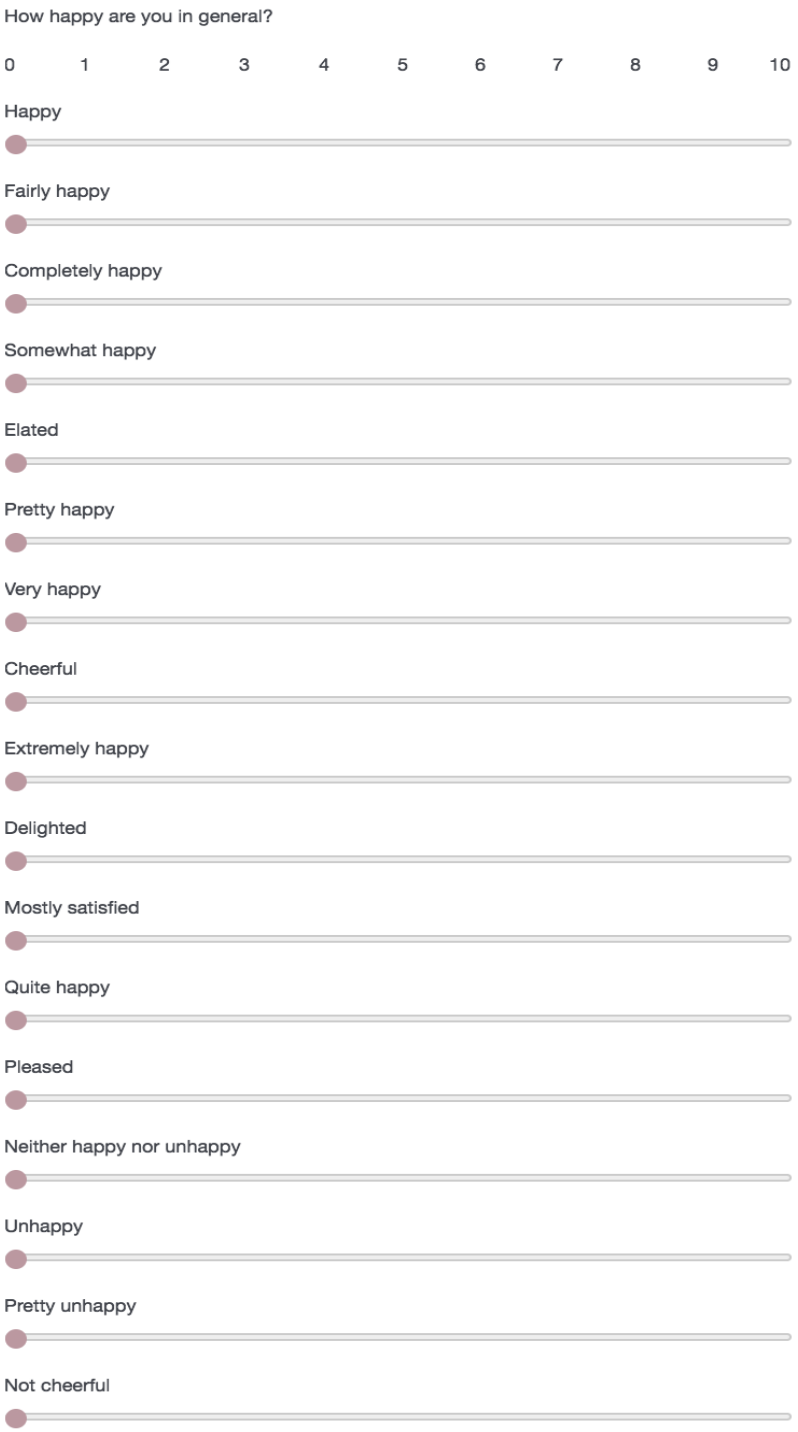


Figure 3.2. The screenshot of the preview for the Semantic Judgement of Fixed Word Value method.

For each question, I averaged all values assigned by the four coders to the same response-option label, and treated the average value as the final value for that response option label on the 0-10 scale. Based on the final values for the response options, the means and standard deviations of the questions were calculated for each question. A screenshot for the rating task for one question is presented below (Figure 3.3).

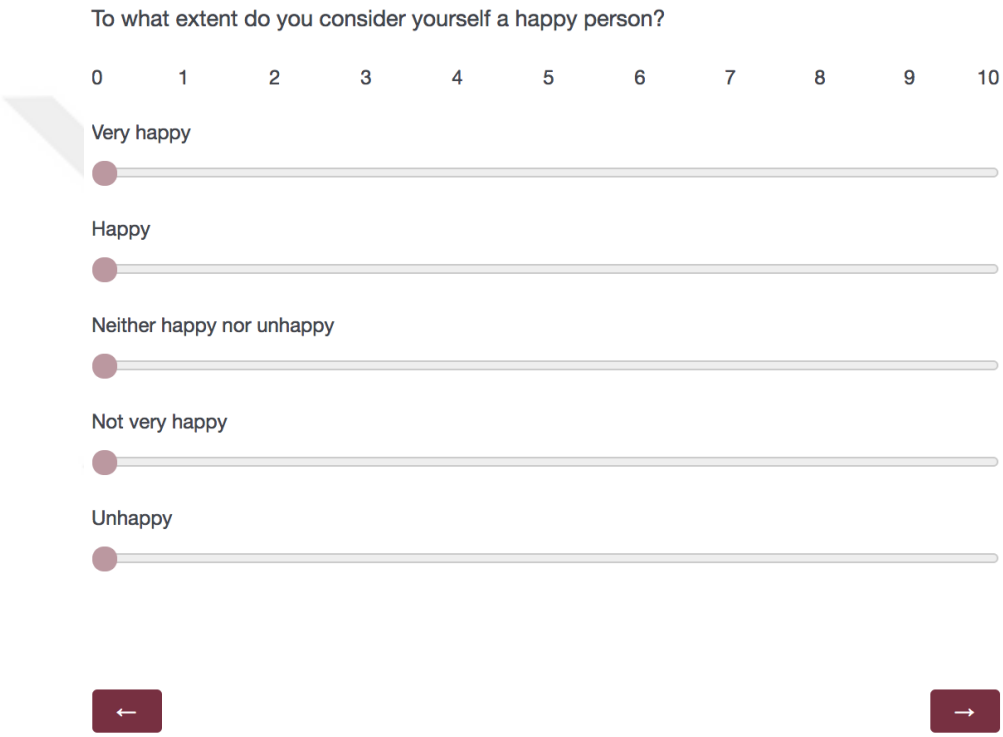


Figure 3.3. Screenshot of one question for the Semantic Judgement of Word Value in Context method.

3.3 A Case Study

In order to illustrate how the transformation methods perform in practice, I also conducted a hypothetical meta-analysis. Eighteen questions about happiness were included the study collection of the meta-analysis. As mentioned, four questions were made-up questions, and fourteen questions were obtained from the happiness database. Table 3.1 presents the question stems and the response option labels for each question.

Table 3.1

The Survey Questions and Response-Option Labels

Question ID	Question Stem	Response-Option Categories						
		1	2	3	4	5	6	7
q01	Taking all things together, would you say you are...?	Not at all happy	Not very happy	Quite happy	Very happy			
q02	If you were to consider your life in general these days, how happy or unhappy would you say you are, on the whole?	Not at all happy	Not very happy	Fairly happy	Very happy			
q03	Overall, would you say you are...?	Not at all happy	Not too happy	Pretty happy	Very happy			
q04	In general, how happy would you say you are?	Not at all happy	Not very happy	Fairly happy	Very happy			
q05	On the whole, now, do you think you are happy or not?	Absolutely unhappy	Not so happy	Somewhat happy	Very happy			
q06*	I am pleased with my life.	Not pleased	NLA	NLA	Pleased			
q07	How happy do you feel as you live now?	Very unhappy	Fairly unhappy	Neither happy nor unhappy	Fairly happy	Very happy		
q08	To what extent do you consider yourself a happy person?	Unhappy	Not very happy	Neither happy nor unhappy	Happy	Very happy		
q09	Are you happy?	Unhappy	NLA	NLA	NLA	Happy		
q10	Overall, how happy would you say you are…?	Extremely unhappy	NLA	NLA	NLA	Extremely happy		
q11	Taking all things together, how happy would you say things are recently?	Pretty unhappy	Fairly unhappy	Neither happy nor unhappy	Fairly happy	Pretty happy		
q12*	When I think of my life, I feel elated.	Not elated	NLA	NLA	NLA	Elated		
q13	If you were to consider your life in general, how happy or unhappy would you say you are, on the whole?	Completely unhappy	Very unhappy	Fairly unhappy	Neither happy nor unhappy	Fairly happy	Very happy	Completely happy
q14	How do you feel about your life as a whole?	Terrible	Unhappy	Mostly dissatisfied	Mixed	Mostly satisfied	Pleased	Delighted
q15	Taking all things together, how would you say things are these days?	Very unhappy	NLA	NLA	NLA	NLA	NLA	Very happy
q16	Generally speaking, are you a happy person?	Very unhappy	NLA	NLA	NLA	NLA	NLA	Very happy
q17*	I am delighted with my life.	Not delighted	NLA	NLA	NLA	NLA	NLA	Delighted
q18*	I feel cheerful about my life.	Not cheerful	NLA	NLA	NLA	NLA	NLA	Cheerful

Note. Question IDs with * refer to the made-up questions. NLA stands for "No Label Attached".

The response scales of the questions differed from each other on a variety of features: Number of response-category options (i.e., four, five, or seven categories), scale direction (i.e., unipolar or bipolar scale in terms of verbal labeling), response-scale labeling (i.e., fully-labeled scale or endpoints-labeled scale), and response-option labeling (i.e., labels varying in intensity of adjectives such as extremely, very, and somewhat). The characteristics of the response scales of the survey questions are reported in Table 3.2.

Table 3.2
The Response-Scale Characteristics of the Questions

Question ID	Country	Year	Number of Categories	Polarity	Scale Labeling
q01	US	2012	4	Unipolar	Fully-Labeled
q02	US	2007	4	Unipolar	Fully-Labeled
q03	US	1974	4	Unipolar	Fully-Labeled
q04	US	1956	4	Unipolar	Fully-Labeled
q05	JP	2012	4	Unipolar	Fully-Labeled
q06*	US	2018	4	Bipolar	Endpoints-Labeled
q07	US	1979	5	Bipolar	Fully-Labeled
q08	NL	2012	5	Unipolar	Fully-Labeled
q09	JP	2010	5	Bipolar	Endpoints-Labeled
q10	GE	2011	5	Bipolar	Endpoints-Labeled
q11	TW	2010	5	Bipolar	Fully-Labeled
q12*	US	2018	5	Bipolar	Endpoints-Labeled
q13	US	2002	7	Bipolar	Fully-Labeled
q14	US	1973	7	Unipolar	Fully-Labeled
q15	US	1987	7	Bipolar	Endpoints-Labeled
q16	NL	1968	7	Bipolar	Endpoints-Labeled
q17*	US	2018	7	Bipolar	Endpoints-Labeled
q18*	US	2018	7	Bipolar	Endpoints-Labeled

Note. Question IDs with * refer to the made-up questions. Country and Year for the made-up questions were set as "US" and "2018", respectively.

3.3.1 Data Generation and Analysis

A sample size of 1000 responses was generated for each of the made-up questions. Depending on the number of response-option categories and the intensity of happiness of the

question stems for each made-up item, the responses were distributed into the response-option categories. A few researchers in the happiness literature suggest that happiness is a latent variable that has an underlying continuous distribution (DeJonge et al., 2015; Kalmijn, 2010). It is believed to follow a left skewed beta distribution, assuming that the horizontal response scale has the negative polarity on the left and the positive polarity on the right in term of presence of happiness (DeJonge, Veenhoven, Kalmijn, & Arends, 2016). Following their suggestion, a sample of 1000 generated responses was distributed to the response-scale categories for the made-up questions so that the primary response scales followed a negatively skewed beta distribution.

Table 3.3 presents the descriptive statistics for the original scales for the questions, including sample sizes, means, standard deviations, and percentage frequencies for the response-option categories. As mentioned earlier, the descriptive statistics for fourteen questions were obtained from the World Database of Happiness. However, the reports of two questions were not complete in the database. The percentage frequencies of the response-response categories for question 15 were not reported. To generate the percentage frequencies, the responses were distributed to the seven categories of the scale, assuming equal intervals between categories, in a way that was consistent with a negatively skewed beta distribution. Given the scale mean and its variance, the α and β parameters of the beta distribution were computed as 3.47 and 1.49 using Equations 6 and 7. Given these parameters and sample size, the percentage frequencies of the scale categories were obtained.

$$\alpha = -\frac{\mu*(\sigma^2+\mu^2-\mu)}{\sigma^2}, \tag{6}$$

$$\beta = \frac{(\sigma^2+\mu^2-\mu)*(\mu-1)}{\sigma^2}. \tag{7}$$

For question 16, the sum of the percentages of answers across the response categories reported in the database summed to 104%; this sum should not have exceeded 100%. As a correction, a portion equal to 0.57% (4% divided by 7, the number of categories) was subtracted from the reported percentages of each of the seven categories so that the adjusted proportions added up to 100%. Then, the mean and standard deviation of the question were recalculated by using the adjusted proportions.

