Measurement Issues in Synthesizing Survey-Item Responses

Betsy Jane Becker and Ahmet Serhat Gözütok

Abstract From questions about politics to queries about candy preferences, survey items ask about matters large and small. While statistical approaches to combining survey estimates have been well studied, less attention has been paid to matters of measurement comparability when survey items are being summarized via meta-analysis. We present an overview of this problem. Meta-analyses begin with a defined problem, and relevant studies (here, surveys) are gathered. Studies and their measures should be scrutinized for validity and comparability as part of data collection and evaluation. When summarizing survey items, meta-analysts must represent item responses using indices that are comparable across surveys. However, survey constructs and the items that tap those constructs differ in diverse ways that challenge the meta-analyst. Cook's concept of "heterogeneous irrelevancies" supports the inclusion of diverse survey items in meta-analysis, but the tasks of construct definition and operationalization are key to a successful synthesis of items. Item variation arises from many sources—differences in construct definition, wording of question stems, direction and labeling of response scales, and number and labeling of response options. We describe approaches to dealing with these features using examples from the World Database of Happiness and raise cautions for various stages of the process.

Parts of this work are derived from Ahmet Serhat Gözütok's Ph.D. dissertation entitled "Critical issues in survey meta-analysis" conducted under the supervision of Betsy Jane Becker at Florida State University.

B. J. Becker (✉)
Measurement and Statistics, Educational Psychology and Learning Systems, College of Education, Florida State University, Tallahassee, FL, USA
e-mail: bbecker@fsu.edu

A. S. Gözütok
Measurement and Evaluation in Education, Educational Sciences, Ereğli Faculty of Education, Zonguldak Bülent Ecevit University, Zonguldak, Turkey
e-mail: gozutok@beun.edu.tr

H. K. T. Ng, D. F. Heitjan (eds.), *Recent Advances on Sampling Methods and Educational Statistics*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-14525-4_6

1 Overview

Surveys are ubiquitous. From polls of political leanings to academic inquiries (Fanelli, 2009) to frivolous studies of candy or soda preferences (e.g., RetailMeNot Editors, 2021), survey items ask us about matters large and small. While statistical approaches to combining quantitative results of surveys have long been of interest (e.g., Kish, 1994, 1999a; Morton, 1999), less attention has been paid to matters of measurement comparability when surveys or survey items are combined in meta-analyses. An early exception to this was Kish's (1999b, p. 131) concern over the measurement challenges faced in cumulating surveys multi-nationally. Of late the scholarship on harmonization of measures has attacked this same problem.

Meta-analyses (Glass, 1976) have the goal of summarizing the "typical" outcome of a set of studies or surveys in terms of strength, direction, and consistency of the findings. In this chapter, we present an overview of conceptual and measurement considerations underlying the synthesis or meta-analysis of survey items, and then briefly characterize the set of techniques called harmonization. We review four survey-item features that impact the quantitative synthesis of survey items. Writings on test validity, item construction, and psychometrics guide this work. To illustrate these ideas, we draw on the World Database of Happiness project by Veenhoven and colleagues (e.g., Veenhoven, 2015; Veenhoven et al., 1993).

2 Introduction to Survey Synthesis

Most of the research to date on cumulating survey results has focused on the nature of the populations to be combined and how their results should be statistically weighted. Kish (1999b) argues that the presence of national surveys (which expanded greatly in the late 1940s) led international entities such as the various agencies of the United Nations to make international comparisons, even when those might not have been statistically justifiable. This growth in cross-national work was followed by many statistical developments including the deliberate design of coordinated national-level studies, derivation of methods for post-stratification weighting, and proposals for new varieties of periodic sampling plans. Kish led the field in this arena, and in 1994 presented five types of multi-population survey designs, based on seven aspects of design. The first three of these aspects relate at least in part to measurement, which is our focus. Kish (2002) later pointed out the connections between his ideas on quantitatively cumulating surveys and meta-analysis—the enterprise of combining studies.

The early focus on statistical analyses for combinations of related surveys may have resulted in part from the fact that early syntheses of surveys estimated parameters based on identical or very similar survey questions. There was little need to consider the nature of the questions asked, avoiding many conceptual and measurement components of the synthesis process. However, such a focus necessarily leads to a narrowed selection of constructs and measures of those constructs compared to what may be seen in the broader literature.

We discuss two classes of approaches to measurement challenges in meta-analysis of surveys. One includes conceptual approaches that deal with the theoretical constructs per se and aim to formalize the meaning behind constructs of interest. van de Water et al. (1996) refer to this as conceptual harmonization. Second are statistical or psychometric approaches that primarily operate on item scale points, distributions of scores, or correlations among items that aim to measure constructs of interest. Properly covering either of these classes of approaches would require a book rather than a book chapter, so we cover only the main aspects of these approaches.

3 The Process of Meta-analysis

Meta-analysis involves the systematic collection of the results of series of related studies, and the eventual quantitative analysis of those results. The process of meta-analysis has components that parallel those of primary research (Cooper, 2017). A simple version of Cooper's steps includes

1. Problem formulation,
2. Literature search,
3. Data evaluation, including representation of study findings,
4. Data analysis,
5. Interpretation of synthesis results, and
6. Public presentation.

We focus on steps 1 and 3, because measurement issues arise primarily at these points.

3.1 Step 1: Problem Formulation for Survey Synthesis

In a typical meta-analysis, a detailed question guides the synthesis process. Meta-analyses often examine the efficacy of interventions, or strengths of relationships. A rationale should be developed for the specific question(s) asked. A successful meta-analysis is based on questions that are not so broad as to be unanswerable, or so narrow that few studies (here, few surveys) address them. In applying this consideration to the synthesis of survey items, we argue that the development of a clear question is critical so that the synthesis team does not end up with a near-infinite set of survey sources, each examining a variation of the true target topic. For example, the World Database of Happiness (WDH; found online at https://worlddatabaseofhappiness.eur.nl) has a bibliography with over 15,500 publications on the topic of happiness, and almost 23,000 distributions of responses to questions on happiness from all over the world. Such a collection of results would swamp an individual meta-analyst; that is why over 100 team members have participated in the accumulation of these results since the 1980s.[1]

The meta-analyst must specify appropriate population(s) of study, develop construct definitions, and delineate an acceptable set of operationalizations of those constructs. The process often begins with an examination of past reviews and existing research; in some fields, scoping reviews (Munn et al., 2018) provide a quick look at the extent of the literature. The creation of lists of keywords and definitions of central concepts are important tasks, as is deciding on the target populations for study, because some constructs will differ by the age, gender, or nationality of respondents.

