Methodological Sources of Bias Affecting On Subjective Well-Being (SWB)

Mehrdad Mazaheri1

1- PhD. Associated Prof. University of Sistan and Baluchestan

E-mail: mazaheri@hamoon.usb.ac.ir

Abstract

A rating scale, like all other measurement tools, is useful only if it provides an unbiased, reliable and valid measure. Several types of biasing from different sources, methodology, personality, and culture-related biases, may affect on SWB measurements. A brief taxonomy of some of these biases specifically response format-related biases, that are considered to affect measurements of SWB is presented in this paper.

key word: Well-Being, methodology

Introduction

Self-report rating scales are one of the most common methods to assess overall SWB. According to Andrich (1978) and Guilford (1954), there are four main reasons for the frequent use of the rating scale method of measurement.

1. Rating scales are relatively easy for researchers to construct and use compared with other scale format.
2. They provide the respondent with a limited number of response options, facilitating data registration for both the subject and researcher.
3. The numbering continuum provides respondents with a ruler upon which they can mentally gauge the intensity and/or direction of their reactions to a statement.
4. Accuracy and reliability of one's ability to communicate mental maps increase with experience in using mental rulers. And, conceivably, the repeated use of similar formats increases the accuracy and reliability of measurement process. One internalized ruler can be used to measure directions and intensities across a variety of sentiments.

Sandvik, Diener, and Seidlitz (1993) suggest that standard self-report measures of SWB are adequate for most research, as there is “a unitary core of experience for well-being, which self-reports reflect to great extent. Thus, researchers using standard well-being scales can generally expect they are obtaining meaningful, interpretable information from these scales under ordinary conditions” (p.337). This is partly attributable to the moderate stability of SWB across situations and over the life span (Diener & Lucas, 2000). For example, SWB has been found to correlate 0.85 over a four-year period (Diener, Suh, & Oishi, 1997).

A rating scale, like all other measurement tools, is useful only if it provides an unbiased, reliable and valid measure. Test and item bias are defined as invalidity or systematic error in a measurement process (Berk, 1982). A particular bias usually affects all measurements in the same direction (Osterlind, 1983), but may concern only particular groups of respondents (Camilli and Shepard, 1994). Several types of biasing from different sources, methodology, personality, and culture-related biases, may affect on SWB measurements. A brief taxonomy of some of these biases specifically response format-related biases, that are considered to affect measurements of SWB is presented below.
Personality and culture-related biases

According to their personality, living circumstances or culture, people may vary in response to questions concerning overall life (dis)satisfactions. Schwarz and Strack (1999) have identified several sources of bias in (conventional) self-reported global assessment of SWB. These include assimilation and contrast effects, when current feelings are coloured or discoloured by the past. Similarly, the mood of the day or events of the moment may distort responses. In addition, responses about well-being may be biased by social acceptability (Schwarz and Strack, 1999).

Personality traits are likely to be co-determinants of expressed SWB. This is obvious for optimism/pessimism, but also applies to extraversion/introversion (Brickman, Coates and Bulman, 1980; Furnham and Cheng, 1997; Schimmack et al., 2002; Diener et al., 2003). Similarly, neuroticism is negatively correlated with SWB (Costa and McCrae, 1980). Insofar as they are 'innate' in people with a common ancestry and cultural background, they are of course legitimate fixed characteristics (traits) of nations. Examples cited by Argyle are that in Ifaluk, a small Pacific island, looking happy is not socially acceptable because it suggests that an individual is showing off, while in China, people focus more on negative events and are less optimistic than Americans (Argyle, 2002).

Extraversion/introversion, which is a strong correlate of SWB (Furnham and Cheng, 1997; Schimmack et al., 2002) may be a genuine determinant of SWB, but also a biasing factor e.g. if the extravert person enjoys being questioned, and the introvert loathes it. American SWB averages, for instance, which are among the highest in the world (Veenhoven, 2006), may pose an interpretation problem if they are influenced by the rather American trait of extraversion. In general, personality traits, as innate features, would resist social change, and therefore tend to obscure the impact of objective factors of well-being on SWB. Personality traits may therefore hamper QOL research in its quest for determinants of happiness.

Proximate or peer-relativity bias occurs when respondents are temperamentally or culturally inclined to answering the SWB question by referring to the criteria of the system or group they belong to rather than to their personal criteria. In health-related QOL research, diseased persons sometimes have similar rating distributions as their healthy socio-demographic counterparts. For instance, it was in some studies difficult to distinguish women with breast cancer from healthy controls, and sick elderly were apparently no less happy than matched healthy people (De Haes, 1992). Moreover, one's response about well-being may be altered by conscious or unconscious comparison with others, often peers. The frame of reference is in most surveys neither made explicit by the respondent nor inquired into by the interviewer (Kahn & Juster, 2002).