Table 3.3

Descriptive Statistics of the Original Response Scales of the Questions

	Descriptive Statistics			Percentage Frequencies (%)						
Question ID	Sample Size	Mean	Std. Dev.	1	2	3	4	5	6	7
q01	2232	3.25	0.66	1.2	8.8	53.7	36.2			
q02	1533	3.32	0.62	1.2	5.0	54.9	38.9			
q03	2248	3.18	0.75	2.8	12.5	48.9	35.7			
q04	1126	3.44	0.63	1.2	3.9	44.4	50.5			
q05	2443	3.22	0.65	1.0	9.7	55.9	33.3			
q06*	1000	3.04	0.67	0.8	17.7	57.8	23.7			
q07	1127	4.27	0.77	0.9	2.4	7.0	47.9	41.7		
q08	3845	3.96	0.70	0.6	3.4	12.7	65.8	17.5		
q09	2507	3.90	0.93	0.9	5.0	28.5	33.9	31.7		
q10	2287	3.70	1.10	4.1	8.2	30.9	26.8	29.9		
q11	1895	3.91	0.90	1.8	8.9	7.9	59.7	21.7		
q12*	1000	3.25	0.88	2.2	17.4	39.6	35.1	5.7		
q13	1160	5.52	0.98	0.1	1.4	2.7	6.0	33.7	44.0	12.1
q14	222	5.31	1.06	0.0	3.0	3.0	10.0	36.0	40.0	8.0
q15	13007	5.38	1.34	0.2	2.5	7.6	14.7	22.6	29.1	23.4
q16	1552	6.20	1.16	1.4	1.4	1.4	7.4	9.4	17.4	61.4
q17*	1000	5.14	1.16	0.1	1.3	7.7	18.5	30.7	31.2	10.5
q18*	1000	5.52	1.24	0.1	1.5	5.6	12.7	24.5	30.3	25.3

Note. Question IDs with * refer to the made-up questions.

For the made-up questions, the means were set by taking into account the strength of the question stems, numbers of response-option categories, and intensity of response labels. The variance of the four-category question was set to be around the average of the variances of the

other four-category questions. The same logic applied to the five-category and seven-category made-up questions. Given these means and variances, the α, and β parameters of the associated beta response-scale distributions were computed. The values of the parameters are reported in Table 3.4. Then, 1000 generated responses were distributed accordingly to the evenly spaced response-option categories of the made-up questions.

Table 3.4

Shape Parameters of the Beta Distributions for the Made-up Questions

Question ID	α	β
q06*	5.99	3.47
q12*	4.29	3.51
q17*	4.95	2.57
q18*	4.01	1.51

Note. Question IDs with * refer to the made-up questions.

When it comes to analyzing data, the meta-analyst makes a decision as to whether the metric of the outcome will be treated as categorical or continuous. This decision mostly depends on what kind of inferences the meta-analyst want to make about the findings. For example, if the meta-analyst wants to conclude whether the population of interest is happy or unhappy, she/he may dichotomize the scale by determining a cut-off point on the scale at which the state of happiness can be distinguished from the state of unhappiness. This operation may result in loss of information because variation in scores within each category is lost. Then, the meta-analyst can conduct a typical meta-analysis of categorical data by using proportions or odds ratios. On the other hand, if the meta-analyst is interested in the average or overall level of happiness in the populations, she/he may want to have a continuous scale on which the outcomes from the primary scales can be compared.

In this study, the analyses were conducted given the assumption that the underlying distribution of the targeted variable is continuous. Therefore, the common scale of happiness was treated as a continuous measurement scale. The raw means from the studies (either as reported or computed from frequencies) were transformed to the common scale and treated as the effect sizes. The analyses were conducted using weights similar to those proposed for the meta-analysis of raw means by Bond, Wiitala, and Richard (2003).

Bond et al. (2003) suggested to weight each raw mean difference by the inverse of its variance estimate,

$$\widehat{W}_j = \frac{1}{\sigma_j^2(\frac{1}{n_{j1}}+\frac{1}{n_{j2}})}. \quad (8)$$

Because this study used means instead of mean differences, I simplified the formula for the weight in Bond et al.'s paper for raw mean differences to the case of single mean (Bond et al., 2003). The weights for means are computed by dividing the study sample sizes by the variances of the transformed means,

$$\widehat{W}_j = \frac{n_j}{s_j^2}. \quad (9)$$

Three meta-analyses were conducted. The analyses were conducted in the R software by using the metafor package (Viechtbauer, 2010). In the first meta-analysis, the means transformed by the Linear-Stretching method were used as the effect sizes. The outcomes in the second and third meta-analyses were the means transformed by the Semantic Judgement of Fixed Word Value, and Semantic Judgement of Word Value in Context methods, respectively.

3.3.1.1 Moderator variables. In order to account for possible between-studies variance across the means of studies, the following variables were tested: the ratings of intensity of question stems, country of survey, year of survey administration, number of response-category

options, scale polarity, and scale labeling. The average ratings of question stems across the raters were entered as a continuous variable. Country of survey was dichotomized to reflect whether the study was done in the US or not. The year of survey was tested as a continuous variable. For the made-up questions, country and year were assigned as the US and 2018, respectively. Number of response-category options was modeled as a categorical variable with three levels (i.e., four, five, and seven categories; as a factor in the meta-regression, the reference category was that for the four-category items). Scale polarity (i.e., unipolar vs. bipolar) and scale labeling (fully labeled vs. endpoints labeled) were modeled as dichotomous variables. Due to collinearity, each moderator variable was tested in a separate mixed-effects meta-regression.

CHAPTER 4

RESULTS

In this section, I briefly summarize the coding tasks that the coders performed. I describe the outcomes of the three applications of coding tasks: Rating the strength of the wordings of question stems, the Semantic Judgement of Fixed Word Value Method, and the Semantic Judgement of Word Value in Context Method. Then, I report descriptive information about the transformed means and standard deviations produced by the three transformation methods. Finally, I describe the results of hypothetical meta-analyses on the outcomes created by the three transformation methods.

4.1 Task One: Rating the Strength of the Wordings of Question Stems

To account for possible variations in the wordings of the question stems, four coders were asked to rate the strength of the question stems considering the intensity of degree of happiness in the wordings of the questions. The ratings 1, 2, and 3 were used to represent the weak, moderate and strong wordings, respectively. Table 4.1 presents the ratings of all coders for all questions. These ratings reveal that at least one coder's rating differed from the others for most of the questions. All coders provided the same rating for only five questions (i.e., q06, q08, q13, q14, and q17). Two of them were made-up questions (i.e., q06 and q17). The differences in the ratings between the coders indicated that the coders interpreted the intensity of degree of happiness in the question stems differently. Table 4.1 also reports the averages of ratings across the coders for each question, and averages of ratings of each coder across the items. The predictive power of the average ratings across the coders as a moderator variable was then tested in a meta-regression that is reported later in this section.

Table 4.1

Task 1: Ratings for the Strength of the Question Stems across Coders

Question ID	Question Stem	Coder 1	Coder 2	Coder 3	Coder 4	Mean
q01	Taking all things together, would you say you are...?	3	2	1	1	1.75
q02	If you were to consider your life in general these days, how happy or unhappy would you say you are, on the whole?	2	3	3	3	2.75
q03	Overall, would you say you are...?	3	1	1	2	1.75
q04	In general, how happy would you say you are?	3	3	2	2	2.5
q05	On the whole, now, do you think you are happy or not?	3	1	1	1	1.5
q06*	I am pleased with my life.	3	3	3	3	3
q07	How happy do you feel as you live now?	3	3	2	2	2.5
q08	To what extent do you consider yourself a happy person?	3	3	3	3	3
q09	Are you happy?	3	3	2	2	2.5
q10	Overall, how happy would you say you are…?	3	3	2	2	2.5
q11	Taking all things together, how happy would you say things are recently?	3	3	1	2	2.25
q12*	When I think of my life, I feel elated.	3	2	1	1	1.75
q13	If you were to consider your life in general, how happy or unhappy would you say you are, on the whole?	3	3	3	3	3
q14	How do you feel about your life as a whole?	1	1	1	1	1
q15	Taking all things together, how would you say things are these days?	3	3	1	2	2.25
q16	Generally speaking, are you a happy person?	3	3	2	2	2.5
q17*	I am delighted with my life.	1	1	1	1	1
q18*	I feel cheerful about my life.	3	2	1	2	2
	Mean	2.72	2.39	1.72	1.94	

Note. Question IDs with * refer to the made-up questions.

In addition, the coders differed from each other in the frequency of use of the ratings. Table 4.2 provides the frequencies of the ratings assigned by the coders. While Coder 1 and Coder 2 gave the highest rating (3) to most of the question stems, Coder 3 gave the lowest rating (1) to most of the question stems. The most frequent rating that Coder 4 provided was 2.

Table 4.2

Frequency of the Ratings Assigned by Coders

	Coder 1	Coder 2	Coder 3	Coder 4
Rating 1	2	4	9	5
Rating 2	1	3	5	9
Rating 3	15	11	4	4

The differences in the frequencies of the use of ratings also indicated that interpretations of the coders about the strength of the question stems were different from each other. The intraclass correlation (ICC) was .74. The bivariate correlations among the ratings assigned by the coders are reported in Table 4.3. Based on the correlations, the levels of agreement between Coder 1 and the others were lower than the levels of agreements among the other coders.

Table 4.3

Correlations between the Ratings for the Strength of Question Stems

	Coder 1	Coder 2	Coder 3	Coder 4
Coder 1	--			
Coder 2	.38	--		
Coder 3	.10	.73	--	
Coder 4	.23	.74	.83	--

Note. Spearman's rank correlation coefficients

Overall, the average rated strength of the question stems across all questions was 2.2. It indicates that the average intensity of degree of happiness in the wordings of the questions was little more than moderate.

4.2 Task Two: The Semantic Judgement of Fixed Word Value Method (SJFW)

In the application of the Semantic Judgement of Fixed Word Value method, all of the response-option labels were presented to the coders at once. The coders assigned a position and its corresponding value to each response-option label on the 0-10 scale. Table 4.4 reports the values assigned by each coder, and the averages (and standard deviations) of those values across the coders and labels (The order of the labels in the table is the same as the order presented in Qualtrics). The average values across the coders were then used to represent (score) the labels on the new scale in the further analyses. The distribution of the means of the values is presented in Figure 4.1.

Table 4.4

Task 2: The Semantic Judgement of Fixed Word Value Method: How Happy Are You in General?

Response-Option Labels	Coder 1	Coder 2	Coder 3	Coder 4	Mean	Std. Dev.
Happy	10	7	7.5	8.1	8.15	1.31
Fairly happy	8	8	5.5	7.1	7.15	1.18
Completely happy	10	9	9.7	10	9.68	0.47
Somewhat happy	8	6	5.9	7.1	6.75	0.99
Elated	10	10	10	10	10.00	0.00
Pretty happy	10	7	8.5	8.2	8.43	1.23
Very happy	10	9	9.4	9.3	9.43	0.42
Cheerful	8	9	9.4	9.1	8.88	0.61
Extremely happy	10	10	10	10	10.00	0.00
Delighted	8	10	10	10	9.50	1.00
Mostly satisfied	7	7	6.9	6.6	6.88	0.19
Quite happy	9	9	8.6	8.3	8.73	0.34
Pleased	7	8	8	7.5	7.63	0.48
Neither happy nor unhappy	5	5	5	5	5.00	0.00
Unhappy	1	4	2	1.4	2.10	1.33
Pretty unhappy	2	2.5	0	2.7	1.80	1.24
Not cheerful	3	4	2.6	4.6	3.55	0.91
Absolutely unhappy	0	0	0	0	0.00	0.00
Not very happy	2	2	4.4	3.5	2.98	1.18
Not at all happy	0	0	0	1.3	0.33	0.65
Not delighted	2	4	6.2	5	4.30	1.78
Very unhappy	0	1	0.8	0.5	0.58	0.43
Completely unhappy	0	0	0	0.1	0.03	0.05
Not so happy	2	2	4.8	3.9	3.18	1.41
Mostly dissatisfied	3	1	0.5	3.6	2.03	1.51
Not pleased	3	3	4.9	4.2	3.78	0.94
Extremely unhappy	0	0	0.2	0	0.05	0.10
Terrible	2	0	0	0.1	0.53	0.98
Not too happy	1	2	5.3	3.5	2.95	1.87
Not elated	0	4	7.6	4.6	4.05	3.13
Fairly unhappy	1	3	1.5	2.9	2.10	1.00
Mixed	5	5	5	5	5.00	0.00
Mean	4.59	4.73	5.01	5.10		
Std. Dev.	3.85	3.45	3.60	3.34		