A key part of problem formulation is to identify the constructs to be studied as the independent and dependent variables. Examination of relevant theories, brainstorming with experts in the field of interest, and use of qualitative research methods such as grounded theory (Wolfswinkel et al., 2013) may help at this stage. Even an idea as simple as happiness may vary in its meaning across cultures (Ye et al. (2015)) or age levels. When several related constructs are of interest (e.g., happiness and life satisfaction in the WDH), the meta-analyst should justify decisions to combine results across those constructs. We posit that the use of frameworks similar to the "blueprints" used in test construction can help outline the components of target constructs (e.g., content focus, behaviors that evidence presence of the construct) and guide collection of desired items. For example, a synthesis on political interest might contain items tapping both engagement/interest and active participation, for different levels of political activity such as local, state, and national campaigns.

Good problem formulation facilitates the creation of inclusion and exclusion rules that help identify appropriate sources of data to address the question of interest. Because aggregate-data meta-analyses synthesize results from completed primary studies, the problem-formulation stage differs from the planning stages of a single survey, or even a multi-site survey program. In a typical survey, measures of the desired construct are developed prior to the survey's administration. In contrast, in meta-analyses one works with existing measures, be they scales or single items. The meta-analyst may aim to gather information about a particular construct only to find it has not been sufficiently studied. For example, in a synthesis of the literature on the management of type 2 diabetes, Brown and colleagues (2016) found that few studies had measured compliance with keeping doctor's appointments.

[1] It should be noted that the WDH is not meant to serve as the source of documents for a single survey synthesis.

3.2 Step 3: Data Evaluation

Efforts to bring together information across independent studies of any kind (including surveys) bring attention to the fact that individual study authors and survey designers have generated a huge diversity of measures on the same or similar topics. This diversity leads to challenges when results are to be accumulated. While any number of authors have commented on the importance of dealing with measurement issues in combining survey results (e.g., Rao et al., 2008, p. 102; Schenker & Raghunathan, 2007, p. 1809), few provide complete solutions for the measurement challenges inherent in the process of combining surveys.

One expects a degree of diversity in outcome measures and study design across studies because the process of science (and the need to publish "new" research) pushes toward uniqueness. Diversity in measures across studies (or surveys) may be great due to differences in construct definitions, or minimal, if construct definitions, item wording, and responses options are similar. Cook (1993) has pointed out that a degree of diversity in measures of a construct can support generalizations. If a varying feature of a set of items, say, strength of item-stem wording, does not relate to how the items function (i.e., to respondent behavior), it tells us that feature is irrelevant to the construct measured. In our example, the meta-analyst would generalize across items of varying strength if wording strength does not relate to response patterns. It is important to identify potential item features at the data-evaluation stage for this reason.

In a typical aggregate-data meta-analysis, reviewers appear to rely on the primary-study authors' claims about what was measured. It is rare to share the exact instruments used in published studies. Also researcher-made measures are often used; these are nearly impossible to obtain. Thus, a great deal of trust, or perhaps mystery, can be involved in construct definition and operationalization in a typical meta-analysis.

In contrast when individual survey items are to be summarized, the exact words used in those items and their response options are obvious. However, this does not remove the necessity for the meta-analyst to assess the nature of the construct(s) tapped, and to ask whether items from different studies measure the same construct. This can be done by way of typical validity-study methods, such as by having experts examine all collected items and rate each on its centrality to the construct of interest and degree of match to the concept definition developed at step 1. These are discussed further below. Only after the constructs are clear should the meta-analyst proceed to the next step of trying to find mathematically sound ways of connecting or comparing the item responses.

4 Harmonization

In part due to this diversity, calls for coordination and (post hoc) harmonization to enable researchers to bring together diverse measures have become more frequent over time, even though this process is rarely reported (Griffith et al., 2015). Early efforts have centered on harmonization as standardization, that is, on creating comparable numerical scoring systems for variables of interest. We refer to this as statistical harmonization, in contrast to conceptual harmonization, discussed above as part of problem formulation. A search of all ProQuest databases for "harmoniz* and measures" in peer-reviewed article titles suggests that attention to harmonization of measures first appeared in the 1990s, setting aside articles on harmonization of physical/scientific indices such as pH and blood counts (Lewis, 1990), or currency-related indices (e.g., Goeltz, 1991). The term harmonization may have grown out of the extensive work on harmonization of laws, regulations, and social policies (e.g., in the European common market), such as in Holloway and Collins (1982) and many other sources.

Initial efforts aimed to make measures of demographics and socioeconomic status more comparable. These were led by ESOMAR—the European Society for Opinion and Marketing Research—which was motivated to aid market researchers (and obviously other commercial entities) to identify "the true diversity of the market place" (ESOMAR, 2003, p. 97) in Europe. Work concerned with the harmonization of measures of human social and cognitive constructs first appeared in the literatures on commerce (e.g., Quatresooz & Vancraeynest, 1992, on demographics) and medicine.

Citations on harmonization grew in the early 2000s as attention was drawn to health-related measures such as activities of daily living that might differ across countries (Nikula et al., 2003) and to measures used in the Health and Retirement Study (e.g., Angrisani & Lee, 2011) and other cross-national health surveys that followed. Interest in the standardization and simplification of measures serves the goals of such cross-national survey efforts. Having similar or related measures and scoring systems facilitates cross-national comparisons (e.g., Bech, 1992, on quality of life), and connections among measures enable researchers to administer fewer measures, thus saving time, money, and the goodwill of participants.

Harmonization may be conducted on measures meant to be of the same construct, or measures of related constructs. The goals of harmonization include avoiding duplication of measurement efforts and ensuring standardization and comparability of measures across different target populations.

4.1 What Is Harmonization?

Harmonization is a process of making definitions and measures of a common construct or variable comparable. While many writings on harmonization focus on the translation of numerical scores to compatible scales, we argue that harmonization must involve two components—a conceptual reckoning and clarification of what evidence is suitable to represent the construct, or conceptual harmonization, and a statistical or psychometric component that accounts for how measures of the construct have been scored. The fourth principle of the National Quality Forum (2010) states that the conceptual component should precede the decision on whether to try to statistically harmonize.

4.2 Conceptual Harmonization

The first step in this process must be consideration of the target concepts of the synthesis and any theoretical frameworks that may underlie the measures at hand. The National Quality Forum argues that harmonization must account for the population or populations to whom measures will be administered as well. This is consistent with the idea of consequential validity of any test or measure from the *Standards for Educational and Psychological Testing* (Joint Committee on the Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, 2014). The measure should be appropriate for all populations to whom it will be administered.

