J Happiness Stud (2014) 15:1407–1423
DOI 10.1007/s10902-013-9484-1

RESEARCH PAPER

More Happy or Less Unhappy? Comparison of the Balanced and Unbalanced Designs for the Response Scale of General Happiness

Pei-shan Liao

Published online: 12 December 2013

Abstract The single-item measure of general happiness has been widely used in questionnaires due to the advantages of easy implementation in surveys for comparison across time and culture. The balanced response scale that includes equal positive and negative response categories based on Likert-type response format has been commonly applied. However, the possibility of using an unbalanced response scale for happiness, for instance, more choices on the happy side, has not been fully examined. This study aims to explore the optimal number of response categories and the corresponding labels for general happiness by using telephone survey data in Taiwan. Six types of response scales with different combinations of response number and response labels were examined to distinguish both the intensity and direction of responses. A completely randomized experimental design using computer-assisted telephone interviewing system was conducted to collect data from representative samples aged 18 years or older. Individual characteristics among the six groups indicated that all of the sub-samples were similar in terms of gender, age, education, marital status, working status, and monthly income. Results of the graded response model suggested that a scale with at least three response categories on the positive side and no more than two on the negative side will be suitable for the single measure of general happiness. Findings of ordered logit regression on happiness were also in favor of an unbalanced response design. A discussion of the result is provided.

Keywords Happiness · Response scale · Unbalanced design · Graded response model

General happiness has been commonly used as a self-reported affective measure for quality-of-life or subjective well-being. A typical single measure of general happiness can be stated as, for example, "Taking all things together, how happy would you say you are

P. Liao (✉)
Academia Sinica, Taipei, Taiwan, ROC
e-mail: psliao@gate.sinica.edu.tw

recently?" It has the advantages of easy implementation in surveys for comparison across time and culture, as well as minimizing the impact of data nonresponses when using multi-item scales (Diener 1984). In the empirical studies on general happiness, the scale points may range from two to one hundred, with one end labeled as "very happy" and the other as "very unhappy" (see Cummins and Gullone 2000; Kalmijn et al. 2011; Lim 2008), with both ordinal and interval scales employed.

Despite of its variation in rating scale, when the single measure of happiness is used in surveys, 70 % or more respondents report that they are happy (Cummins 2003; Veenhoven 1993). This is a common phenomenon disregarding regions, cultures, or survey topics. It may seem plausible that people are generally happy all around the world. From the view of survey data analysis, however, this is a feature of the negatively skewed distribution. It is even more obvious when both of the numeric points and response labels are included in a rating scale, in which case the scale points are usually limited to seven or fewer.

Similar to other aspects of subjective quality of life (SQoL), the commonly-used Likert-type rating scales usually formed skewed-to-the-happy-side data rather than near-normal distributions (see Cummins and Gullone 2000). Such a skewed distribution is mostly resulted from a balanced response design that includes an equal number of positive and negative responses, with or without a mid-point response (e.g., Lim 2008; Kalmijn et al. 2011). On the one hand, the negatively skewed data are often transformed into a near-normal distribution since common statistical techniques, such as *T* test, ANOVA, Chi square test, and regression models, just to name a few, ought to be applied under the assumption of normal distribution of data. A near-normal distribution of a modified measuring scale on general happiness or SQoL may diminish the burden of data management.

On the other hand, some have considered the possibility of using an unbalanced response scale (see Kalmijn et al. 2011; Liao et al. 2005; Tourangeau et al. 1991), although the reasons of using such an unbalanced response design are not clear. With respect to the methodological issues of response format, many have thoughtfully considered various statistical approaches to the single measure of general happiness to account for its comparability across data sets (Cummins 2003; Ferrer-i-Carbonell and Frijters 2004; Fordyce 1988; Kalmijn et al. 2011; Lim 2008). However, few have examined the issue of unbalanced response design from the perspective of survey methodology. It is of interest to explore the possibility of alternative measurement scales.

This study aims to compare balanced and unbalanced designs of response scales that include the combinations of number of response categories and the corresponding labels by using the single item of general happiness as an example. Survey data collected by a completely randomized experiment in a computer-assisted telephone interviewing (CATI) system are analyzed. It is essential to consider whether a midpoint response should be included for odd-number categories as a balanced response design. When studying certain issues that tend to have negative skew of data, such as general happiness, it may be more appropriate to use an unbalanced response design to obtain a near-normal distribution. The corresponding labels that distinguish both of the direction and intensity are examined as well.