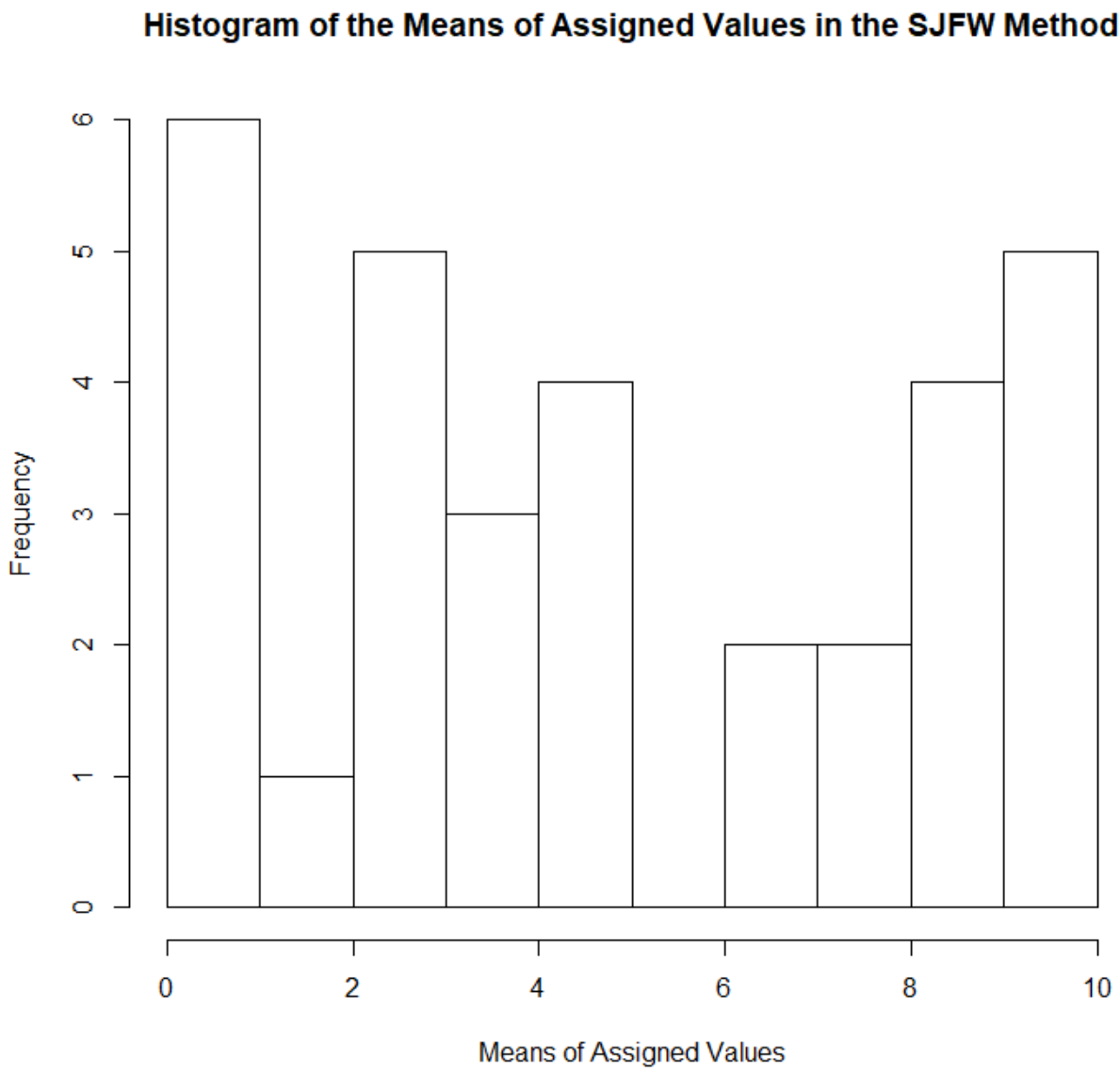


Figure 4.1. Distribution of the means of assigned values in the SJFW method.

The coders made different judgments in assigning values for many of the labels. The largest difference between the raters' mean values across labels was only 0.51, indicating that no rater was much stricter or more lenient than the others. The most striking discrepancy between the assigned values occurred when they judged the label "not elated", especially between Coder 1 who assigned 0, and Coder 3 who assigned 7.6. The coders were in total agreement in judging

five of the labels (i.e., elated, extremely happy, neither happy nor unhappy, absolutely unhappy, mixed), and were in almost total agreement for the two labels completely unhappy and extremely unhappy, with standard deviations of 0.05 and 0.1, respectively. Overall, the correlations among the values assigned by coders were strong, as seen in Table 4.5. The intraclass correlation was .98.

Table 4.5

Correlations between the Values Assigned by Coders in the SJFW Method

	Coder 1	Coder 2	Coder 3	Coder 4
Coder 1	--			
Coder 2	.92	--		
Coder 3	.83	.91	--	
Coder 4	.93	.96	.94	--

Note. Pearson product-moment correlation coefficients

4.3 Task Three: The Semantic Judgement of Word Value in Context Method (SJWC)

In performing the task for the Semantic Judgement of Word Value in Context method, the coders saw each question individually with its response-option labels. The coders assigned values to the labels by considering the characteristics of the questions such as question stem, number of response-categories, and other response-option labels of that question. The assigned values were then averaged across the coders for each question separately, and the average values were used to represent the specific labels of the separate individual questions on the 0-10 scale. The values assigned by coders, and the means and standard deviations across the coders and labels are reported in Table 4.6. The order of the questions in the table is the same as the order presented in Qualtrics. The distribution of the values across all questions is depicted in Figure 4.2.

Table 4.6

Task 3: The Semantic Judgement of Word Value in Context Method

Question Stem	Response-Option Labels	Coder 1	Coder 2	Coder 3	Coder 4	Mean	Std. Dev.
q01-Taking all things together, would you say you are...?	Very happy	10	9	10	10	9.75	0.50
	Quite happy	8	7	8	8.7	7.93	0.70
	Not very happy	4	4	5.5	3	4.13	1.03
	Not at all happy	1	1	0	0.5	0.63	0.48
q02-If you were to consider your life in general these days, how happy or unhappy would you say you are, on the whole?	Very happy	10	9	9	10	9.50	0.58
	Fairly happy	6.1	7	6.6	8.2	6.98	0.90
	Not very happy	4	3	5.4	3	3.85	1.14
	Not at all happy	1	1	0	0.1	0.53	0.55
q03-Overall, would you say you are...?	Very happy	10	10	9.2	10	9.80	0.40
	Pretty happy	8	8	8.2	8.2	8.10	0.12
	Not too happy	4	3	5.2	3	3.80	1.05
	Not at all happy	1	0	0	0.1	0.28	0.49
q04-In general, how happy would you say you are?	Very happy	10	10	9.2	10	9.80	0.40
	Fairly happy	8	8	6.6	8.3	7.73	0.76
	Not very happy	4	3	5.3	2.5	3.70	1.24
	Not at all happy	1	1	0	0	0.50	0.58
q05-On the whole, now, do you think you are happy or not?	Very happy	10	9	9.1	10	9.53	0.55
	Somewhat happy	7	6	5.1	7.5	6.40	1.07
	Not so happy	3	3	4.9	3	3.48	0.95
	Absolutely unhappy	1	0	0	0	0.25	0.50
q07-How happy do you feel as you live now?	Very happy	10	10	9.2	10	9.80	0.40
	Fairly happy	7	7	6.2	7.5	6.93	0.54
	Neither happy nor unhappy	5	5	5	5	5.00	0.00
	Fairly unhappy	3	3	2	2.5	2.63	0.48
	Very unhappy	1	1	0.8	0	0.70	0.48
q08-To what extent do you consider yourself a happy person?	Very happy	10	10	9.4	10	9.85	0.30
	Happy	7	8	8	7.5	7.63	0.48
	Neither happy nor unhappy	5	5	5	5	5.00	0.00
	Not very happy	2	3	5.1	2.5	3.15	1.36
	Unhappy	1	2	0	0	0.75	0.96
q11-Taking all things together, how happy would you say things are recently?	Pretty happy	9	7	8.6	8.5	8.28	0.88
	Fairly happy	7	6	6.5	7	6.63	0.48
	Neither happy nor unhappy	5	5	5	5	5.00	0.00
	Fairly unhappy	3	3	1.7	2.5	2.55	0.61
	Pretty unhappy	1	2	0.3	2.5	1.45	0.99

Table 4.6 *Continued*

Question Stem	Response-Option Labels	Coder 1	Coder 2	Coder 3	Coder 4	Mean	Std. Dev.
q13-If you were to consider your life in general, how happy or unhappy would you say you are, on the whole?	Completely happy	10	10	10	10	10.00	0.00
	Very happy	10	9	9.1	9	9.28	0.49
	Fairly happy	7	7	6.9	7	6.98	0.05
	Neither happy nor unhappy	5	5	5	5	5.00	0.00
	Fairly unhappy	3	3	1.7	2.9	2.65	0.64
	Very unhappy	2	2	0.5	1	1.38	0.75
	Completely unhappy	1	0	0	0	0.25	0.50
q14-How do you feel about your life as a whole?	Delighted	9	10	10	10	9.75	0.50
	Pleased	7	8	8.6	9	8.15	0.87
	Mostly satisfied	7	9	6.1	8	7.53	1.25
	Mixed	5	5	5	5	5.00	0.00
	Mostly dissatisfied	3	1	0.4	1.5	1.48	1.11
	Unhappy	3	2	2.4	1.5	2.23	0.63
	Terrible	1	0	0	0	0.25	0.50
q15-Taking all things together, how would you say things are these days?	Very happy	10	9	9.3	10	9.58	0.51
	NLA	8.5	10	7.8	8.5	8.70	0.93
	NLA	7	7	5.8	6.8	6.65	0.57
	NLA	5	5	5	5.1	5.03	0.05
	NLA	3.5	2	3.6	3.3	3.10	0.74
	NLA	2	0	1.8	1.5	1.33	0.91
	Very unhappy	1	1	0.4	0	0.60	0.49
q18*-I feel cheerful about my life.	Cheerful	10	10	8	10	9.50	1.00
	NLA	8.5	9	6.9	8	8.10	0.90
	NLA	7	7	5.6	6.4	6.50	0.66
	NLA	5	5	5	5	5.00	0.00
	NLA	3.5	3	3.6	3.4	3.38	0.26
	NLA	2	1	2.9	1.7	1.90	0.79
	Not Cheerful	1	0	2.5	0.1	0.90	1.16
q16-Generally speaking, are you a happy person?	Very happy	10	10	9.4	10	9.85	0.30
	NLA	8.5	9	8.2	8.5	8.55	0.33
	NLA	7	7	6.8	6.9	6.93	0.10
	NLA	5	5	5	5.4	5.10	0.20
	NLA	3	3	3.6	3.6	3.30	0.35
	NLA	1	1	1.7	2	1.43	0.51
	Very unhappy	0	0	0.5	0	0.13	0.25

Table 4.6 *Continued*

Question Stem	Response-Option Labels	Coder 1	Coder 2	Coder 3	Coder 4	Mean	Std. Dev.
q06*-I am pleased with my life.	Pleased	8	9	8.1	10	8.78	0.93
	NLA	6	6	5.6	7	6.15	0.60
	NLA	4	3	2.4	3	3.10	0.66
	Not pleased	2	1	1	0	1.00	0.82
q10-Overall, how happy would you say you are?	Extremely happy	10	10	9.7	10	9.93	0.15
	NLA	7	7	7.4	7.5	7.23	0.26
	NLA	5	5	5	5	5.00	0.00
	NLA	3	2	2.7	2.5	2.55	0.42
	Extremely unhappy	0	0	0.4	0	0.10	0.20
q17*-I am delighted with my life.	Delighted	7.5	10	8.8	10	9.08	1.19
	NLA	6.7	9	8	8.5	8.05	0.99
	NLA	5.6	7	6.4	7	6.50	0.66
	NLA	4.4	5	5	5	4.85	0.30
	NLA	3.3	3	3.5	3.3	3.28	0.21
	NLA	2.3	1	1.8	1.7	1.70	0.54
	Not delighted	1	0	0.5	0	0.38	0.48
q09-Are you happy?	Happy	10	8	8.1	10	9.03	1.13
	NLA	7.5	6	6.4	7.5	6.85	0.77
	NLA	5	5	5	5.1	5.03	0.05
	NLA	2.5	4	3.4	2.5	3.10	0.73
	Unhappy	0	2	2.5	0	1.13	1.31
q12*-When I think of my life, I feel elated.	Elated	9	10	10	10	9.75	0.50
	NLA	7	8	7.7	7.5	7.55	0.42
	NLA	5	5	5	5	5.00	0.00
	NLA	3	2	3.2	2.5	2.68	0.54
	Not elated	1	0	1.3	0	0.58	0.68
	Mean	5.15	5.02	4.98	5.09		
	Std. Dev.	3.17	3.39	3.16	3.58		

Note. Question IDs with * refer to the made-up questions. NLA stands for "No Label Attached".