Cultural biases have been shown not to suffice to explain the major differences in SWB between countries (Veenhoven, 1997), but still pose a major problem to the international comparison of QOL data (Schimmack et al., 2002). Asians, for instance, usually have intermediate mean ratings, with a much narrower distribution than Westerners (Diener et al., 1995; Lau & Cummins, 2003). This phenomenon has been attributed to the cultural valuation of moderation (Diener et al., 1995). Another factor possibly contributing to the distinctive distribution of SWB among Asians is less individualism and the persistent closeness of family ties. Such ways of life are characterized by increased control, and thus tend to limit the impact of perturbations on SWB and the affect balance, whether by buffering negative events or by blunting positive affects.

Methodological sources of bias affecting on SWB

A variety of methodological biases may affect on responding to question of SWB. For example, the phenomena of regression to the mean or ceiling or floor effects may explain differences in scale responsiveness (Bruin et al., 1997). The 'scaling-effects bias' includes the phenomenon that respondents who assume that the mid-point of the scale reflects an 'average', will compare themselves to that average (Schwarz and Hippler, 1987). Interviewer attitudes and character are other obvious sources of bias.

A summary of some of the common methodological biases that are considered to affect measurements of QOL is presented below;

Sample selection bias takes place if participants are systematically different from non-participants (i.e., individuals selected for research can, and frequently do, refuse participate). This occurs when one or more group(s) of people is over- or under-represented in the study population because of the way in which participants were selected – for example, selecting participants over the telephone means that poorer people will be underrepresented in the study (Donald, 2001). Drop-out bias is when so many participants leave the study (usually more than 15%) that the results from those remaining in the study are no longer representative of the original group (Donald, 2001). Subject or participant bias occurs when participants
behave in some way so that their responses are systematically higher or lower than they would have been in normal circumstances.

Experimenter, Interviewer and observer bias refer to experimenters’ expectations influence results. This occurs when the measurement of some aspect of quality of life systematically skews the responses in a particular direction – for example, if weight gain is used to measure improvement in health status following cancer treatment, a set of scales that started at 1 kg instead of 0 kg could yield misleading results (Donald, 2001). Interviewer attitudes and character are other obvious sources of bias.

Measurement bias occurs when the measurement of some aspect of quality of life systematically skews the responses in a particular direction – for example, if weight gain is used to measure improvement in health status following cancer treatment, a set of scales that started at 1 kg instead of 0 kg could yield misleading results (Donald, 2001).

Agreement bias, also referred to as acquiescence response style (Martin, 1964), is a tendency to agree with statements, irrespective of the content of the item. Agreement bias occurs when individuals differ in their tendency to agree with item statements. Such bias adds individual variation over and above variation in the construct being measured. Moreover, agreement bias can be caused by several factors (Baumgartner & Steenkamp, 2001): stimulus-seeking extroverts (Couch & Keniston, 1960); lower status or cognitive ability of respondents (Knowles & Nathan, 1997; Schuman & Presser, 1981); ambiguous, vague, or neutral items (Paulhus, 1991); or distraction, time pressure, or other such factors (McGee, 1967).

Disacquiescence response style, also referred to as disagreement bias or nay-saying, is the opposite of acquiescence response style and could be caused by stimulus-avoiding introverts (Couch & Keniston, 1960). Net acquiescence response style (Baumgartner & Steenkamp, 2001) is the sum of these two response styles and is also referred to as direction bias. In studies of response style effects (Baumgartner & Steenkamp, 2001; Martin, 1964), such response styles are assessed by the degree of agreement (or disagreement) with heterogeneous items from multiple scales without much in common, or from the extent of agreement with positively and negatively worded items from one scale before reverse scoring.

Non-contingent responding is the tendency to be careless, random, or non-purposeful in responding (Baumgartner & Steenkamp, 2001; Watkins & Cheung, 1995) and may occur because of lack of motivation. It has been measured by summing the absolute differences between pairs of items that are highly correlated and have similar means across respondents, and that are worded in the same direction (Bachman & O’Malley, 1984; Baumgartner & Steenkamp, 2001; Watkins & Cheung, 1995).

Social desirability, a tendency to present oneself in a favorable light, can similarly lead to additive or correlational systematic error. Standard deviation error, a tendency to use a wide or narrow range of responses, can increase or reduce spread. If individuals vary consistently in standard deviation, this pattern leads to within-measure correlational systematic error (i.e., consistent differences across individuals over and above the construct being measured). If a constant, pervasive effect leads to greater or lesser spread, this pattern is also an example of within-measure correlational systematic error.