Another relevant concept drawn from the Standards is that of concept underrepresentation—the idea that a measure taps into "less or more than its proposed construct" (p. 12). Assessment of concept representation would be facilitated by the use of a survey blueprint, as we describe above. Many constructs encompass a broad range of measures and diverging conceptualizations, for example, activities of daily living (ADL; Pluijm et al., 2005) and quality of life (QOL; Bech, 1992). Some fields have moved toward the use of common or "core" measures, but we argue that in syntheses, key benefits also arise from diversity in instrument use. Cook (1993) has noted that having a diverse set of measures (or study designs, or populations) assists with generalizability. In particular, if such differences are not associated with study outcomes, our findings can be stated unconditionally. Finding empirical evidence that those features do not matter means that simpler but broader conclusions can be stated. If we narrow our constructs or measures to a select few, we cannot even assess how generalizable our results might be.

Some researchers have provided conceptual models for these variously measured constructs or used frameworks such as classifications of measures (e.g., of attitudes) into affective, cognitive (Crites et al., 1994), and behavioral aspects (Ostrom, 1969). Often theoretical models can assist the meta-analyst in judging whether single items or longer measures "fit" a construct definition. For example, Bech (1992) described a model for health-related quality of life based on six diagnostic components: physical, cognitive, affective, social, economic, and ego-function aspects (PCASEE). Bech's Table 1 relates the six components to specific variables such as sleep (P), concentration (C), depression (A), and so on. However, for others quality of life may be represented in terms of self-assessments: Joyce et al. (1999) listed the individual's assessment of subjective health as the manifestation of Bech's physical aspect, their decision-making capacities as representing cognition, warm feelings toward others as an index of affective QOL, and so on.

Remarkably, though papers can be found with personal definitions of happiness and well-being provided by Veenhoven, few that we examined connect to other scholarship, and many are overly glib and simplistic about defining happiness.[2] Veenhoven (2007, p. 3) states that "'well-being' denotes that something is in a good state." Veenhoven (2009) stipulated that happiness and quality of life and well-being are the same. Further musings dissect quality of life into components, but do not tie the ideas to other literatures that conceptualize or theorize on these constructs, and those literatures are each extensive. Veenhoven (2009) is the most thorough, though its provision of evidence is haphazard, with little attention to the different populations and cultural groups that appear in the database. One finds the conceptual basis for harmonization could be stronger in the happiness realm.

Examination of items is another component of conceptual harmonization. For example, Wang et al. (2014) provide a compendium of items tapping health-related behaviors across a set of surveys that aimed to be comparable to the US Health and Retirement Study (HRS). These reveal the vast array of questions asked with content deemed "similar enough" to be harmonized. Lengthy concordance tables are given for items on smoking, drinking, and physical activity.

Chen et al. (2021) conducted a similar process that they called "pre-statistical harmonization" which involved close inspection of all items, reviews of scoring procedures, and inspection of populations assessed in surveys of behavioral symptoms of dementia. Individual patient data from eight samples allowed them to also conduct statistical harmonization, including psychometric analyses using item response theory, model-fit analyses, and examination of inter-item correlations and cross-tabulations. Experts in the content matter at hand would be critical to this process.

Given sufficient data, one could conduct empirical validity studies of collected items such as investigations of inter-item correlations or factor analyses. However, if each study contributes only a few items to the collection, such studies would require additional new primary data. Seemingly sensible quantitative analyses should be preceded by conceptual harmonization. Even when harmonization is a reasonable goal, surveys intending to include comparable measures may still show notable variation in wording and content. Some of these features are described in the following sections.

[2] Other papers in Veenhoven's extensive writings may present more thorough analyses of relevant theories and evidence.

4.3 Item Features Critical to Harmonization

We next consider four important survey-item features that impact the synthesis of survey items. Findings on item writing and psychometrics guide this work. These features are candidates for coding because they may relate to the responses of participants and may also play a role if statistical harmonization is to take place. We discuss:

- Variation in wording of the question stem or statement
- Number of response-option categories
- Scale direction: Unipolar vs. bipolar scales
- Nature of response options: Labeling and wording of response options

Other features may play roles in cross-survey variation in items (e.g., the use of negatively and positively worded items, per Pilotte and Gable (1990), among many others), but we believe these four features are most important and are moderately easily addressed.

4.3.1 Wording of Item Stems

Differences in the wording of item stems can result in variation in the focus and the strength of the questions asked, and affect the strength of responses from recipients. The impact of wording in surveys is parallel to the way that test-item wording and stem complexity affect difficulty in standard educational exams (e.g., Ascalon et al., 2007). More extreme stem statements are expected to be harder to endorse. Schuman and Presser (1996) discuss various aspects of wording including intensity, centrality, and tone in several chapters in their classic book; there are too many to properly cover here. Additionally, the exact wording differences that are important will surely vary from one meta-analysis to another, so we mention here only the general idea and its importance.

A multifaceted example of stem differences was reported in the RAND project to harmonize measures of health behavior in the elderly. Examining ten different longitudinal studies of aging, Wang et al. (2014) noted that questions about drinking alcoholic beverages varied in the time frames they asked about, and the amounts and specificity of beverages consumed. Items on the frequency of drinking asked about time ranges including the last 7 days, last month, last 3 months, last 6 months, and last year. Some asked about consumption of “any alcohol,” whereas others differentiated wine versus beer versus spirits, and the most focused questions probed for consumption of “normal beer” versus strong beer, or listed specific types of liquor (e.g., wine, beer, and whiskey). Cross-national comparisons are difficult when particular beverages are more popular and readily available in their country of origin (e.g., sake, soju, makgeolli); such wording matters are idiosyncratic but may be key to understanding a particular population.

As an example, suppose we want to summarize information on the portion of the elderly population that is engaged in heavy drinking. One might expect the two features of time frame and amount of alcohol consumed per unit time to work together to represent total alcohol consumption. Thus, it would be important for the meta-analyst to capture such features during data extraction. Some meta-analyses have used coded variables to create additional variables. For example, multiplying the number of treatment occasions for an intervention by the typical session length provides a total-exposure-to-treatment measure. A similar approach could be taken for assessing alcohol consumption. To synthesize data on items that do not allow for similar computations, the meta-analyst could ask expert raters to assess the strength of the item-stem statements and use those assessments as moderators of diversity in the responses.