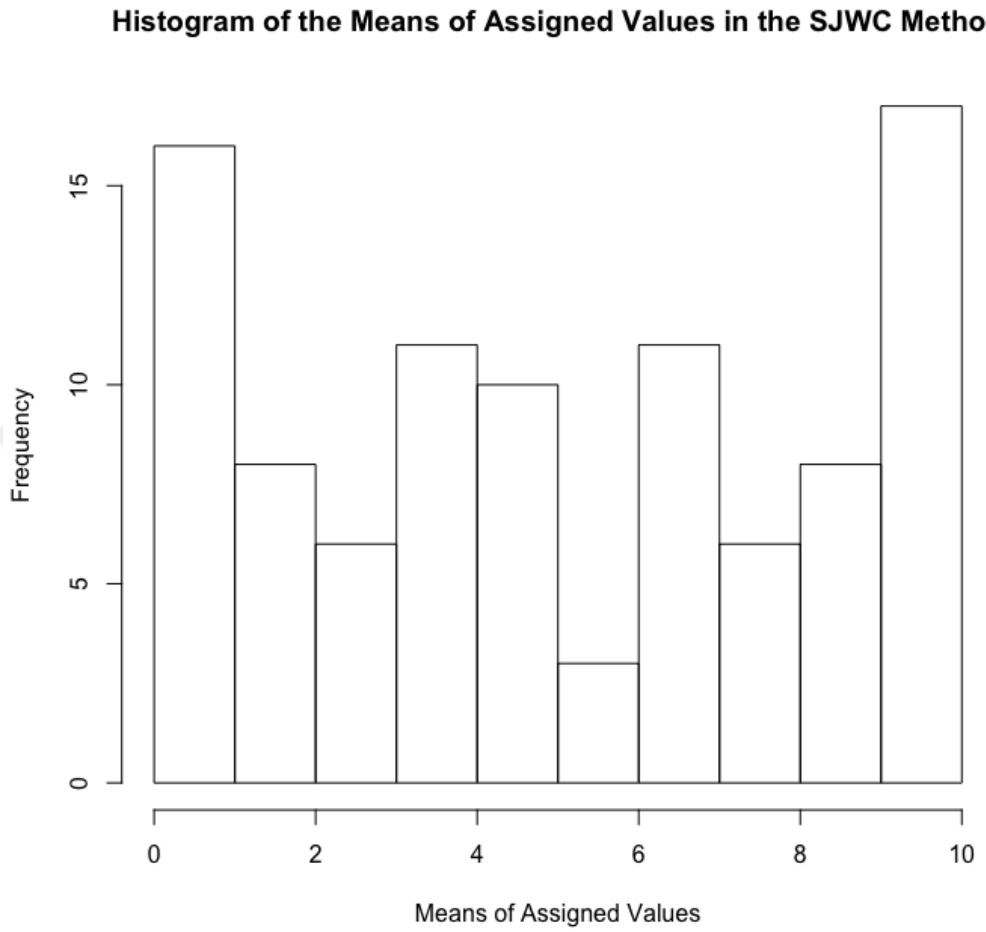


Figure 4.2. Distribution of the means of assigned values in the SJWC method.

With this method, the coders appeared to assign more similar values to labels as compared to the SJFW method. The largest difference between the raters' mean values across labels in the SJWC method (0.17) was even smaller than that for the SJFW method. The correlations among the values assigned by coders are reported in Table 4.7. Relatively large discrepancies between the coders appeared when assigning values for the label "not very happy" in multiple questions (i.e., q01, q02, 04, q08, with standard deviations of 1.03, 1.14, 1.24, and 1.36, respectively). The coders also tended to assign the values so that the assigned values on the

0-10 scale followed a linearly increasing (evenly spaced) trend. This trend was more often seen for the questions with endpoints-labeled scales (e.g., q06, q09, q10).

Table 4.7

Correlations between the Values Assigned by Coders in the SJWC Method

	Coder 1	Coder 2	Coder 3	Coder 4
Coder 1	--			
Coder 2	.96	--		
Coder 3	.95	.95	--	
Coder 4	.98	.98	.96	--

Note. Pearson product-moment correlation coefficients

4.4 The Outcomes of the Transformation Methods

The original means and standard deviations of the eighteen questions were transformed to the common 0-10 scale by applying three transformation methods: The Linear-Stretching (with two versions, LS-1 and LS-2), the Semantic Judgement of Fixed Word Value (SJFW), and the Semantic Judgement of Word Value in Context (SJWC) methods.

As mentioned earlier, the Linear-Stretching method can be applied in two ways. In the first way (LS-1), the original means were directly transformed by applying Equation 4, which is parallel to Equation 1 with the original (X) mean substituted for X_{orig}. Standard deviations were obtained using Equation 5. In the second way (LS-2), the original response-option category points (e.g., 1, 2, 3, and 4 for a 4-point scale) were transformed to the 0-10 scale using Equation 1. The transformed points for labels used more than once were averaged across all questions. The average transformed point was then assigned as the transformed value for that label across all questions. The transformed values were then weighted by the proportions of responses in each of the response-option categories for each question using Equations 2 and 3.

For the SJFW and SJWC methods, the coders' ratings were also weighted by the proportions. The methods for obtaining the ratings for the SJFW and SJWC methods were described earlier. As was done for the LS-2 method, the transformed means and standard deviations for the SJFW and SJWC methods were calculated using Equations 2 and 3, respectively.

Table 4.8

The Transformed Unweighted Means and Standard Deviations from the Four Transformations

Question ID	LS-1		LS-2		SJFW		SJWC	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
q01	7.50	2.20	7.42	2.17	8.37	1.96	8.16	1.75
q02	7.73	2.07	7.83	1.95	7.75	1.76	8.45	1.63
q03	7.27	2.50	8.01	2.40	7.87	2.38	7.76	2.24
q04	8.13	2.10	8.19	1.95	8.05	1.75	8.53	1.67
q05	7.40	2.17	7.33	2.11	7.23	1.97	7.09	2.00
q06*	6.81	2.22	6.25	2.04	6.40	0.86	6.19	1.87
q07	8.18	1.93	7.88	1.92	7.77	1.76	7.83	1.92
q08	7.40	1.75	8.22	1.79	7.76	1.59	7.49	1.64
q09	7.25	2.33	6.51	1.91	6.49	1.41	6.78	1.85
q10	6.75	2.75	6.75	2.76	6.77	2.75	6.66	2.70
q11	7.28	2.25	6.63	1.74	6.71	1.82	6.40	1.65
q12*	5.62	2.21	5.62	2.21	7.39	1.31	5.66	2.12
q13	7.53	1.63	8.28	2.09	8.38	2.16	8.03	1.88
q14	7.18	1.77	7.48	2.11	6.54	1.32	7.36	1.76
q15	7.30	2.23	7.29	2.10	7.04	1.97	7.28	2.22
q16	8.67	1.72	8.58	2.03	8.25	1.91	8.87	1.59
q17*	6.90	1.93	6.90	1.93	7.89	1.00	6.63	1.75
q18*	7.54	2.07	7.53	2.07	7.57	1.10	7.30	1.89
Average of Means	7.36		7.37		7.46		7.36	
Std. Dev.	0.65		0.80		0.66		0.87	

Note. Question IDs with * refer to the made-up questions.

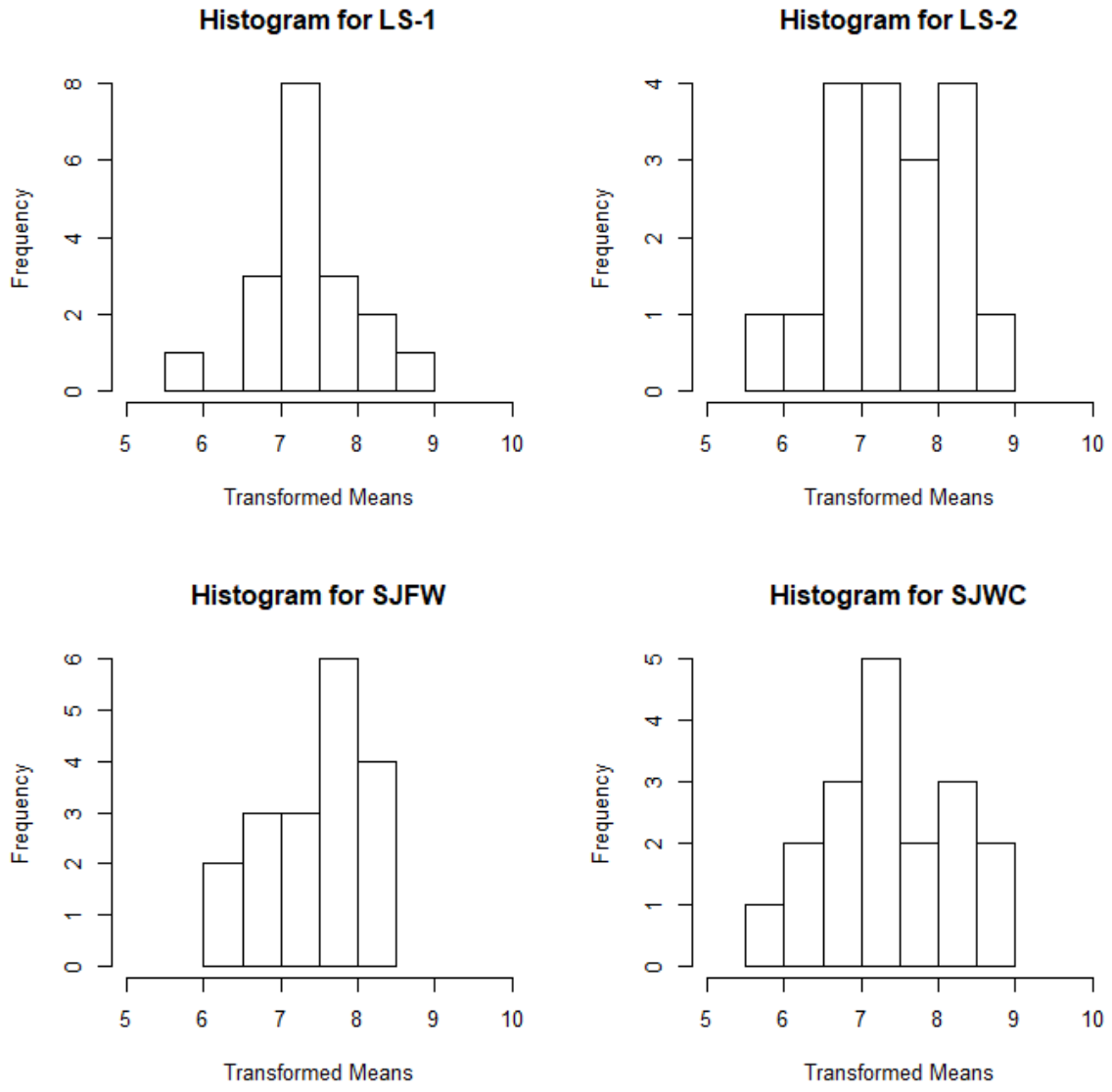


Figure 4.3. Distributions of means from the four transformation methods.

Table 4.8 reports the transformed means and standard deviations from the four transformation methods. The distributions of the transformed item means from the methods are also depicted in Figure 4.3. The average values of the transformed means across all questions appeared to be very similar to each other across the transformation methods (i.e., 7.37, 7.36, 7.46, 7.36 for LS-1, LS-2, SJFW, and SJWC, respectively). However, both Figure 4.3 and the standard deviations in Table 4.8 show differences in variation across the four methods. The SJFW method was the less

variable of the rating methods. In the SJFW method, each label was assigned one value by each coder. On the other hand, in the SJWC method, labels that appeared in multiple questions were rated several times, often with different values. In addition, in the SJWC method, the highest and lowest labels for every item could potentially be assigned 0 and 10, the most extreme values on the numerical scale. In contrast, in the SJFW method, the extreme numerical values were assigned only to the most extreme verbal labels across all items. Other less extreme verbal labels (that may have been the highest or lowest for a particular item) would then receive less extreme values on the numerical scale. This would reduce overall variability in the means of the SJFW method.