Primacy effects occur when an item placed at the beginning of a list has a higher likelihood of being selected over other items in the list and recency effects occur when items at the end of a list are more likely to be selected. Based on their review of the literature, Krosnick and Alwin (1987) suggest that when items are presented visually, primacy effects should be expected for several reasons. First, items presented early are given importance for later judgments because they establish a standard of comparison. Second, items seen first are processed at a deeper level because they are not in competition with other alternatives. Finally, many individuals will, in many cases, select the first acceptable answer or satisfied to minimize cognitive costs, rather than diligently search a list for an optimal answer.

Location bias occurs when individuals differ in the manner in which they use response scale categories (e.g., a tendency to scale upward or use extremes). Leniency is the tendency of a respondent to rate too high or too low. Severity (or stringency) is the opposite of leniency.

End-aversion bias, which also called the central tendency bias, refers to the reluctance of some people to use the extreme categories of a scale. The effect of this bias is to reduce the range of possible response. The ‘scaling-effects bias’ includes the phenomenon that respondents who assume that the midpoint of the scale reflects an ‘average’, will compare themselves to that average (Schwarz and Hippler, 1987).

Extreme response style refers to choosing extreme responses irrespective of content (Greenleaf, 1992b). It could be caused by several factors (Baumgartner & Steenkamp, 2001): an intolerance for ambiguity or dogmatism (Hamilton, 1968); anxiety (Hamilton, 1968); respondents lacking appropriate cognitive schemas (Shulman, 1973); or stimuli that are meaningful, important, or involving to respondents.
Response range is the tendency to use response categories in a narrow or wide range and may be caused by factors similar to those that cause extreme response style (Baumgartner & Steenkamp, 2001). It has been measured by the standard deviation in an individual's responses across items (Greenleaf, 1992; Hui & Triandis, 1985; Wyer, 1969).

The halo effect is the tendency to rate someone high or low in all categories because he or she is high or low in one or two areas. Halo effect (is a phenomenon first recognized 80 years ago) also causes within-measure correlational systematic error and is a tendency to provide similar responses across items that are thought to be related. Such error, if restricted to a small proportion of individuals, is similar to idiosyncratic random error.

Midpoint responding, a tendency to use the middle scale point irrespective of content (Baumgartner & Steenkamp, 2001), may be caused by evasiveness, indecision, or indifference (Messick, 1968; Schuman & Presser, 1981). This response style has been measured by the proportion of use of midpoints (Chen, Lee, & Stevenson, 1995). Midpoint responding leads to use of middle alternatives and is likely to cause within measure correlational systematic error. In other words, although not affecting means, this type of error can lead to consistent differences over and above the construct in question. If any such error affects a small proportion of individuals, it may be identified as idiosyncratic random error. However, if a factor such as the use of extreme wording in the end anchors has a more pervasive effect, then the result is within-measure correlational systematic error.

Standards of Evaluation; Using categories such as fair, good, excellent as the scales verbal points may affect the way applicants responding to the questions because the meanings of these words can be differed from person to person.

In addition to the distortions already mentioned, scales which are scored on a continuum, such as visual analogue and Likert scales, may prone to other types of biases such as the type of response format and scale’s orientation (Scott & Huskisson, 1976, 1979; Frisbie & Brandenburg, 1979; Gift, 1989; Friedman & Friedman, 1994; Schwartz et al. 1991; Paul-Dauphin et al., 1999; Stephenson & Herman, 2000; Sangster et al. 2001).

Response format related-factors

The debate among researchers as to the “ideal” rating format has an extensive history. A desired effect of the rating scale method is to provide subjects with a format that allows them to make equal interval judgments thus meeting statistical assumptions of an interval scale of measurement. However, while the rating scale provides a powerful tool for investigating a wide variety of phenomenon, investigations of rating scale function reveal performance anomalies across scale formats.

Ratings scales differ in the number of categories as well as number and placement of labels to aid in selection of a category. “Label” is verbal, descriptive statements placed at various locations along vectors of possible response options. Frequently, these options are numbers of increasing and/or decreasing magnitude. The respondent’s task is to select the numerical response option associated with the appropriate label that he/she perceive to be the best representation of his/her attitude or belief on a latent trait. There are several characteristics of response formats that are of relevance to the quality of survey data, ranging from the labeling of response categories and the issue of administering scales with or without midpoints, to the question of whether response categories are ordered from positive to negative or the other way around.