4.3.2 Number of Response Options

Another feature of item responses that leads to between-surveys variation is the number of response options. Differing numbers of options are the leading reason for using the statistical harmonization methods described below, because the scales of item scores correspond to the number of options available. In some cases, the options are nominally or qualitatively different and for those, scale changes are not needed. In others item responses represent an implied underlying continuum. This distinction has implications for the choice of quantitative approaches to scale harmonization. It is tempting to separate dichotomous items from items with three or more options, but when an item measures an underlying continuum, the difference between two and three or more options is simply one of the granularities of responses. Often surveys require respondents to reply using ordinal scales with varying numbers of categories. Others may allow for continuous responses, for instance, by placing a mark on a line. In any case when synthesizing findings from individual items, the meta-analyst may want to consider the number of options as a potential moderator of between-items differences in results, as there is no way to add response options after the fact.

To enable sensible comparisons, scale conversions have been used to rescale individual item outcomes for ordinal- and interval-scale items, to locate them on a common metric. Many of these are listed in handbooks for statistical harmonization (e.g., Griffith et al., 2013), and we show several examples below. At its simplest, rescaling may entail collapsing responses or applying simple linear transformations of scale points; however, these can lead to potentially idiosyncratic translations across items. More complicated approaches may require assigning labeled points differently across studies, or adopting more sophisticated latent variable models which involve more assumptions and computation (e.g., van den Heuvel et al., 2020).

4.3.3 Nature of Response Options

Responses to items may be partially or fully labeled and may be graphical, numeric, or verbal. When survey-item responses are verbally labeled, the nature of the responses available as answers must be considered. Response-option-label differences have been the focus of efforts to harmonize measures across surveys, especially in the work of Veenhoven and his team (DeJonge et al., 2017), and can be very tricky when different formats appear across studies (e.g., what words are associated with the array of frowning faces on pain-scale items?).

Differences in response labeling are presumed to lead to different choices by participants, and are known to vary across different surveys, contributing to further between-surveys variation. Thus, this feature is one that should be coded or characterized by the meta-analyst. When a finite number of categories is offered, response options may be fully or partially labeled; this has long been noted to affect the reliability of responses (Endig, 1953). Similarly even when respondents are offered a continuum to mark, labels may be specified at different points along the length of the response line, and continuous-looking scales may be assigned integer scores by survey software. Coding these features enables the meta-analyst to empirically assess whether such variations in design affect participant responses.

4.3.4 Item Polarity or Direction

Some surveys use items with responses organized along continua with endpoints that are meant to be opposites; these are bipolar items. For example, ratings may range from "happy" to "unhappy" (or if endpoints have modifiers, "very happy" to "very unhappy," etc.). In contrast unipolar item responses may run from "not at all happy" to "very happy," with no coverage of the range representing degrees of unhappiness. This approach to labeling may reflect a potential belief that, say, happiness and unhappiness reflect two separate dimensions, rather than two ends of a continuum. It may be possible to link similar option choices across items of different polarities if verbal labels are assigned to all items.

To combine unipolar and bipolar items presents a challenge to the meta-analysis. Using simple linear transformations (e.g., that move responses to a common scale) will not address the fact that the endpoint of a unipolar scale may correspond more closely to the middle of a bipolar scale than to its end. This is why numerical transformations cannot be blindly applied without consideration of the constructs per se that are tapped by individual items. At the very least, the polarity of items must be coded, and it may be sensible to separately analyze items with different polarity, unless a translation can be found that soundly matches response options across these two item types.

4.4 Statistical Harmonization

Hofer and Piccinin (2009) of the Integrative Analysis of Longitudinal Studies on Aging (IALSA) project have pioneered the idea of harmonization within the study of the psychology of aging. They and others have provided a guide to statistical harmonization aimed largely for use in individual participant meta-analyses (Griffith et al., 2013). Many approaches to this statistical harmonization, along with supportive software (e.g., Adhikari et al., 2021; Fortier et al., 2011; Winters and Netscher, 2016), have been developed. We touch on some of the most common methods here and discuss their weaknesses.

4.4.1 Linear Transformations

A conventional method to locate item responses on a common scale is to place the scale points of the diverse items (i.e., responses to individual items) onto a common secondary scale by applying linear transformations. These date far back (Hull, 1922) and have numerous instantiations. The simplest linear transformations include linear stretching that converts primary-scale response-option scores to a common scale running between prespecified endpoints (e.g., 1–10) and standard linear transformations that shift scores to a scale with a known mean and standard deviation (SD), such as the T score with a mean of 50 and SD of 10, or the well-known z score.

4.4.2 Linear Stretching

If the number of response options of a primary scale is smaller than that of the common scale, the transformation is done by linearly stretching the scale points from the smaller scale onto 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 (e.g., from a 10-point scale to 5-point scale).

The equation below can be used to stretch or shrink the scale points from one scale X to another, say Y:

$$Y = \left[(X - \min(X)) \times \frac{\max(Y) - \min(Y)}{\max(X) - \min(X)}\right] + \min(Y), \text{ or}$$

$$Y = \left[X \times \frac{\max(Y) - \min(Y)}{\max(X) - \min(X)}\right] - \min(X) \times \frac{\max(Y) - \min(Y)}{\max(X) - \min(X)} + \min(Y), \quad (1)$$

where X is a scale point of the original item; $\min(X)$ and $\max(X)$ are the minimum and maximum possible scale points of that item (not the observed min and max values), respectively; and $\min(Y)$ and $\max(Y)$ are the analogous values on the transformed scale (Card, 2011). This is easily seen to be a linear transform $Y = a + bX$ where

$$a = -\min(X) \times \frac{\max(Y) - \min(Y)}{\max(X) - \min(X)} + \min(Y) \text{ and } b = \frac{\max(Y) - \min(Y)}{\max(X) - \min(X)}.$$

4.4.3 Linear Transformation to a Target Mean and Variance

An obvious second transformation assigns a new mean (say μ_Y) and variance (σ_Y^2) to the scale responses. This is attained by using the linear transform $Y = a + bX$, with

$$a = (\mu_Y S_x - \bar{X}\sigma_Y)/S_x \text{ and } b = \sigma_Y/S_x.$$

Target values are denoted using Greek characters to distinguish from the sample values for the original scales.

4.4.4 Assumptions

Once the scale points across a set of items are on a common scale, a meta-analyst can directly summarize their results across studies. The assumptions of linear translations are that response options are equidistant (i.e., interval scaling), that the most extreme possible responses on all items should be scored as min(Y) and max(Y), and that identical verbal labels do not need to be assigned the same numerical value across items or surveys. Griffith et al. (2015) state without support that these methods also assume normality (we doubt this condition is needed), but rightly note that such translations may run into trouble if the measures have very non-normal distributions (e.g., skewness due to ceiling effects). Indices used may include mean scores if one is willing to assume an underlying continuum, or proportions of participants scoring above (or below) a set cutoff on the new scale.