In addition, the Pearson product-moment correlations between the transformed means produced by different methods showed some variations. The strongest correlation appeared to be between the means of the LS-2 and SJWC, r = .91. This supports the previous observation that the coders tended to assign the values linearly in the SJWC method. The weakest correlation was found to be between the means of LS-1 and the means of SJFW, r = .47. The correlation between the means of SJFW and SJWC was 0.69. All bivariate Pearson product-moment correlations among the means are reported in Table 4.9. The outcome variables were on the whole linearly related to each other, as shown in Figure 4.4.

Table 4.9

The Pearson Product-Moment Correlations between Means from the Four Transformations

	LS-1	LS-2	SJFW	SJWC
LS-1	--			
LS-2	.84	--		
SJFW	.47	.67	--	
SJWC	.87	.91	.69	--

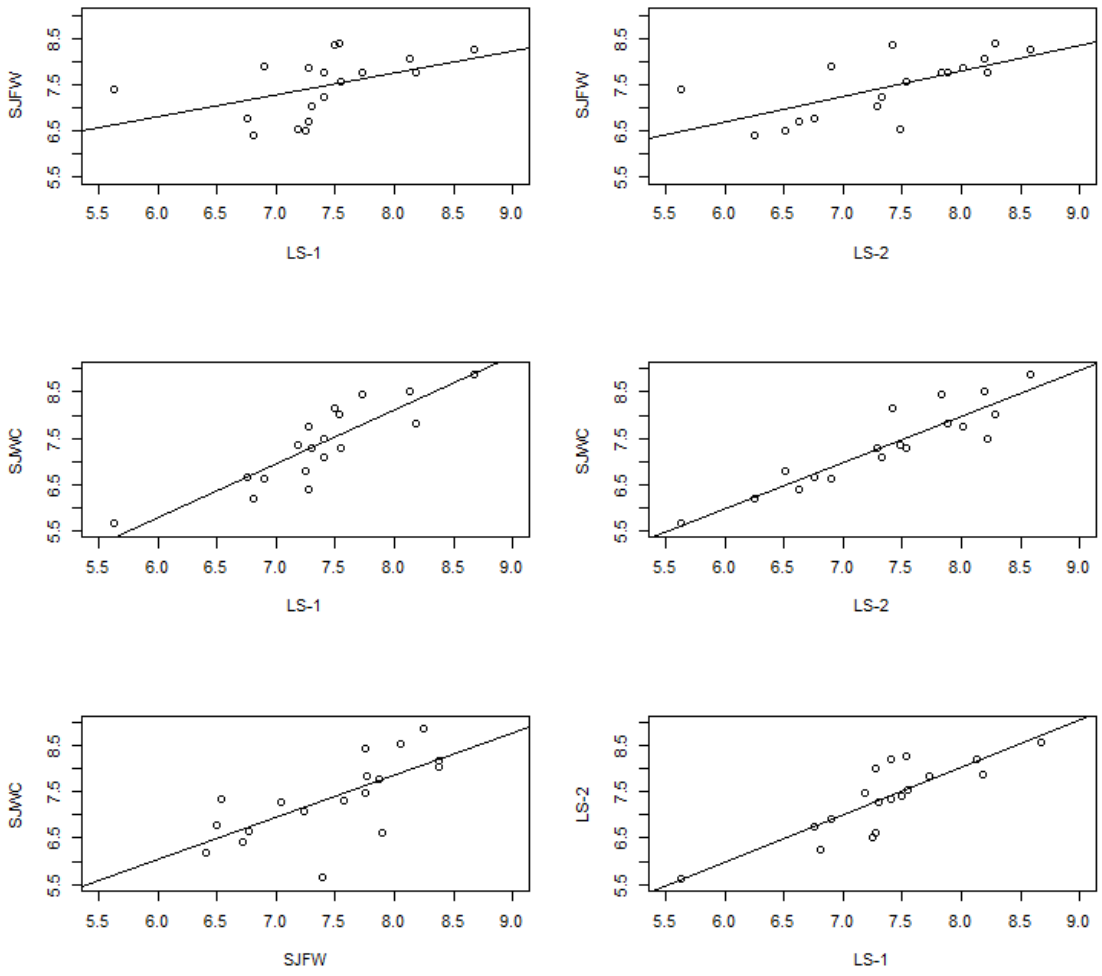


Figure 4.4. Scatterplots among the means from the four transformations.

4.5 The Results of Meta-Analyses

The meta-analyses were conducted on the means from the LS-2, SJFW, and SJWC transformations. I excluded one of the linear-transformation methods, LS-1, from further analyses because both linear transformation methods yielded very similar transformed scores. The fixed-effects models for each outcome revealed a great amount of between-studies heterogeneity among the effect sizes. The survey studies did not share a common mean value for each set of outcomes. The forest plots for the fixed-effects models for each outcome are presented in Figures 4.5, 4.6, and 4.7, respectively. The forest plots for the fixed-effects models

show that the effect estimates were very precise; the standard errors of the estimates were so small that the 95% confidence intervals around the estimated effects were all very narrow. This was expected given that the sample sizes for the primary studies were very large. The *Q* tests and overall effect estimates based on the fixed-effects models for each outcome are reported in Table 4.10.

Table 4.10

The Fixed-effects Models

Outcome	*Q* (*df* = 17)	Weighted Mean	Standard Error	Z value	95% CI
LS-2	4130.21*	7.41	0.01	734.43*	[7.39 - 7.43]
SJFW	5109.66*	7.30	0.01	875.83*	[7.28 - 7.32]
SJWC	5676.23*	7.42	0.01	778.08*	[7.40 - 7.44]

*$p < .05$

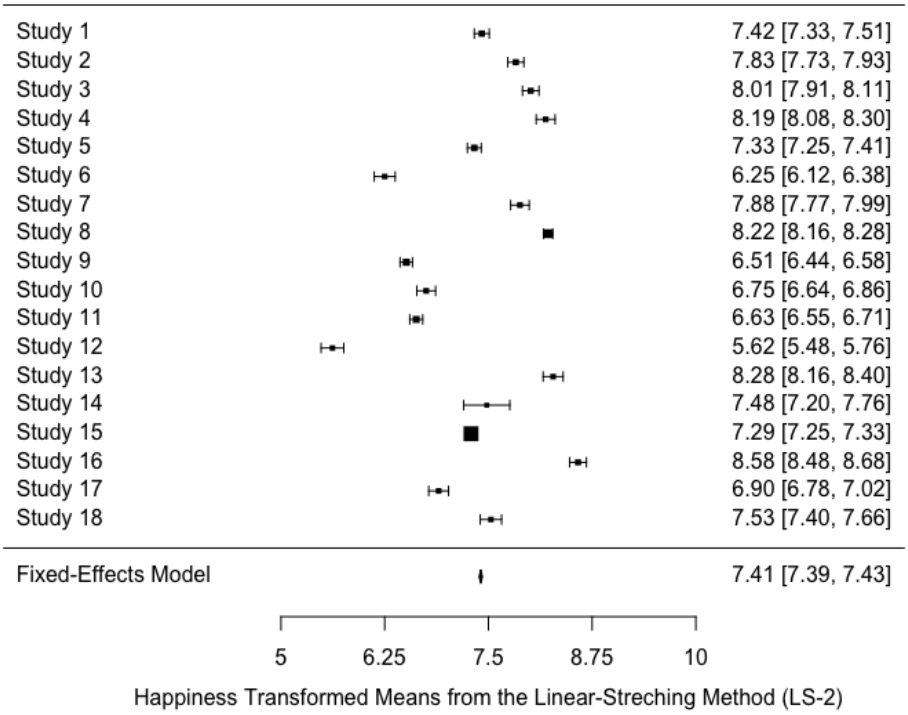


Figure 4.5. Fixed-effects forest plot for the Linear-Stretching method.

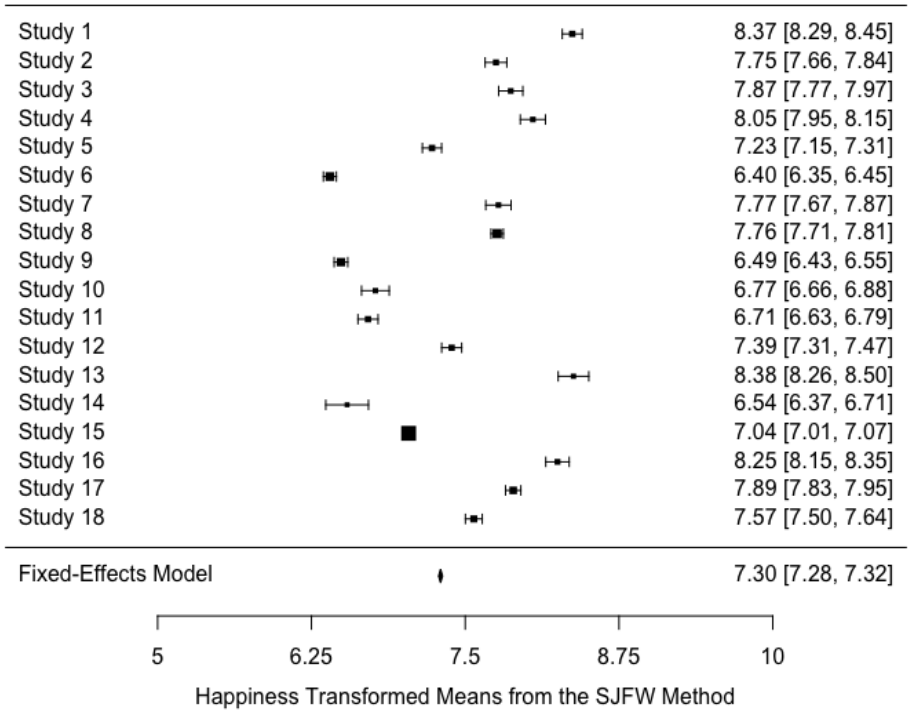


Figure 4.6. Fixed-effects forest plot for the Semantic Judgement of Fixed Word Value method.

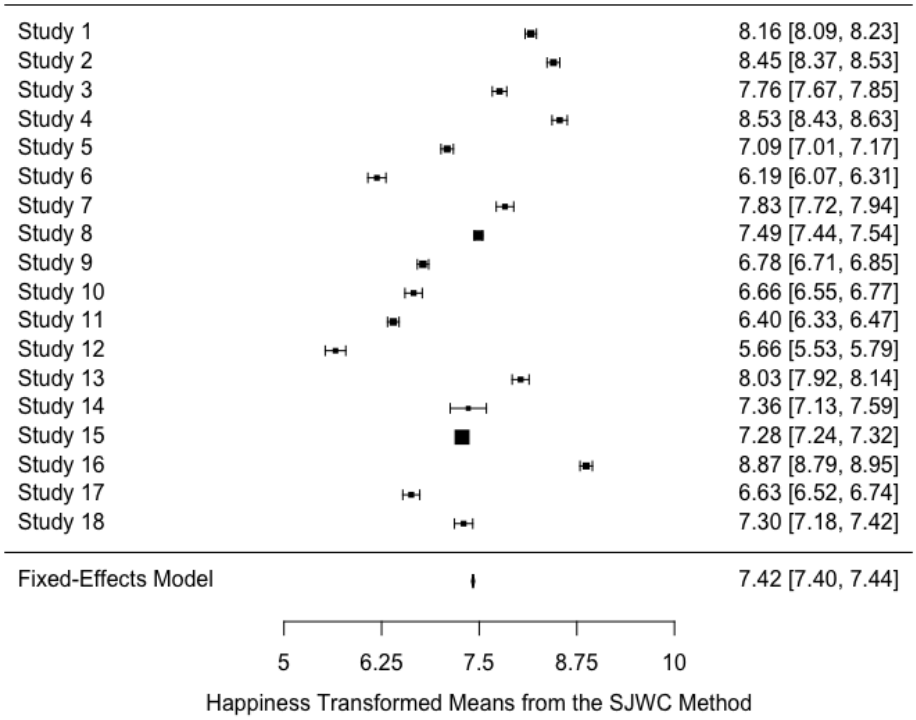


Figure 4.7. Fixed-effects Forest Plot for the Semantic Judgement of Word Value in Context method.

For each of the three outcomes, the effect-size estimates were obtained by incorporating the between-studies variability ($\hat{\tau}^2$) along with the sampling variability in the weights for the random-effects analyses. Table 4.11 reports the random-effects-model estimates of mean happiness for each outcome. Between-studies variances were estimated to be 0.63, 0.44, and 0.76 for the LS-2, SJFW, and SJWC data, respectively. I^2 values indicated almost all variability in the means of the outcomes was due to the between-studies variance rather than sampling error.