1 Perspectives from Survey Methodology

For measures of subjective phenomena and attitudes, the Likert-type rating scale is the most commonly used one in surveys, compared to other forms of response design

(Cummins and Gullone 2000; Schaeffer and Presser 2003). The measure of happiness unexceptionally shares the common ground and faces the same issues of the design for survey responses. Previous studies on response design of surveys address the issues of the provision of a midpoint category, its effect on response patterns and distributions, and the true meaning of such a category (Bishop 1987; Moors 2007; Raaijmakers et al. 2000; Schaeffer and Presser 2003; Schuman and Presser 1981). The inclusion of a midpoint category indeed involves a decision regarding having an odd or even number of response categories. Although midpoint categories are often classified as non-substantial responses due to their uncertain nature (Francis and Busch 1975), empirical studies have demonstrated their middle position between two anchor points when rating is employed in the scale (Kroh 2007; Wildt and Mazis 1978).

In addition to the inclusion of a midpoint response, the number of response categories and their corresponding labels, which is also seen as verbal responses (Cummins and Gullone 2000), have also been important issues for the design of response scales. Previous researches suggest using a larger number of response categories in order to improve reliability of measurement (Alwin 1992, 1997; Alwin and Krosnick 1991; Chang 1994; Masters 1974; Saris and Gallhofer 2007; Weng 2004), since binary responses tended to decrease estimation accuracy. Variations in measurement may, therefore, be limited (Comrey 1988). The estimation of measurement reliability included coefficient alpha reliabilities, test–retest stability, and multi-trait multi-method (MTMM) approach. A number of no less than three is suggested (Preston and Colman 2000) and, more specifically, five to nine categories are recommended when conducting surveys on the general public (Alwin 1997). Cummins (2003) further concluded that the happiness measurement is best with at least an 11-point scale, but he did not take the response labels into account. Researchers need to be aware, however, that a large number of categories can be difficult to administer in telephone interviews, where fewer categories are often used instead (Dawes 2008; Schaeffer and Charng 1991; Schaeffer and Presser 2003). The number of response categories may vary depending on the target sample due to their comprehension and cognitive abilities. Since responding to survey questions is a multi-task process, the elderly and those with lower education level may find it difficult to choose from a large number of response categories (Fox et al. 2007; Holbrook et al. 2006; Weng 2004).

With regard to response labels, previous studies have focused on the linguistic intensity and the results suggest the changes of response patterns. Alwin and Krosnick (1991) compare the 7-point scales with full labels with those without full labels and their results support the practice of labeling response scales extensively for higher reliability. Lam and Klockars (1982) manipulate response labels at the end points and indicate that the intensity of response categories other than the end points and their positive/negative label wording directly affect the distribution of scale scores. The effect of label intensity is decided by sensitivity of the content and intensity of the questions.

Few studies on the response formats of general happiness and life satisfaction have examined the ratings from a range of measurement scales with end points labeled (Mazaheri and Theuns 2009a, b; Wedell and Parducci 1988). Little differences are found for those balanced response scales, whether a midpoint category is indicated, A comparison among different measures of quality-of-life across nations suggest using bi-directional Likert-type rating scales with at least 11 points to account for the values that are normally found within a narrow positive range (Cummins 2003). These findings pave the way for the use of an unbalanced response design. When measuring quality of life or constructs with a highly skewed distribution, an unbalanced design for response scale may be more useful. In practice, a scale with more positive than negative choices has been applied for the measure of general happiness (see Kalmijn et al. 2011; Liao et al. 2005; Tourangeau et al. 1991). Unfortunately, those studies with a 3-point scale, where two are on the happy side and one is on the unhappy, could not accurately reflect people's true levels of quality-of-life (Cummins 2003; Cummins and Gullone 2000). Instead, a 4- to 7-point scale with both number and labels of categories may work well (see Kalmijn et al. 2011). It is of interest to see whether an unbalanced response scale with more than two positive points can function well for the measure of general happiness, when compared to a balanced design. The numbers on response scale and corresponding labels are examined as well to take the effect of label intensity into account.

2 Research Design

This research employed a completely randomized experimental design to collect survey data on different combinations of response numbers and labels to compare balanced and unbalanced response designs for general happiness. Computer-assisted telephone interview (CATI) system was used to collect data during March and April, 2011. Three-stage stratified systematic sampling was also used to select respondents aged 18 years and older in Taiwan. The first stage was to divide listed residential numbers into 23 district clusters at the city/county level. Next, a systematic sampling method was used to select numbers based on the principle of population proportional to size (PPS) in each of the 23 clusters, with the last four digits being replaced by random numbers. After the telephone numbers were dialed, a probability-based within-household sampling approach developed by Hung (2001) was employed to select one qualified individual within each household.

With the consideration of midpoint category, response labels, and the comparison between balanced and unbalanced response scales, six different designs were developed in this research. The questionnaire included items on socio-demographic variables and subjective quality-of-life (SQoL), such as general happiness and satisfaction with different aspects of life. The response scales varied for SQoL items across the six designs, while the item wordings and their position in the questionnaire were all the same. The respondents were randomly assigned to one of the six groups of SQoL items by the use of CATI. A total of 3,682 surveys were collected, with a response rate of 53.78 %.