Zumbo and Woitschach (2021) endorse the concerns we raise and have raised several additional concerns about the family of linear transforms by examining a more stringent mathematical formalization. These authors concur with our assessment that when a scale is fundamentally only ordinal, simplistic translations cannot magically change that fact.

4.4.5 Example

We demonstrate using two survey questions measuring a respondent's degree of happiness, taken from the World Database of Happiness (Veenhoven, n.d.). The first question is "In general, how happy would you say you are these days?" Its response scale runs from 1 to 7 with response-option labels Not happy at all (1), Very unhappy, Somewhat unhappy, Neither happy nor unhappy, Pretty happy, Very happy, and Extremely happy (7). The second question is worded slightly differently: "How happy do you feel as you live now?" This question has four response-option labels scored from 1 to 4: Very unhappy, Not too happy, Pretty happy, and Very happy, respectively. We use the linear-stretching method to transform the ratings of both scales to a common metric running from 0 to 10. After applying this method, the ratings for response labels of the first question become 0, 1.67, 3.33, 5, 6.67, 8.33, and 10, respectively, and those of the second question are 0, 3.33, 6.67, and 10, respectively.

Once the original ratings are linearly transformed to the same metric, the observed means ($\hat{\mu}_y$) and variances (S_y^2) can be calculated for each transformed scale with an underlying continuum by entering the transformed ratings y_j and the proportions of respondents choosing each of the ratings ($P(y_j)$) in the following equations:

$$\hat{\mu}_y = E(Y) = \sum_{j=1}^{J} y_j P(y_j), \tag{2}$$

$$S_y^2 = \sum_{j=1}^{J} (y_j - \hat{\mu})^2 P(y_j), \tag{3}$$

where we sum across all J response categories, $j = 1$ to J. Alternatively, one may calculate the linearly transformed mean and standard deviation directly from the means and standard deviations of the original scales (Kalmijn, 2010). Specifically,

$$\hat{\mu}_y = \left[\hat{\mu}_x - \min(X)) \times \left(\frac{\max(Y) - \min(Y)}{\max(X) - \min(X)}\right)\right] + \min(Y), \tag{4}$$

$$S_y = S_x \times \left(\frac{\max(Y) - \min(Y)}{\max(X) - \min(X)}\right). \tag{5}$$

The linear-stretching method has a few drawbacks especially when used to transform items with verbal response labels. One is that the meta-analyst must assume equidistance between response categories of all the scales being transformed (i.e., assume they have interval-scale properties). This assumption implies that the differences between successive response-option categories are the same within the old and new metrics. For the first question in our example, for instance, the difference in the degree of happiness between "Pretty happy" and "Very happy" should be the same as the difference in the degree of happiness between "Very happy" and "Extremely happy". This is impossible to verify as the measure is inherently ordinal even if the underlying construct is not.

The other issue is that the transformation does not consider either the verbal anchoring of response options or any differences in strength or focus of the question stems. For instance, after linear stretching has been applied, the score of 3.33 on the common 0–10 scale is associated with both the “Somewhat unhappy” verbal label of the first scale and “Not too happy” of the second scale, which is unsatisfying, and implies that they have the same meaning. In contrast at the top end of the scale, 6.67 is assigned to “Pretty happy” on both items, but “Very happy” is assigned 8.33 for the 7-point item versus 10 for the 4-point item. This lack of correspondence is both problematic and confusing.

4.4.6 Nonlinear Transformations

Two nonlinear and nonmathematical transformation approaches can be used to avoid these issues associated with the linear-transformation method. Both approaches are based on using subjective judgments of individuals (e.g., coders or expert judges) to determine corresponding values for possible response-option labels on a secondary numerical scale. The use of raters or coders to rate prompts (e.g., Eagly & Carli, 1981) or other study features (e.g., quality features; Atkins et al., 2004; Guyatt et al., 2008) is common in meta-analysis, and the use of ratings as moderators of study effects is also common. Here raters are asked to evaluate the semantics of the response-option labels.

In one approach, coders are presented with a list of all response-option labels from the primary scales and are asked to assign values to each label on a second common metric that is bounded by predetermined values. In the other approach from Veenhoven et al. (1993), coders evaluate the semantics of each response-option label after reading the question stem and the other response-option labels of the item.

In the first approach, which we call the semantic-judgment-out-of-context method, each response option is rated irrespective of the other response-option labels of the original item and its relative position on the common scale that contains all other response-option labels. This approach does not take into account the differences in the wordings of the question stems or the number or nature of the response options of the items. A weakness of this approach is that differences in question stems may impact how judges interpret the response-option labels being rated. Also, the semantic intensity of a response-option label may be interpreted differently depending on whether it is presented with many other options (e.g., seven categories) or few (e.g., three categories).

One early application of this approach was as the first step in a longer process of scaling items proposed by Jones and Thurstone in 1955. Respondents rated the semantic strength of 51 phrases (response options) used to indicate like or dislike of various foods (e.g., strongly like, tasty, bad). Each phrase was presented independently with no stem (i.e., no prompt of a specific food), and the raters assigned to each an integer value between -4 and 4, inclusive. Scale endpoints were labeled with “Greatest Dislike” and “Greatest Like” and the midpoint (0) was labeled with “Neither Like Nor Dislike.” The authors then applied a modified version of the successive-intervals scaling method (Edwards, 1952) to determine the values of the phrases on the common scale. Consequently, all the phrases were placed on an interval where a neutral label has the value 0.

With the second approach, the semantic-judgment-in-context method, coders assign values to each response-option label considering all aspects of the original item, including the relative position of the label, other response-option labels, the number of response options, and the wording of the question stem. Each individual question and its corresponding response-option labels are presented to the coders separately. For each response option, the coders assign a value that they consider the most appropriate on a secondary scale, say running from 0 to 10. One coder might assign 1 for the option label “very unhappy,” 3 for “unhappy,” 6 for “happy,” and 9 for “very happy.” Another coder could rate the same labels differently. Also, coders may assign different values to the same response-option label when it is presented in the context of different questions (e.g., with different question stems and accompanying response-option labels). The final ratings for each item’s response options are computed by averaging assigned values across coders. Consequently, response-option labels have unique values specific to each survey item on a set secondary scale.

In the extensive project on happiness, Veenhoven et al. (1993) used this approach to place values of the response options of nine survey items tapping happiness on a common scale. The items had almost identical question stems but differed in numbers and labeling of response options. Ten content experts evaluated the semantics of response-option labels by assigning values on a 0 to 10 scale. The means of those values across experts determined the final ratings of the response options. If a response option appeared in multiple items, its ratings were also averaged across items. Response options used only once retained their original rated values. Consequently, the authors came up with a single common scale having all possible options.