Table 4.11

The Random-effects Models

Outcome	Weighted Mean	Standard Error	Z value	95% CI	$\hat{\tau}^2$	I^2 (%)
LS-2	7.37	0.19	39.18*	[7.00 - 7.74]	0.63	99.69
SJFW	7.46	0.16	47.79*	[7.15 - 7.76]	0.44	99.70
SJWC	7.36	0.21	35.77*	[6.96 - 7.76]	0.76	99.77

*$p < .05$

To account for the between-studies variability in the outcomes, mixed-effects models were applied. Each moderator variable was tested in a separate the meta-regression analysis for each outcome. The correlations among the moderator variables are reported in Table 4.12. Scale labeling and polarity are strongly correlated with each other. A strong correlation between these variables was expected because the response-option scales of all unipolar questions were fully labeled. Also, scale polarity was strongly correlated with the number of response-option categories because most of the four-category scales were bipolar scales, and most of the seven-category scales were unipolar. In addition, because the characteristics of all made-up questions were set as bipolar, endpoint-labeled, US, and 2018, the variable that represented whether questions were made up was correlated with scale polarity, labeling, country, and year. These strong correlations suggest that models that include more than one predictor may have a high potential for multicollinearity. If a meta-analyst wants to examine multi-predictor models, he/she

must use standard collinearity checks to make sure that collinearity problems have not affected the model.

Table 4.12

Correlations among the Predictor Variables

	Number of Category	Polarity	Labeling	Stem Ratings	Made-up	Country	Year
Number of Category	--						
Polarity	.64[c]	--					
Labeling	.42[c]	-.71[b]	--				
Stem Ratings	-.17[d]	-.21[a]	.01[a]	--			
Made-up	.19[c]	.43[b]	-.60[b]	.23[a]	--		
Country	.50[c]	-.08[b]	.08[b]	.21[a]	.38[b]	--	
Year	-.04[d]	-.28[a]	.31[a]	.02[d]	-.51[a]	.17[a]	--

Note. a = Point-biserial correlation coefficient, b = Phi coefficient, c = Cramer's V coefficient, d = Pearson product-moment correlation coefficient

4.5.1 Number of Response-Option Categories

Number of response-option categories did not have a significant impact on explaining the between-studies variances in the means for all methods. In other words, for all transformed scales the means were not related to whether the number of response-option categories of the scales was four, five, or seven.

Table 4.13

The Mixed-effects Model - Number of Response Categories

Outcome	Intercept (S.E.)	Slope (S.E.)	Q_M	Q_E	$\hat{\tau}^2$	I^2 (%)	Pseudo-R^2 (%)
LS-2	7.51* (0.317)	Five: -0.57 (0.448) Seven: 0.17 (0.448)	3	3987.28*	0.60	99.61	5.57
SJFW	7.61* (0.271)	Five: -0.46 (0.383) Seven: 0.002 (0.383)	1.96	4988.88*	0.44	99.67	0
SJWC	7.70* (0.335)	Five: -0.89 (0.474) Seven: 0.12 (0.474)	4.19	4503.51*	0.67	99.71	11.46

*$p < .05$

Note. The reference group for the number of response categories is “Four categories”.

4.5.2 Scale Polarity

The weighted mean of the effect-size estimates of the bipolar scales (7.06) was significantly smaller than the weighted mean of the estimates of the unipolar scales (7.83) in the set of means from the SJWC method. The Pseudo R^2 shows the reduction in the between-studies variance ($\hat{\tau}^2$) between the random-effects model (with no predictor) and the mixed-effects model (with the predictor). The $\hat{\tau}^2$ decreased from 0.76 to 0.64, which corresponded to a 15% change in the residual between-studies variance. Scale polarity was not a significant predictor in explaining the variance in the transformed means from the LS-2 and SJFW methods. For all outcome scales the test of the residual-variance component (Q_E) was still significant, indicating that a significant amount between-studies variance was left unexplained.

Table 4.14

The Mixed-effects Model - Polarity

Outcome	Intercept (S.E.)	Slope (S.E.)	Q_M	Q_E	$\hat{\tau}^2$	I^2 (%)	Pseudo-R^2 (%)
LS-2	7.78*	-0.67					
	(0.281)	(0.361)	3.48	2998..60*	0.55	99.62	12.84
SJFW	7.65*	-0.32					
	(0.250)	(0.320)	1.02	3866.51*	0.44	99.69	0.23
SJWC	7.83*	-0.78*					
	(0.304)	(0.389)	3.99*	4673.17*	0.64	99.72	15.03

*$p < .05$

Note. The reference group for scale polarity is "Unipolar scale".

4.5.3 Scale Labeling

For outcomes based on the LS-2 and SJWC methods, the weighted means of the effect-size estimates of the questions with fully-labeled scales were significantly larger than those of the questions with endpoints-labeled scales. The magnitude of the impact of scale labeling is very similar across LS-2 and SJWC. Scale-labeling type accounted for about 22% and 16% of the total variabilities in the outcomes of the LS-2 and SJWC methods, respectively. On the other

hand, the weighted mean effect size was not found to be different between the fully-labeled scales and endpoints-labeled scales for the SJFW method, indicating that scale labeling did not influence variability in the means transformed by the SJFW method.

Table 4.15

The Mixed-effects Model - Labeling

Outcome	Intercept (S.E.)	Slope (S.E.)	Q_M	Q_E	$\hat{\tau}^2$	I^2 (%)	Pseudo-R^2 (%)
LS-2	6.93* (0.250)	0.80* (0.336)	5.65*	3243.07*	0.5	99.56	21.58
SJFW	7.22* (0.228)	0.42 (0.306)	1.88	3898.84*	0.41	99.67	5.01
SJWC	6.92* (0.282)	0.79* (0.379)	4.33*	5127.26*	0.63	99.71	16.41

*$p < .05$

Note. The reference group for scale labeling is "Endpoints-labeled scale".

4.5.4 Ratings of the Question Stem Strength

The coded ratings of the question-stem strength were found to be one of the weakest predictors in accounting for the variations in the transformed means across the questions. Pseudo R^2 values indicated that for all transformed outcomes, the ratings failed to explain any heterogeneity among the effect estimates.

Table 4.16

The Mixed-effects Model - Ratings of Question Stem Strength

Outcome	Intercept (S.E.)	Slope (S.E.)	Q_M	Q_E	$\hat{\tau}^2$	I^2 (%)	Pseudo-R^2 (%)
LS-2	6.73* (0.707)	0.29 (0.310)	0.89	3648.21*	0.64	99.69	0
SJFW	7.33* (0.601)	0.06 (0.264)	0.05	4769.22*	0.46	99.71	0
SJWC	6.66* (0.772)	0.32 (0.339)	0.87	5452.24*	0.76	99.77	0

*$p < .05$

4.5.5 Question Type

Question type was also tested to see if the weighted means of effect estimates differed between the set of made-up questions and the set of the questions from the World Happiness Database. The results showed for the LS-2 and SJWC means, the averages of effect estimates were smaller in the set of made-up questions (6.58 for LS-2, 6.44 for SJWC) than in the database questions (7.60 for LS-2, 7.62 for SJWC). This result was not surprising because the made-up questions were generated for the purpose of increasing heterogeneity in the question stems, and were made up to be more difficult to endorse. Thus, it is sensible that their statistics turned out to be different from those of the rest of the questions. Interestingly, the effect estimates were not significantly different across the two sets of questions with the means from the SJFW method. While question type accounted for about 26% to 29% of the total variability in the mean-effect estimates of the LS-2 and SJWC methods, including this predictor in the model for the means of SJFW method did not explain any between-studies heterogeneity.

Table 4.17

The Mixed-effects Model – Question Type

Outcome	Intercept (S.E.)	Slope (S.E.)	Q_M	Q_E	$\hat{\tau}^2$	I^2 (%)	Pseudo-R^2 (%)
LS-2	7.60* (0.184)	-1.02* (0.390)	6.89*	3466.69*	0.47	99.59	25.8
SJFW	7.50* (0.181)	-0.18 (0.383)	0.24	5059.26*	0.46	99.7	0
SJWC	7.62* (0.197)	-1.18* (0.417)	7.93*	4603.82*	0.54	99.69	29.04

*$p < .05$

Note. The reference group for question type is "Database question".

4.5.6 Country

The survey-administration location was also examined to see if the weighted means of the survey questions that were administered in the US were different from the weighted means of

questions administered in any countries other than the US. The results showed that the effect estimates were not influenced by country for any of the transformation methods. The weighted means of the effect estimates were very similar across the survey questions of the US and those of the non-US respondents.

Table 4.18

The Mixed-effects Model - Country

Outcome	Intercept (S.E.)	Slope (S.E.)	Q_M	Q_E	$\hat{\tau}^2$	I^2 (%)	Pseudo-R^2 (%)
LS-2	7.33*	0.05					
	(0.335)	(0.41)	0.02	4127.51*	0.67	99.68	0
SJFW	7.20*	0.38					
	(0.267)	(0.327)	1.37	5045.32*	0.43	99.68	2.18
SJWC	7.22*	0.22					
	(0.364)	(0.446)	0.24	5521.03*	0.79	99.77	0

*$p < .05$

Note. The reference group for question type is "Non-US".

4.5.7 Year of Administration

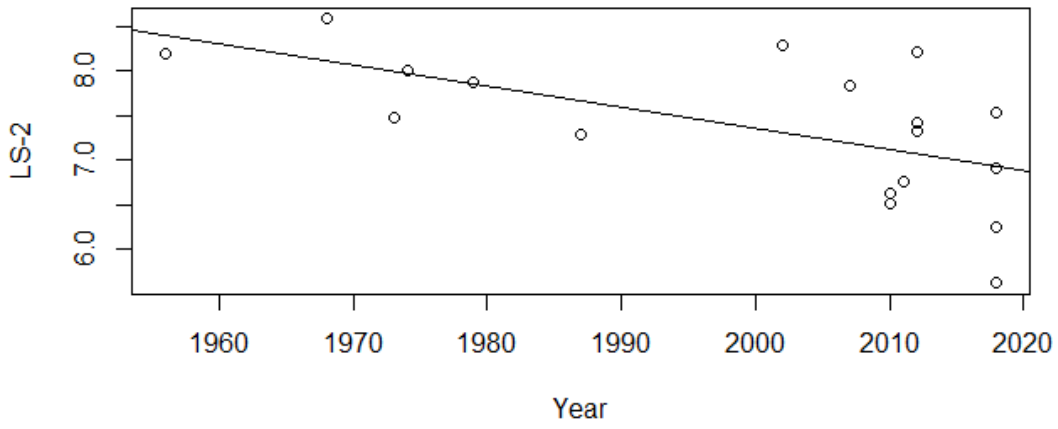
The range for the year of survey administration was from 1956 to 2018. I centered the variable around 2000 (close to the mean value, 1999.17) so that the value of the intercept could be interpreted easily. Year as a predictor helped explaining some of the variation in the means of LS-2 and SJWC methods. The scatter plot of the weighted means from the LS-2 and SJWC methods predicted by year of administration was depicted in Figure 4.8. The estimated change in the weighted means with a one unit increase in year was very small though (0.02 for both conditions). That is, it would take 50 years to see a shift of 1 point on the 0-10 scale. The $\hat{\tau}^2$ values were reduced about 35% thanks to the predictive power of year on the means of LS-2 and SJWC. However, year of administration was not found be a significant moderator in the SJFW method, indicating that it did not explain much heterogeneity in the means transformed by the SJFW method.

Table 4.19

The Mixed-effects Model - Year

Outcome	Intercept (S.E.)	Slope (S.E.)	Q_M	Q_E	$\hat{\tau}^2$	I^2 (%)	Pseudo-R^2 (%)
LS-2	7.35*	-0.02*					
	(0.154)	(0.008)	9.51*	3447.62*	0.42	99.51	33.57
SJFW	7.45*	-0.008					
	(0.156)	(0.008)	1.10	4952.08*	0.43	99.69	0.68
SJWC	7.34*	-0.02*					
	(0.164)	(0.008)	10.67*	4057.85*	0.48	99.64	36.42

*$p < .05$



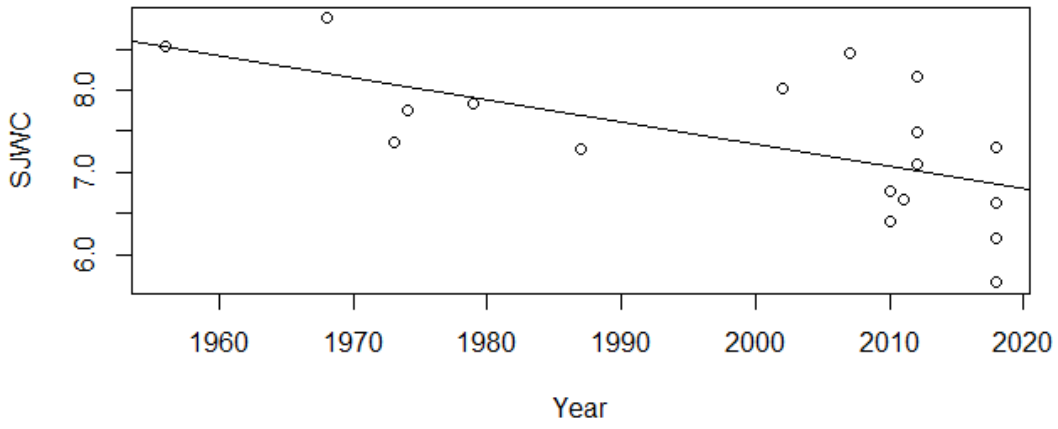


Figure 4.8. Scatter plots of the weighted means from LS-2 and SJWC, predicted by year of administration.