Five and seven points of response categories were included for both of the balanced and unbalanced design. The five-point scales included two different types of response labels. The midpoint category in the balanced scales was not read out as a choice by using the unfolding technique, in which interviewers first asked about direction and then about the degree of attitudes (Schaeffer and Presser 2003). Only when the respondents insisted on such an answer would the interviewer recode it. The six designs for the response scales of general happiness are illustrated in Table 1.

Socio-demographic characteristics include the respondent's age, gender, education, marital status, employment status, and monthly income. For the purpose of testing the representativeness of the samples, age was grouped mainly based on 10-year intervals, except for the respondents who were less than 20 years old. The categories of education level, marital status, working status, and monthly income are listed in Table 2. Educational level is recoded in years in the multivariate analysis reported later. Along with these demographic variables, self-rated health and perceived fairness are included in ordinal logit regression models to examine criterion validity of the different response design of happiness. The contribution of self-rated health has been found closely associated with general happiness (Argyle 1997; Michalos et al. 2000). Perceived fairness, with roots in

Table 1 Distribution of six designs for balanced and unbalanced response scale

Group labels	A	B	C	D	E	F
1	Pretty happy (21.8 %)	Pretty happy (13.0 %)	Very happy (13.1 %)	Very happy (11.7 %)	Very happy (7.6 %)	Very happy (6.1 %)
2	Fairly happy (52.8 %)	Happy (38.4 %)	Happy (59.3 %)	Happy (36.2 %)	Pretty happy (16.3 %)	Pretty happy (27.3 %)
3	Neither happy nor unhappy (.7 %)	Fairly happy (27.1 %)	Neither happy nor unhappy (1.7 %)	Fairly happy (31.1 %)	Happy (28.8 %)	Fairly happy (45.0 %)
4	Somewhat unhappy (17.4 %)	Somewhat unhappy (15.1 %)	Unhappy (18.0 %)	Unhappy (13.7 %)	Fairly happy (28.8 %)	Neither happy nor unhappy (.3 %)
5	Pretty unhappy (7.3 %)	Pretty unhappy (6.4 %)	Very unhappy (7.9 %)	Very unhappy (7.3 %)	Somewhat unhappy (11.9 %)	Somewhat unhappy (14.3 %)
6					Pretty unhappy (3.7 %)	Pretty unhappy (3.9 %)
7					Very unhappy (2.9 %)	Very unhappy (3.1 %)
Skewness	.891	.453	.896	.437	.332	.994
Total	593	576	605	563	590	593

the relative deprivation theory, was also found to be an important factor affecting happiness (Liao et al. 2005). Respondents were asked whether one felt that current living standards fairly reflected his or her efforts. Scale points of self-rated health and perceived fairness both correspond to the measures of general happiness in one of the six designs. In addition, age and age square were both examined in the regression models for its commonly found U-shape association with general happiness (Frijters and Beatton 2012; Tsou and Liu 2001).

Socio-demographic variables were first compared across groups to examine the dis/similarity of the six sub-samples. The distributions and the skewness of the response scales were then calculated to see the response patterns of different designs. Furthermore, Samejima's (1969) graded response model (GRM) derived from the item response theory (IRT) was employed to examine the characteristics of item responses. Ordinal logit regression models were conducted to see the effects of self-rated health, perceived fairness, and demographic variables on general happiness in the difference response designs. Since the analyzed data were not nationally representative, as usually the case in telephone surveys, sample weights were employed by adjusting population distribution of gender, age, and educational level in the survey year.

The graded response model is one type of polytomous IRT model, a logistic model specifically designed for ordinal manifest variables. The GRM is used when item responses can be characterized as ordered categorical responses such as Likert scales (Embretson and Reise 2000; Samejima 1997). It allows the items to not to have the same response format in order to fit the GRM within a measure. Each item (i) is described by one item slope parameter (a_i) and $j = 1 \ldots m_i$ between category "threshold" parameters (b_{ij}). The slope or