A concern with this approach is that it damps down spread that may be explained by other item features. Consider an item with response-label ratings that all exceed the ratings of the same labels when used with other items. Because the mean response score will replace those higher-than-average ratings, the process has the feature of moving extreme labels closer to the center, and in theory could reorder verbal labels within items, especially if ratings adjacent to the focal option are numerically close.

4.4.7 Comparisons of Approaches

Gözütok (2018) made use of survey items and their descriptive statistics (i.e., original means, standard deviations, proportions of respondents on response options) from Veenhoven’s World Database of Happiness collection to illustrate how three scale-transformation methods can be used in conducting a meta-analysis. The original response options of items were transformed to a secondary numerical scale running from 0 to 10 by the transformation methods described above. Then means and standard deviations of the items on the new scale were computed. The means obtained from the transformed scales were treated as study outcomes in three hypothetical meta-analyses based on raw-means synthesis (Bond et al., 2003).

In the three pseudo-meta-analyses, Gözütok (2018) included the wording of the question stems as a moderator variable, along with other survey-item characteristics such as number of response-category options, scale polarity (i.e., unipolar vs. bipolar scales), and scale labeling (i.e., endpoints labeled vs. all points fully labeled). To capture differences in the wording of question stems, he used ratings of the strength of the statement or question about the construct (i.e., happiness). This rating task may be done by meta-analysis coders, content experts, or a sample of target respondents. For example, coders may assign a higher rating of strength to the question stem "Do you feel elated?" and give a lower value to "Do you feel happy?". If so, part of the potential between-studies variance in the study outcomes will be explained by the differences in question-stem strength. The strength ratings did not relate to the mean happiness ratings in these pseudo-meta-analyses, possibly because the actual items on happiness were very similar (i.e., there was little between-items variation in wording). Also concern was raised due to an idiosyncratic rater Gözütok identified in his rater pool.

As part of the World Database of Happiness project, a great deal of work regarding the comparability of survey items of the same construct has been done by Veenhoven and his colleagues. Veenhoven (2008) reported on the International Happiness-Scale Interval Study. Its participants provided interval boundaries on a 0 to 10 scale for each response-option label of a set of country-specific happiness items. Each item stem plus its associated set of J response labels was presented to the participants. A web-based tool called the Scale Interval Recorder (Veenhoven & Hermus, 2006) allowed the participant to slide $(J - 1)$ markers that defined the boundaries between the J verbal labels that were associated with each question. Midpoints of the resultant summarized intervals were used to represent each response option.

This study inspired further innovations such as the continuum approach of Kalmijn (2010), and the reference-distribution method from DeJonge et al. (2014). The continuum approach postulates a latent happiness variable in the population. It is assumed to be continuous and was set arbitrarily to have scores in the interval [0, 10]. Beta-distribution shape parameters that best match the observed data are then found. Kalmijn recommends using a beta distribution which is left-skewed to reflect high levels of happiness in the population.

The reference-distribution method builds heavily on other harmonization methods in that it aims to define intervals to represent ranges for response options. In the reference-distribution method, the boundaries between the response options of the primary scale are derived from a reference distribution instead of from ratings by judges on a scale-interval recorder. Again this approach assumes an underlying latent distribution and uses the beta distribution that fits best to the survey results of the responses of a given sample and item. For interested readers, details and implications of these approaches can be found in DeJonge et al. (2017).

Other methods have been proposed to harmonize data from different survey items. Griffith et al. (2015) provided a summary of statistical approaches used in systematic reviews of cognitive variables. Transformation and other score-conversion techniques were common, but these authors argue for more sophisticated latent variable techniques including factor analysis and item response theory that would be difficult to achieve without individual participant data.

5 Conclusion

Methods and software tools have been developed to support the practical task of harmonization, but few focus on the detailed conceptual components that we argue are crucial. One exception is Fortier et al. (2017) whose step 2a concerns variable definition. Also on the whole, the practice of statistical harmonization remains simplistic. Griffith et al. (2015) and Zumbo and Woitschach (2021) have argued for the use of latent variable modeling in statistical harmonization, which is consistent with mainstream work in measurement and assessment, yet it is rarely used. The problem for the meta-analyst is that unless extensive individual-level data are available, these analyses cannot be conducted. For example, few surveys have simultaneously administered more than one or two of the hundreds of items in the Happiness Database. Also, such approaches assume that the first step of conceptual harmonization has occurred and identified a set of measures worthy of calibration and linking. It is unclear how often this has been done.

One possible route for meta-analysts, though a labor-intensive one, would be to include in the meta-analysis what are called bridge studies (e.g., Perie et al., 2005, which examined changes in the test structure of the National Assessment of Educational Progress or NAEP). The goal would be to map different items onto one calibrated scale. After conceptual harmonization of target items, the meta-analyst would administer those survey items to a new sample from the population of interest. When one considers the massive numbers of highly similar items in the World Database of Happiness, the task seems impossible. However, if multiple subsamples responded to smaller structured subsets of items (e.g., using incomplete blocks designs as have been used in NAEP and other large-scale assessments), a set of calibrations with common anchor items could allow for various forms of equating or linking (Kolen & Brennan, 2004) to be applied. Various designs for effective linking already exist. This would also allow for item analyses to be conducted to check on the structure of the construct as a whole, thus enhancing the logical examinations and construct-validity analyses done as part of conceptual harmonization.

References

Adhikari, K., Patten, S. B., Patel, A. B., Premji, S., Tough, S., Letourneau, N., Giesbrecht, G., & Metcalfe, A. (2021). Data harmonization and data pooling from cohort studies: A practical approach for data management. *International Journal of Population Data Science, 6*(1), 21. https://doi.org/10.23889/ijpds.v6i1.1680

Angrisani, M., & Lee, J. (2011). *Harmonization of cross-national studies of aging to the health and retirement study income measures (WR-861/5)*. RAND Corporation: Santa Monica, Calif. https://doi.org/10.7249/WR861.5

Ascalon, M. E., Meyers, L. S., Davis, B. W., & Smits, N. (2007). Distractor similarity and item-stem structure: Effects on item difficulty. *Applied Measurement in Education, 20*(2), 153–170. https://doi.org/10.1080/08957340701301272

Atkins, D., Best, D., Briss, P. A., Eccles, M., Falck-Ytter, Y., Flottorp, S., Guyatt, G. H., Harbour, R. T., Haugh, M. C., Henry, D., Hill, S., Jaeschke, R., Leng, G., Liberati, A., Magrini, N., Mason, J., Middleton, P., Mrukowicz, J., O'Connell, D., Oxman, A. D., . . . GRADE Working Group. (2004). Grading quality of evidence and strength of recommendations. *BMJ (Clinical Research Ed.), 328*(7454), 1490–1494. https://doi.org/10.1136/bmj.328.7454.1490