Summing up the results in the mixed-effect models, number of response-option categories, ratings of question stems, and location of the survey administration (i.e., the US and non-US) did not have significant relationships with the effect estimates for all transformation methods. Scale labeling (i.e., fully-labeled and endpoints-labeled), question type (i.e., made-up and database), and year of survey administration were found to be significant moderator variables in explaining between-studies variance in the LS-2 and SJWC outcomes. Scale polarity (i.e., unipolar and bipolar) was found to be a significant predictor in the SJWC outcome only. The between-studies variance in means transformed by the SJFW method was not significantly explained by any of the moderator variables tested in the mixed-effects models above.

In addition to the models described above, a few mixed-effects models with multiple moderator variables were tested. For the LS-2 outcomes, scale labeling, question type and year of administration (significant predictors from previous models) were modeled together. For the SJWC outcomes, the mixed-effects model included scale labeling, question type, year of administration, and polarity. For both models, year of administration was the only significant predictor. The other variables were not significant when they were entered in a model with other moderator variables.

CHAPTER 5

CONCLUSIONS AND DISCUSSION

In summarizing survey questions about the same construct, meta-analysts often encounter situations in which the questions from different surveys differ from each other in a variety of ways. These may include differences in wording of question stems, number of response-option categories, scale format and direction, and response-option labeling. Consequently, meta-analysts face a problem in comparing and summarizing the outcomes of such diverse survey questions. In this study, I illustrated how meta-analysts can deal with such issues that hamper comparability. To compare and synthesize outcomes of the survey questions appropriately, one may use scale transformations to locate different scales of survey questions on a common scale. I studied three transformation methods, and applied them to a set of survey questions about happiness. Also I recommended evaluating the strength of question stems, and using strength as a potential moderator variable. Last, I illustrated a hypothetical meta-analysis that synthesizes the mean responses to survey questions about happiness.

5.1 Transforming Outcomes of Survey Questions

The linear-transformation method transformed the scales with different numbers of response-option categories to a common scale running from 0 to 10. I presented two ways of applying the linear-transformation method. In one way, the means of the original scales were directly transformed by applying the linear-transformation formula (Kalmijn, 2010). The transformed values were considered as the means on the 0-10 scale for each question. In the other way, instead of transforming means directly, the response-option points of the original scales were linearly transformed. In situations where response option-labels appeared in more than one question scale, the transformed points of that response-option-labels were averaged

across the questions in which the label appeared. In this way, the points assigned to the response-option scales became the same across all questions. The two types of linear-transformation methods yielded very similar outcomes. The correlation between the two sets of transformed means in my hypothetical meta-analyses example was very strong. In the analyses, I used the outcomes obtained by the second method of linear transformation, as it dealt with the inconsistencies between the points assigned to the same response labels across the questions.

The linear-transformation method is a straightforward technique to scale the different questions on a common metric. With the linear transformation, one can easily transform the means of different survey questions and take them to a 'comparable' level. However, as mentioned earlier, the method is limited in many ways. It assumes that the interval distances between the response-option categories are the same. This is not a realistic assumption in most cases. Take question 14 in the dataset as an example. The question has unique labels to represent the response options. One could easily argue about the accuracy of assuming that the difference between category 1 'terrible' and category 2 'unhappy' is the same as the difference between category 6 'pleased' and category 7 'delighted'. Also, the linear method does not take into account the differences in the question stems and labeling of the response options. For example, regardless of the differences in the wordings of questions 7 and 8, the point scored 4 on both scales will be transformed to 7.5 on the 0-10 scale. Moreover, the point 7.5 will then represent both the label of the response scored 4 on question 7 ('fairly happy') and the label scored 4 on question 8 ('happy') on the 0-10 scale (assuming the labels appeared only in these questions). This assignment of equal scores implies that the labels mean the same thing. This is also arguable, and on the surface is not sensible.

In addition, it is not clear how to apply the linear-transformation method with unipolar response scales. Unlike the ranges of bipolar scales (e.g., for unhappiness-happiness; very unhappy-very happy), unipolar scales typically cover one spectrum of the construct (e.g., for happiness; not too happy to very happy). Therefore, the lower endpoints of the responses for unipolar scales often would fall somewhere close to the middle of the potential response range of a common unhappiness-happiness scale. If this is the case the direct linear transformation may not work as expected. If the meta-analyst believes that the response option labels of unipolar scales should not cover the full range of happiness, she/he may be able to assign an appropriate covered range within the 0-10 scale, given the response-option labels.

For example, the meta-analyst may make a subjective judgment and decide that the lowest response option should be transformed to 4, instead of 0, on the 0-10 common scale. The linear-transformation method could then be applied to transform the primary scale onto the common scale between 4 and 10. Alternatively, the meta-analyst may have study coders make the decision about the position of the lowest response option on the 0-10 scale. For example, raters could be asked "On a new scale from 0 to 10 representing the range from unhappiness to happiness, where do you think the label "lowest response option" from the primary scale should fall?". The wording of the lowest response option from the original scale would be shown inside of the quotation marks in place of "lowest response option".

A second transformation technique was the Semantic Judgement of Fixed Word Value method. With this approach all response-option labels from the collection of survey questions are placed on the 0-10 scale by a set of coders without regard to the question-stem wordings. The advantage of this method over the linear-transformation method is that the degrees of happiness denoted by the response-option labels are evaluated by the coders. Thus, transformed values of

the labels on the 0-10 scale are determined after taking into account the perceived differences in the labels. However, this method is not capable of dealing with differences in the wordings of the primary questions, and the numbers of response options of the primary scales. As mentioned above, coders might interpret the response-option labels differently when they see the question stems to which the labels belong. Also, they might assign different values to the same response-option label depending on whether it is in a long scale or a short scale.

Some of these limitations were addressed in discussions of the Semantic Judgement of Word Value in Context method above. In this method, coders made their judgments about the intensity of the response-option labels by considering the wording of the questions and other-response options for each scale. However, as noted above, my observation was that the coders may not have fully paid attention to the wordings of the questions when they were assigning values to the response options. With the SJWC method, the coders assigned values to 96 response options. They might have found the task too repetitive and boring. Consequently, they might have assigned the values with less-than-optimal attention to the intensity of question stems and the response-option labels.

In the SJFW method, the coders were presented 32 response-option labels, all of which were presented on the same screen. The coders were able to make relative comparisons among the response-option labels when assigning values. In this sense, the SJFW method was easier to apply in practice than the SJWC method. Consequently, fatigue effects appear to be less likely in the SJFW method than the SJWC method. In general, I recommend using the SJFW method when the meta-analyst has many survey questions to summarize. If the survey questions differ drastically from each other in wordings of the question stems, then I recommend using the SJWC method.

5.2 Example Meta-analyses

Each transformation method was applied to the scales of eighteen survey questions about happiness in order to locate them on a common metric. Once they were on the same metric, they could be summarized in a meta-analysis. In this study, I conducted examples of meta-analyses on the outcomes of all three transformation methods. The meta-analyses used the transformed mean values of answers to the survey questions as effect sizes. The precisions of the estimates were extremely high within each set (e.g., small standard errors and confidence intervals around weighted means). The values of *Q* statistics were very large in all three analyses, indicating large between-studies variances. This was due in part to the large sample sizes of the survey studies, which led to extremely small within-study variances. The point estimates (i.e., means) for the individual questions seemed to sit apart from each other, as seen in the forest plots. In the random-effects models, the I^2 values were also very large, indicating that almost all the variation in observed mean effects was due to the heterogeneity in true mean effects rather than sampling error. In the mixed-effect models, even though some of the moderator variables were found to be significant in accounting for the between-studies variances in the transformed means (discussed below), the I^2 did not drop much at all. The changes in I^2 values were not larger than 1%. This suggests that even significant predictors could account for only tiny proportions of between-studies variances.

Mixed and random-effects meta-analyses models took into account the fact that the estimates did not share a common effect in each set of the outcomes. The random-effects weighted mean estimates were obtained by accounting for the between-studies variances in each outcome set. The weighted mean effect estimates for the data from all transformation methods were very similar to each other. Expectedly, the ratios of between-studies variance to the total

variability were extremely large for each outcome because the within-study variances were so small (i.e., I^2 values were around 99%). Mixed-effects models were fitted in an attempt to explain the between-studies variances: moderators included survey characteristics studied such as number of response options, scale polarity, labeling, ratings of question-stem strength, an indicator of the made-up questions, country, and year.

The mixed-effects model for the ratings of the question-stem strength revealed that the ratings did not explain any heterogeneity in the means for data from all three methods. This result suggests that the intensity of degree of happiness used in the wordings of questions doesn't have an impact on the variations between the means of the happiness survey questions in this collection. Looking at Table 4.2, Coder 1 and 2 found many question stems to be strong statements, and Coder 3 rated most of them as weak statements. Most of the question stems were called moderate statements by Coder 4 in terms of intensity of happiness. The ratings showed clear discrepancies across the coders. Take question 12 as an example. This question was one of the made-up questions. I used the word "elated" to introduce a strong statement in the question stem. The initial expectation was that the coders would tend to rate this item as a 3 reflecting a strong statement. However, as seen in the ratings of this question in Table 4.1, two coders assigned 1, one coder assigned 2, and one coder assigned 3. It may be possible that the coders perceived the meaning of "intensity of happiness" differently.

To avoid possible misunderstandings about the rating task aimed at differentiating the question stems, the meta-analyst should conduct a training session with the coders. In the training session, the meta-analyst can explain the procedures of the coding task and discuss the conceptual meaning of "strength" of the question stem (in this study, it was the intensity of degree of happiness). Then, she/he can address any questions and comments to reduce possible

misunderstandings, and clarify any concerns. The meta-analyst may want to also demonstrate an example of the rating task with a sample of questions. The demonstration may be followed up with a practice session in which coders rate another set of questions, and discuss them to achieve greater inter-rater consistency in the real rating task.

This type of practice can be useful for question-stem ratings and the ratings of labels in the transformation methods. The mean ratings for each coder can be used to assess whether any coder is deviating from the rest during practice or in the context of further coding tasks. In cases where one or more coders show extreme ratings, the meta-analyst may ask the coders to discuss their discrepancies and resolve them.

In this study, the ratings of question stems were averaged across the coders. Then, the average values were used as a predictor variable in the mixed-effects models. Alternatively, as just discussed, the meta-analyst may ask coders to resolve any discrepancies in ratings, and come up with one rating for each question that they all agree on. Then, the meta-analyst may introduce the ratings of stem strength as a categorical variable (e.g., with 3 levels; weak, moderate, and strong) in the mixed-effects models, and test its predictive power as a grouping variable.

The happiness questions examined in this study were very simple, thus were very similar to each other. Even though the question stems sounded similar (especially the questions from the World Database of Happiness), coders used the whole range of ratings (1, 2, and 3) when rating the questions. With other psychological concepts, the survey questions may differ from each other to a greater extent. The meta-analyst may consider using more ratings to capture possible differences in the strength of question stems. Indeed, at the end of the rating session in this study, one of the coders stated that had a wider range of rating points been available, she would have

used four or five different rating scores to better differentiate the strengths of the question stems from each other.

The analyses of the mixed-effect models for individual moderators revealed that some variation in the means transformed by the Linear-Stretching method and Semantic Judgement of Word Value in Context method were accounted for by the year of administration, polarity, and question type. Nevertheless, none of the predictors were significant in explaining variation in the outcomes transformed by the Semantic Judgement of Fixed Word Value method.

These findings should be interpreted with caution, and cannot be used for real inferences about happiness. Most particularly, the items do not all represent real survey results. The setup of the four made-up questions has had some impact on the obtained results. These questions were made up to be relatively stronger than the rest of the questions in terms of the intensity of the degree of happiness (i.e., the stems included the words pleased, elated, delighted, and cheerful). Therefore, the original-scale means of these questions were set to be relatively smaller than those of the database questions. Consequently, after the transformation methods were applied, the transformed means of these questions not surprisingly appeared to be smaller than those of the database questions. The means transformed by the LS-2, SJFW, and SJWC methods for the made-up questions were 6.58, 7.31, and 6.45, respectively. For the database questions, the means were 8.28, 7.50, and 7.62, respectively. The mixed-effect models for question type also showed that weighted means of the made-up questions were smaller than those of the database questions.