Number of rating point

In light of the extensive use of rating scales, it would be also useful to have a clear understanding of how to optimize reliability and validity through use of the scale. Number of rating points used on the “ruler” can vary from 2 to 100 or more. There is a general consensus that the optimal number is from 5 to 7 points. But the specific number of points eludes researchers. Whether to use an even or odd number of categories is another source of debate among practitioners (e.g., Gable & Wolf, 1993). Cox (1980; p.408) provides the following definition of optimal number of rating points: “At a general level, a scale with the optimal number of response alternatives is refined enough to be capable of transmitting most of the information available from respondents without being so refined that it simply encourages response error. At that optimal number, the ratio of meaningful or systematic variation on total variation is maximized. At
an operational level, the optimal number depends on the purpose of the scale and thus the nature of its systematic variation."

**Scale labels**

Compared to the research on the number of scale points necessary to maximize reliability, effects of scale labels on rating responses have not been examined as extensively. Frisbie and Brandenburg (1979) varied the type of evaluation scale labels (verbal and numerical) and the degree of labeling (all the ordered response points labeled versus only the end points labeled). They collected responses from a large sample (approximately 900 students in each of the four conditions) on an eight-item measure of college freshmen expectations. Each form contained a mix of 4- and 5-ordered response categories. No differences were noted in the scale means for the verbal and numerically anchored forms, but the forms with end-only anchoring resulted in higher mean ratings (more positive) compared to when all the ordered response categories were anchored. They concluded that: Bipolar adjective scales may tend to yield higher (more positive) ratings than content-parallel item stems with fully defined scales. In the absence of a suitable criterion it is impossible to determine which of the two formats is more valid. (p. 47)

**Intermediate and endpoint-labeled response format**

Lam and Kiocars (1982) proposed that the difference in ratings between scales having all intermediate points labeled and those having only endpoints labeled, as observed by Frisbie and Brandenburg (1979), is a function of the locations of intermediate labels on the continuum being measured. In the case of the scale where the intermediate points labeled, subjects will choose a label at the point on the continuum closest to their actual position on the same continuum because subjects are primarily content bound in their choice of response alternatives. However, when the psychological continuum has only the endpoints labeled, subjects will attach meaning to the alternatives by partitioning the distance between the endpoints into equal units. In other words, scales with only endpoints labeled produced results similar to scales with equally spaced response labels. Subjects tend to mentally divide the response categories into equal units, thereby creating an interval scale. The equal interval properties of a rating scale were dependent on an appropriate choice of labels to anchor the points on the rating scale.

Dixon, Bobo and Stevick (1984) reviewed several studies where different techniques were used to label the response categories. Techniques consisted of various combinations, such as labeling all categories versus labeling only categories at the end of the response continuum, verbal versus numerical labels and vertical versus horizontal labels. For the data gathered by Dixon et al., no significant differences were found between the end-defined and all-category defined versions of items from Rotter's (1966) locus of control scale. Two Likert-type formats (29 items per format), one with all choice points defined and the other with only endpoints defined were administered to 121 college students. The label defined format used a 6-point scale (strongly agree, agree very much, tend to agree, tend to disagree, disagree very much, strongly disagree). The end-defined format had strongly agree on the extreme left under each question with six blanks on the continuum with strongly disagree on the extreme right. Each subject completed half of the items in the all-category-defined condition and the other half in the end-defined condition. Results did not show a significant difference between formats, nor did subjects indicate a format preference. Significantly more items had a large standard deviation in the end-defined format than in the defined, suggesting a greater variability in the end-defined format.

French-Lazovik and Gibson (1984) demonstrated that the distribution parameters (means, skewness) of rating scale data were influenced by the labels used. They hypothesized that the negative skew found in most distributions of performance ratings is dependent upon the choice of verbal labels used to anchor rating scale points. Using evaluative rating labels, they were able to systematically change the ratings in the predicted (or hypothesized) direction. The degree of negative skew in distributions of behavioral rating measures was altered by the verbal labels used as anchors. Both means and skewness coefficients were affected by the evaluation labels used in the study. Results presented also showed that a rating scale anchored by a set of more positive evaluative labels (those having higher descriptor indices), results in a shift of the mean numerical values toward the less positive end of the scale. On the other hand, a rating scale anchored by a set of less positive evaluative labels results in a shift of the mean numerical values toward the more positive end of the scale.