Bech, P. (1992). Issues of concern in the standardization and harmonization of drug trials in Europe: Health-related quality of life, ESCT Meeting, Strasbourg, 23–24 May 1991. *Quality of Life Research: An International Journal of Quality of Life: Aspects of Treatment, Care and Rehabilitation, 1*(2), 143–145. https://doi.org/10.1007/BF00439722

Bond, C. F., Jr., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods, 8*(4), 406–418. https://doi.org/10.1037/1082-989X.8.4.406

Brown, S. A., García, A. A., Brown, A., Becker, B. J., Conn, V. S., Ramírez, G., Winter, M. A., Sumlin, L. L., Garcia, T. J., & Cuevas, H. E. (2016). Biobehavioral determinants of glycemic control in type 2 diabetes: A systematic review and meta-analysis. *Patient Education and Counseling, 99*(10), 1558–1567. https://doi.org/10.1016/j.pec.2016.03.020

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

Chen, D., Jutkowitz, E., Iosepovici, S. L., Lin, J. C., & Gross, A. L. (2021). Pre-statistical harmonization of behavrioal [sic] instruments across eight surveys and trials. *BMC Medical Research Methodology, 21*(1), 227. https://doi.org/10.1186/s12874-021-01431-6

Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relations. In L. B. Sechrest & A. G. Scott (Eds.), *Understanding causes and generalizing about them. New Directions for program evaluation* (Vol. 57). Jossey-Bass.

Cooper, H. M. (2017). *Research synthesis and meta-analysis: A step-by-step approach* (5th ed.). Sage.

Crites, S. L., Fabrigar, L. R., & Petty, R. E. (1994). Measuring the affective and cognitive properties of attitudes: Conceptual and methodological issues. *Personality and Social Psychology Bulletin, 20*(6), 619–634. https://doi.org/10.1177/0146167294206001

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. https://doi.org/10.1007/s11205-013-0335-6

DeJonge, T., Veenhoven, R., & Kalmijn, W. (2017). *Diversity in survey questions on the same topic: Techniques for improving comparability*. Springer.

Eagly, A. H., & Carli, L. L. (1981). Sex of researchers and sex-typed communications as determinants of sex differences in influenceability: A meta-analysis of social influence studies. *Psychological Bulletin, 90*(1), 1–20. https://doi.org/10.1037/0033-2909.90.1.1

Edwards, A. L. (1952). The scaling of stimuli by the method of successive intervals. *Journal of Applied Psychology, 36*(2), 118–122. https://doi.org/10.1037/h0058208

Endig, A. W. (1953). The reliability of self-ratings as a function of the amount of verbal anchoring and of the number of categories on the scale. *Journal of Applied Psychology, 37*, 38–41. https://doi.org/10.1037/h0057911

ESOMAR European Society for Opinion and Marketing Research. (2003). The ESOMAR standard demographic classification. In J.H.P. Hoffmeyer-Zlotnik & C. Wolf, (Eds.), *Advances in cross-national comparison*. Boston, MA: Springer. https://doi.org/10.1007/978-1-4419-9186-7_6

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One, 4*(5), e5738. https://doi.org/10.1371/journal.pone.0005738

Fortier, I., Doiron, D., Little, J., Ferretti, V., L'Heureux, F., Stolk, R. P., Knoppers, B. M., Hudson, T. J., & Burton, P. R. (2011). Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. *International Journal of Epidemiology, 40*(5), 1314–1328. https://doi.org/10.1093/ije/dyr106

Fortier, I., Raina, P., van den Heuvel, E. R., Griffith, L. E., Craig, C., Saliba, M., Doiron, D., Stolk, R. P., Knoppers, B. M., Ferretti, V., Granda, P., & Burton, P. (2017). Maelstrom research guidelines for rigorous retrospective data harmonization. *International Journal of Epidemiology, 46*(1), 103–115. https://doi.org/10.1093/ije/dyw075

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher, 5*(10), 3–8. https://doi.org/10.3102/0013189X005010003

Goeltz, R. K. (1991). International accounting harmonization: The impossible (and unnecessary?) dream. *Accounting Horizons, 5*(1), 85.

Gözütok, A. S. (2018). Critical issues in survey meta-analysis. Unpublished doctoral dissertation. Florida State University.

Griffith, L., van den Heuvel, E., Fortier, I., Hofer, S. M., Raina, P., Sohel, N., Payette, H., Wolfson, C., & Belleville, S. (2013). Harmonization of cognitive measures in individual participant data and aggregate data meta-analysis. methods research report. (Prepared by the McMaster University Evidence-based Practice Center under Contract No. 290-2007-10060-I.) AHRQ Publication No.13-EHC040-EF. Rockville, MD: Agency for Healthcare Research and Quality. https://www.ncbi.nlm.nih.gov/books/NBK132553/

Griffith, L. E., van den Heuvel, E., Fortier, I., Sohel, N., Hofer, S. M., Payette, H., Wolfson, C., Belleville, S., Kenny, M., Doiron, D., & Raina, P. (2015). Statistical approaches to harmonize data on cognitive measures in systematic reviews are rarely reported. *Journal of Clinical Epidemiology, 68*(2), 154–162. https://doi.org/10.1016/j.jclinepi.2014.09.003

Guyatt, G. H., Oxman, A. D., Kunz, R., Vist, G. E., Falck-Ytter, Y., Schünemann, H. J., & GRADE Working Group (2008). What is "quality of evidence" and why is it important to clinicians? *BMJ (Clinical Research ed.), 336*(7651), 995–998. https://doi.org/10.1136/bmj.39490.551019.BE

Hofer, S. M., & Piccinin, A. M. (2009). Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological Methods, 14*(2), 150–164. https://doi.org/10.1037/a0015566

Holloway, J., & Collins, D. (1982). Social policy harmonization in the European community. *Journal of Social Policy, 11*, 144–144.

Hull, C. L. (1922). The conversion of test scores into series which shall have any assigned mean and degree of dispersion. *Journal of Applied Psychology, 6*(3), 298–300.

Joint Committee on the Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. (2014). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

Jones, L. V., & Thurstone, L. L. (1955). The psychophysics of semantics: an experimental investigation. *Journal of Applied Psychology, 39*(1), 31–36. https://doi.org/10.1037/h0042184

Joyce, C. R. B., McGee, H. M., & O'Boyle, C. A. (Eds.) (1999). *Individual quality of life*. Routledge.