In addition, the year of administration for the made-up questions was arbitrarily set as 2018. Thus the slope coefficient for year of the survey also turned out to be negative in the mixed-effects models for the outcomes, indicating that the mean happiness level decreased over

the years. The results might have been the opposite if the year for the made-up questions had been assigned randomly or set to be a different year than 2018.

Moreover, the scales of the made-up questions were all bipolar. The mixed-effects models for the scale-polarity predictor revealed that the weighted means of the bipolar scales were smaller than those of the unipolar scales for the outcomes of the LS-2 and SJWC. Again, the results might have been the opposite if the some or all of scales of the made-up questions had been created to be bipolar.

5.3 Implications and Limitations

My work is the first research study that provides meta-analysts with a framework demonstrating how to handle variations across survey studies in meta-analysis due to wordings of questions about the same construct, numbers of response options, response-scale labeling (endpoints labeled vs. fully labeled), scale polarity, and response-option labeling. In my examples I used response scales commonly used in the happiness literature. However, the methods described here can also be used for other latent constructs in education and psychology that are typically measured by self-report rating scales and survey questions, as well as for measures of other constructs in areas other than education. Meta-analysts can combine the findings of such survey questions on the same topic by using these methods, according to the specific objectives of their meta-analytic study. Consequently, by dealing with such survey diversity, meta-analysts will be able to integrate more findings in their analyses by including more diverse studies in their collections. Also, taking into account the variations due to the aforementioned factors may help the meta-analyst to understand observed differences in study outcomes and potential population differences. These practices will in turn enhance the

comprehensiveness of data collection, the validity of survey meta-analyses, and the generalizability of the results of survey meta-analyses.

Having said that, meta-analysts need to also take into account other methodological considerations when they attempt to summarize findings of survey data across studies. These considerations include the survey-administration mode, administration duration, survey language, use of complex survey-sampling designs, variance-estimation techniques (e.g., use of sampling weights), specifications of target populations and sub-populations, and the psychometric quality of the included surveys, just to name a few. These issues are not covered in this study. For interested readers, discussions of such important issues can be found elsewhere (Fox, 2011; Kish, 1999; Rao, Graubard, Schmid, Morton, Louis, Zaslavsky, & Finkelstein, 2008; Saris & Gallhofer, 2007).

APPENDIX A

IRB APPROVAL MEMORANDUM

Florida State
UNIVERSITY

Office of the Vice President for Research
Human Subjects Committee
Tallahassee, Florida 32306-2742
(850) 644-8673 · FAX (850) 644-4392

APPROVAL MEMORANDUM

Date: 07/19/2018

To: Ahmet Serhat GOZUTOK <[redacted]>

Address: [redacted]

Dept.: EDUCATIONAL PSYCHOLOGY AND LEARNING SYSTEMS

From: Thomas L. Jacobson, Chair

Re: Use of Human Subjects in Research
Critical Issues in Survey Meta-Analysis

The application that you submitted to this office in regard to the use of human subjects in the proposal referenced above have been reviewed by the Secretary, the Chair, and two members of the Human Subjects Committee. Your project is determined to be Expedited per 45 CFR § 46.110(7) and has been approved by an expedited review process.

The Human Subjects Committee has not evaluated your proposal for scientific merit, except to weigh the risk to the human participants and the aspects of the proposal related to potential risk and benefit. This approval does not replace any departmental or other approvals, which may be required.

If you submitted a proposed consent form with your application, the approved stamped consent form is attached to this approval notice. Only the stamped version of the consent form may be used in recruiting research subjects.

If the project has not been completed by 07/18/2019 you must request a renewal of approval for continuation of the project. As a courtesy, a renewal notice will be sent to you prior to your expiration date; however, it is your responsibility as the Principal Investigator to timely request renewal of your approval from the Committee.

You are advised that any change in protocol for this project must be reviewed and approved by the Committee prior to implementation of the proposed change in the protocol. A protocol change/amendment form is required to be submitted for approval by the Committee. In addition, federal regulations require that the Principal Investigator promptly report, in writing any unanticipated problems or adverse events involving risks to research subjects or others.

By copy of this memorandum, the chairman of your department and/or your major professor is reminded that he/she is responsible for being informed concerning research projects involving human subjects in the department, and should review protocols as often as needed to insure that the project is being conducted in compliance with our institution and with DHHS regulations.

This institution has an Assurance on file with the Office for Human Research Protection. The Assurance Number is IRB00000446.

Cc: Betsy Becker <bbecker@fsu.edu>, Advisor
HSC No. 2018.25208

APPENDIX B

IRB CONSENT FORM

FSU Behavioral Consent Form

Critical Issues in Survey Meta-Analysis

You are invited to be in a research study of dealing with diversity across survey questions on the same construct in meta-analysis. I ask that you read this form and ask any questions you may have before agreeing to be in the study. This study is being conducted by Ahmet Serhat Gozutok, doctoral candidate in the Department of Educational Psychology and Learning Systems at FSU.

Background information:

The purpose of this study is to shed light on how best to deal with diversity across survey questions on the same construct in meta-analysis.

Procedures:

If you agree to be in this study, I would ask you to do the following things:
You will be presented with a series of survey questions about happiness that differ slightly from each other in the wording of their question stem. You will be asked to rate the strength of the question stems. The purpose of rating the wordings of the question stems across different questions is to determine whether the strengths of the question stems sound similar. If not, I want to quantify the differences. Your task is to rate the strength of the wording of the question stems for each question, based on your personal judgment. The term "strength" refers to "intensity of degree of happiness". Your ratings will be used as a moderator variable in conducting a meta-analysis.
You will also be presented with various survey questions about happiness and their corresponding response-option labels. You will be asked to locate the response-option labels on a 0-10 scale given the specific wordings of the questions and their corresponding response-option labels. Your task is to assign each response-option label a position and corresponding number on the 0-10 scale. You will be asked to move a slider bar for each label by using a cursor until you feel that the position of the slider bar appropriately represents the degree of happiness denoted by the label.
Note that you are NOT asked to reveal your own levels of happiness, but rather to simply order and place verbal labels that indicate degrees of happiness on a 0-10 scale. Your input will be used to locate the verbal labels on a common 0-10 scale in a meta-analysis.
All tasks will be completed on a computer screen. You will be emailed a Qualtrics website link, which leads you to the tasks. Completion of the tasks will take about 15 minutes.

Risks and benefits of being in the study:

There is no anticipated risk or benefit to you for participating in this research.

Confidentiality:

The records of this study will be kept private and confidential to the extent permitted by law. In any sort of report I might publish, I will not include any information that will make it possible to identify you as a participant. However, research information that identifies you may be shared with the FSU Institutional Review Board (IRB) and others who are responsible for ensuring

compliance with laws and regulations related to research, including people on behalf of the Office for Human Research Protections.

Voluntary nature of the study:

Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University. If you decide to participate, you are free withdraw at any time without affecting those relationships.

Contacts and questions:

The researcher conducting this study is Ahmet Serhat Gozutok. You may ask any question you have now. If you have a question later, you are encouraged to contact him at 3204F Stone Building, Tallahassee, FL, 32306-4453, email: [redacted] and phone number: [redacted]. You can also contact his academic advisor Betsy Jane Becker at 3204F Stone Building, Tallahassee, FL, 32306-4453, email: bbecker@fsu.edu and phone number (850) 645-2371.

If you have any questions or concerns regarding this study and would like to talk to someone other than the researcher, you are encouraged to contact the FSU IRB at 2010 Levy Street, Research Building B, Suite 276, Tallahassee, FL 32306-2742, or 850-644-8633, or by email at humansubjects@fsu.edu

You will be given a copy of this information to keep for your records.

Statement of consent:

I have read the above information. I have been able to ask questions and have received answers to all questions I have asked. I consent to participate in the study.

_____ _____

Signature Date

_____ _____

Signature of Investigator Date

FSU Human Subjects Committee approved on 07/19/2018, void after 07/18/2019. HSC #2018.25314

REFERENCES

American Psychological Association Publication and Communication Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *The American Psychologist, 63,* 839–851.

Batz, C., Parrigon, S., & Tay, L. (2016). The impact of scale transformations on national subjective well-being scores. *Social Indicators Research, 129*(1), 13-27.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*. *50*(4), 1088–1101.

Bond Jr, C. F., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods, 8*(4), 406.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: Wiley.

Card, N. A. (2011). *Applied meta-analysis for social science research*. Guilford Press.

Cooper, H. M. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (5th ed.). Los Angeles, CA: SAGE Publications.

Duval, S. J., & Tweedie, R. L. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455–463.

DeJonge, T., Kalmijn, W., Veenhoven, R., & Arends, L. (2015). Stability of boundaries between response options of response scales: Does 'very happy' remain equally happy over the years? *Social Indicators Research, 123*(1), 241-266.

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

DeJonge, T., Veenhoven, R., & Arends, L. (2015). 'Very Happy' is not always equally happy on the meaning of verbal response options in survey questions. *Journal of Happiness Studies, 16*(1), 77-101.

DeJonge, T., Veenhoven, R., Kalmijn, W., & Arends, L. (2016). Pooling time series based on slightly different questions about the same topic forty years of survey research on happiness and life satisfaction in The Netherlands. *Social Indicators Research, 126*(2), 863-891.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*. *315*(7109), 629–634.

Eutsler, J., & Lang, B. (2015). Rating scales in accounting research: The impact of scale points and labels. *Behavioral Research in Accounting*, *27*(2), 35-51.

Fox, K. M. P. C. (2011). *A framework for the meta-analysis of survey data* (Doctoral dissertation).

Friedman, H. H., Cohen, D., & Amoo, T. (2003). Label or position: Which has the greater impact on subjects' responses to a rating scale? *Journal of International Marketing and Marketing Research*, *28*(2), 77-81.

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher, 5*(10), 3–8.

Hamby, T., & Levine, D. S. (2016). Response-scale formats and psychological distances between categories. *Applied Psychological Measurement*, *40*(1), 73-75.

Jürges, H., Avendano, M., & Mackenbach, J. P. (2008). Are different measures of self-rated health comparable? An assessment in five European countries. *European Journal of Epidemiology*, *23*(12), 773-781.

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

Kieruj, N. D., & Moors, G. (2010). Variations in response style behavior by response scale format in attitude research. *International Journal of Public Opinion Research*, *22*(3), 320-342.

Kish, L. (1999). Cumulating/Combining population surveys. *Survey Methodology, 22*(2), 129-138.

Krosnick, J. A. & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, L. Decker, E. de Leeuw, C. Dippo, et al. (Eds.), *Survey measurement and process quality* (pp. 141–164). New York: John Wiley & Sons, Inc.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (Vol. 2, pp. 263–314). Bingley, UK: Emerald Group Publishing Limited.

Lietz, P. (2010). Research into questionnaire design: A summary of the literature. *International Journal of Market Research, 52*(2), 249-272.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine, 20*(4), 641-654.

Rao, S. R., Graubard, B. I., Schmid, C. H., Morton, S. C., Louis, T. A., Zaslavsky, A. M., & Finkelstein, D. M. (2008). Meta-analysis of survey data: application to health services research. *Health Services and Outcomes Research Methodology*, *8*(2), 98-114.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, *118*(2), 183.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustment*. West Sussex, UK: John Wiley & Sons.

Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation and analysis of questionnaires for survey research.* New York: Wiley.

Schwarz, N., Knäuper, B., Hippler, H. J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, *55*(4), 570-582.

Veenhoven, R. (n.d.). *World Database of Happiness*, Erasmus University Rotterdam, The Netherlands. Assessed on May 2018 at: http://worlddatabaseofhappiness.eur.nl

Veenhoven, R. (2009). The international scale interval study: Improving the comparability of responses to survey questions about happiness. In *Quality of Life and the Millennium Challenge* (pp. 45-58). Dordrecht, Netherlands: Springer.

Veenhoven, R. (2014). World database of happiness. In *Encyclopedia of Quality of Life and Well-Being Research* (pp. 7257-7260). Netherlands: Springer.

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

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

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*. *36*(3), 1–48.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*(3), 236-247.

Yan, T., & Keusch, F. (2015). The effects of the direction of rating scales on survey responses in a telephone survey. *Public Opinion Quarterly, 79*(1), 145-165.

BIOGRAPHICAL SKETCH

Ahmet Serhat Gozutok completed his undergraduate study in Elementary Education program at Gaziosmanpasa University, Turkey in 2008. He worked as an elementary school teacher for a short time. He started the Measurement and Statistics program at Florida State University in 2011. He received the Master of Science degree in Measurement and Statistics in 2012. Following completion of his master's degree, he enrolled in the doctoral program in the Measurement and Statistics at Florida State University.

During his doctoral study, he has been a graduate research assistant in various research projects. He worked as a psychometrician in the Florida Department of Juvenile Justice. He also worked as a psychometric intern in the K-12 Assessment Office at the Florida Department of Education. His research interests include meta-analysis of survey studies. He is also interested in test development and psychometric analyses in large scale assessments.