**Scale's orientation**
Some researchers prefer using a horizontal format of rating scale, while others opt for a vertical presentation of the response format. Some investigators have argued that the orientation of rating scales can affect the respondent’s responses to rating scales (Scott & Huskisson, 1979; Grau, 1998). Sriwatanakul and co-workers (1983) compared five different types of rating scales to assess pain and concluded that horizontal scales with gradations were both more reliable and significantly preferred by participants whereas vertical linear scales were found to have higher coefficients of variation and distributions that divert more from normal. In spite of the high correlation between vertical and horizontal formats of rating scale (Stephenson & Herman, 2000; Gift, 1989), few studies comparing horizontal and vertical orientations of rating scales have produced contradicting results. For example some researchers demonstrated that subjects usually report a higher mean score on vertical formats of VAS than on horizontal formats (Stephenson & Herman, 2000; Gift, 1989). But, Breivik and Skoglund (1998) did not find any significant differences between horizontal and vertical presentations of rating scale and concluded that vertically and horizontally oriented rating scale are equally sensitive in assessing the intensity of present pain in oral surgery.

**Bipolar and Unipolar scales**

Rating scales can be presented as a bipolar or unipolar format. There are two ways in which we may signal to respondents whether we wish them to treat a response scale as unipolar or bipolar. The usual way is by using verbal anchors which are either unipolar (eg [no more power, much more power], [not having any success, having great success]) or bipolar (eg [much more power, much less power], [much success, much failure]). The second way is to use numeric labels which either imply a unidimensional construct (eg [0 to 10], [0 to 5], [0 to 6], [-5 to 0]) or bipolar construct (e.g. [+5 to -5], [+3 to -3], {+2 to -2}) (O’Muircheartaigh, Gaskell, and Wright, 1993).

While the numeric values are often included only for coding and response convenience, Schwartz and co-workers (1991) have demonstrated that they carry more, sometimes unintended, meanings. For a particular question, "How successful have you been in life, so far?", they showed that a scale with numeric values ranging from 0 to 10 was not the same as a scale whose values ranged from -5 to +5. The verbal anchors were "not at all successful" (0 or -5) and "extremely successful" (10 or +5). They argued that when a 0 to 10 scale is used respondents infer that 0 stands for the absence of any amount; the scale becomes unipolar. In contrast, respondents infer that the scale is bipolar when the numeric values range from -5 to +5. For example, when asking people how successful they had been in their life, if a 0 to 10 scale is offered, they will assume that the low anchor (0) corresponds to not having any success. This contrasts with the interpretation of the lowest point on the -5 to +5 scale as being unsuccessful (being a failure). Sangster and co-workers (2001) in a study on the evaluation of the universities by their students and faculty found the greater proportion of responses above the mid-value (zero) on the scale with bipolar format (-3 to +3) than for the scale with unipolar format (1 to 7), and a more frequent use of the mid-point value in the unipolar than for the scale with bipolar format, and the evaluations of universities by students were more positive when using the bipolar scale than with the unipolar scale. Mazaheri & Theuns (2006) used three different formats of a VAS scale for measuring SWB and found that, in spite of non-significant differences between score means of SWB when using equivalent horizontal and vertical formats, the higher percentages of respondents was shown to range in the upper part of the distribution (mid-point to top) of the bipolar formats (with set of anchor-points -5 to +5) in comparison to the distribution for the unipolar format (with set of anchor-points 0 to 10).

**Conclusions**

A rating scale, like all other measurement tools, is useful only if it provides an unbiased, reliable and valid measure. Some investigators suggested that the type of response format and the scales' orientation may affect the respondent’s response (Scott & Huskisson, 1976, 1979; Frisbie & Brandenburg, 1979; Gift, 1989; Friedman & Friedman, 1994; Schwartz et al. 1991; Paul-Dauphin et al., 1999; Stephenson & Herman, 2000; Sangster et al. 2001). The debate among researchers as to the "ideal" rating format has an extensive history. A desired effect of the rating scale method is to provide subjects with a format that allows them to make equal interval judgments thus meeting statistical assumptions of an interval scale of measurement. However, while the rating scale provides a powerful tool for investigating a wide variety of phenomenon, investigations of rating scale function reveal performance anomalies across scale formats. There are several characteristics of response formats that are of relevance to the quality of survey data,
ranging from the labeling of response categories and the issue of administering scales with or without midpoints, to the question of whether response categories are ordered from positive to negative or the other way around. This paper listed some of the most common sources of bias which influences the respondent’s response on self-report rating scales for measuring overall life satisfaction.

References

Donald A, 2001 ‘What is quality of life?’, Hayward Medical Communications, Vol 1, Number 9, http://www.evidence-based-medicine.co.uk/
French-Lazovik U & Gibson CL,1984. Effects of verbally labeled anchor points on the distributional
change the meaning of scale labels. Public Opinion Quarterly, 55, 570-582.