Kalmijn, W. M. (2010). Quantification of happiness inequality. Unpublished doctoral dissertation. Erasmus University Rotterdam.

Kish, L. (1994). Multipopulation survey designs: Five types with seven shared aspects. International Statistical Review, 62(2), 167–186.

Kish, L. (1999a). Combining surveys: A framework. Bulletin of the International Statistical Institute: Proceedings of the ISI 52nd Session, Finland. https://www.stat.fi/isi99/proceedings/arkisto/varasto/kish0135.pdf

Kish, L. (1999b). Cumulating/combining population surveys. *Survey Methodology, 25*(2), 129–138.

Kish, L. (2002). Combining multipopulation surveys. *Journal of Statistical Planning and Inference, 102*, 109–118.

Kolen, M.J., & Brennan, R.L. (2004). Test equating, scaling, and linking. Springer

Lewis, S. M. (1990). Standardization and harmonization of the blood count: The role of International Committee for Standardization in Haematology (ICSH). *European Journal of Haematology. Supplementum, 3*, 9–13. https://doi.org/10.1111/j.1600-0609.1990.tb01520.x

Morton, S. (1999). Combining surveys from a meta-analysis perspective. Bulletin of the International Statistical Institute: Proceedings of the ISI 52nd Session, Finland. https://www.stat.fi/isi99/proceedings/arkisto/varasto/mort0275.pdf

Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology, 18*, 143. https://doi.org/10.1186/s12874-018-0611-x

National Quality Forum. (2010). Guidance for measure harmonization: A consensus report. Washington, DC: NQF. https://www.qualityforum.org/Publications/2011/05/MeasureHarmonization_full.aspx

Nikula, S., Jylhä, M., Bardage, C., Deeg, D. J., Gindin, J., Minicuci, N., Pluijm, S. M., Rodríguez-Laso, A., & CLESA Working Group (2003). Are IADLs comparable across countries? Sociodemographic associates of harmonized IADL measures. *Aging Clinical and Experimental Research, 15*(6), 451–459. https://doi.org/10.1007/BF03327367

Ostrom, T. M. (1969). The relationship between the affective, behavioral, and cognitive components of attitude. *Journal of Experimental Social Psychology, 5*(1), 12–30. https://doi.org/10.1016/0022-1031(69)90003-1

Perie, M., Moran, R., & Lutkus, A. D. (2005). *NAEP 2004 Trends in academic progress: Three decades of student performance in reading and mathematics*. (NCES 2005–464). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Washington, DC: Government Printing Office.

Pilotte, W. J., & Gable, R. K. (1990). The impact of positive and negative item stems on the validity of a computer anxiety scale. *Educational and Psychological Measurement, 50*(3), 603–610. https://doi.org/10.1177/0013164490503016

Pluijm, S. M., Bardage, C., Nikula, S., Blumstein, T., Jylhä, M., Minicuci, N., Zunzunegui, M. V., Pedersen, N. L., & Deeg, D. J. (2005). A harmonized measure of activities of daily living was a reliable and valid instrument for comparing disability in older people across countries. *Journal of Clinical Epidemiology, 58*(10), 1015–1023. https://doi.org/10.1016/j.jclinepi.2005.01.017

Quatresooz, J., & Vancraeynest, D. (1992). Harmonisation of demographics in Europe 1991: The state of the art; Part 2: Using the ESOMAR Harmonised Demographics: External and internal validation of the results of the EUROBAROMETER Test. *Marketing and Research Today, 20*(1), 41.

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. https://doi.org/10.1007/s10742-008-0032-0.

RetailMeNot Editors. (2021). RetailMeNot Study Finds Reese's and M& M's Are STILL the Most Popular Halloween Candies This Year. RetailMeNot. https://www.retailmenot.com/blog/favorite-halloween-candy-revealed.html

Schenker, N., & Raghunathan, T.E. (2007). Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in Medicine, 26*(8), 1802–1811. https://doi.org/10.1002/sim.2801

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Academic Press.

van de Water, H. P., Perenboom, R., J., & Boshuizen, H. C. (1996). Policy relevance of the health expectancy indicator; an inventory in European Union countries. *Health Policy, 36*(2), 117–129. https://doi.org/10.1016/0168-8510(95)00803-9

van den Heuvel, E. R., Griffith, L. E., Sohel, N., Fortier, I., Muniz-Terrera, G., & Raina, P. (2020). Latent variable models for harmonization of test scores: A case study on memory. *Biometrical Journal, 62*(1), 34–52. https://doi.org/10.1002/bimj.201800146

Veenhoven, R. (2007). Subjective measures of well-being. In M. McGillivray (Ed.) Human well-being: Concept and measurement. Palgrave/McMillan.

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

Veenhoven, R. (2009). How do we assess how happy we are? Tenets, implications and tenability of three theories. In A. K. Dutt & B. Radcliff (Eds.), *Happiness, economics and politics: Towards a multi-disciplinary approach* (pp. 45–69). Edward Elger.

Veenhoven, R. (2015). Concept of happiness. Downloaded from https://worlddatabaseofhappiness-archive.eur.nl/hap_quer/introtext_measures2.pdf

Veenhoven, R. (n.d.) World Database of Happiness, Erasmus University Rotterdam, The Netherlands. http://worlddatabaseofhappiness.eur.nl

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.

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

Wang, S., Min, J., & Lee, J. (2014). Harmonization of cross-national studies of aging to the Health and Retirement study: USER GUIDE, Health behavior, Version A. (WR-861/8) Santa Monica, Calif.: RAND Corporation. https://doi.org/10.7249/WR861.8

Winters, K., & Netscher, S. (2016). Proposed standards for variable harmonization documentation and referencing: A case study using QuickCharmStats 1.1. *PLoS One, 11*(2), e0147795. https://doi.org/10.1371/journal.pone.0147795

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. M. (2013). Using grounded theory as a method for rigorously reviewing literature. *European Journal of Information Systems, 22*(1), 45–55. https://doi.org/10.1057/ejis.2011.51

Ye, D., Ng, Y. K., & Lian, Y. (2015). Culture and happiness. *Social Indicators Research, 123*(2), 519–547. https://doi.org/10.1007/s11205-014-0747-y

Zumbo, B., & Woitschach, P. (2021). A critique of the conventional methods of survey item transformations, with an eye to quantification. In Michalos, A. C. (Ed.), *The Pope of Happiness—A Festschrift for Ruut Veenhoven* (pp. 303–313). Springer. https://doi.org/10.1007/978-3-030-53779-1_30