Table 2 Sample description of six response designs

Variable	Group											
	A		B		C		D		E		F	
	f	%	*f*	%	*f*	%	*f*	%	*f*	%	*f*	%
Gender												
Male	282	44.5	278	46.4	301	48.9	260	44.3	268	43.7	296	46.7
Female	352	55.5	321	53.6	314	51.1	327	55.7	345	56.3	338	53.3
Pearson Chi square	$\chi^2(5) = 4.789, p = .442$											
Age												
18–19 years	4	.6	8	1.3	7	1.1	3	.5	5	.8	6	.9
20–29 years	64	10.1	56	9.3	62	10.1	66	11.2	70	11.4	59	9.3
30–39 years	100	15.8	113	18.9	120	19.5	112	19.1	124	20.2	120	18.9
40–49 years	172	27.1	157	26.2	139	22.6	141	24.0	150	24.5	158	24.9
50–59 years	149	23.5	126	21.0	154	25.0	152	25.9	139	22.7	143	22.6
60 years or older	145	22.9	138	23.0	133	21.6	112	19.1	125	20.4	148	23.3
Refusal	0	.0	1	.2	0	.0	1	.2	0	.0	0	.0
Pearson Chi square	$\chi^2(30) = 25.076, p = .721$											
Educational level												
Elementary or less	92	14.5	86	14.4	98	15.9	81	13.8	91	14.8	96	15.1
Junior high school	79	12.5	74	12.4	85	13.8	59	10.1	74	12.1	71	11.2
Senior high school	207	32.6	194	32.4	189	30.7	189	32.2	171	27.9	189	29.8
Associate	98	15.5	85	14.2	94	15.3	91	15.5	93	15.2	98	15.5
College/ University	119	18.8	116	19.4	115	18.7	139	23.7	157	25.6	149	23.5
Graduates or higher	36	5.7	43	7.2	31	5.0	28	4.8	26	4.2	30	4.7
Others	2	.3	0	0	2	.3	0	0	0	0	1	.2
Refusal	1	.2	1	.2	1	.2	0	0	1	.2	0	0
Pearson Chi square	$\chi^2(35) = 35.865, p = .428$											
Marital status												
Single	119	18.8	100	16.7	112	18.2	112	19.1	118	19.2	114	18.0
Married	419	66.1	412	68.8	406	66.0	421	71.7	429	70.0	447	70.5
Divorced	34	5.4	30	5.0	36	5.9	13	2.2	23	3.8	15	2.4
Separate	6	.9	2	.3	5	.8	4	.7	3	.5	5	.8
Widowed/ widowered	45	7.1	44	7.3	46	7.5	33	5.6	32	5.2	46	7.3
Living with a partner	10	1.6	9	1.5	7	1.1	3	.5	7	1.1	6	.9
Refusal	1	.2	2	.3	3	.5	1	.2	1	.2	1	.2
Pearson Chi square	$\chi^2(30) = 35.233, p = .234$											
Total	634	100.0	599	100.0	615	100.0	587	100.0	613	100.0	634	100.0
Working status												
Full time	347	54.7	351	58.6	352	57.2	319	54.3	361	58.9	358	56.5
Part time	54	8.5	39	6.5	50	8.1	55	9.4	42	6.9	45	7.1
Student	14	2.2	9	1.5	13	2.1	11	1.9	15	2.4	11	1.7
Homemaker	83	13.1	59	9.8	61	9.9	67	11.4	69	11.3	57	9.0

Table 2 continued

Variable	Group											
	A		B		C		D		E		F	
	f	%	*f*	%	*f*	%	*f*	%	*f*	%	*f*	%
Retired	78	12.3	90	15.0	79	12.8	75	12.8	76	12.4	86	13.6
Unemployed	50	7.9	44	7.3	50	8.1	51	8.7	44	7.2	68	10.7
Others	7	1.1	7	1.2	10	1.6	9	1.6	5	.8	9	1.4
Refusal	1	.2	0	0	0	0	0	0	1	.2.	0	0
Pearson Chi square	$\chi^2(35) = 34.658$, $p = .709$											
Monthly income												
NT9,999 or less	121	19.1	104	17.4	116	18.9	106	18.1	105	17.1	130	20.5
NT10,000–NT29,999	209	33.0	198	33.1	202	32.8	202	34.4	209	34.1	196	30.9
NT30,000–NT49,999	144	22.7	135	22.5	145	23.6	131	22.3	168	27.4	150	23.7
NT50,000–NT69,999	76	12.0	82	13.7	83	13.5	60	10.2	69	11.3	72	11.4
NT70,000 and more	56	8.8	57	9.5	41	6.7	58	9.9	39	6.4	61	9.6
Don't know	9	1.4	7	1.2	13	2.1	15	2.6	9	1.5	10	1.6
Refusal	19	3.0	16	2.7	15	2.4	15	2.6	14	2.3	15	2.4
Pearson Chi square	$\chi^2(30) = 27.586$, $p = .592$											
Total	634	100	599	100.	615	100.	587	100.	613	100.	634	100.
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Self-rated health	3.84	1.21	3.53	1.21	3.64	1.16	3.49	1.08	4.79	1.38	5.01	1.46
Perceived fairness	3.34	1.35	3.05	1.20	3.23	1.25	3.06	1.12	4.21	1.39	4.44	1.62

discrimination parameter describes the strength of the relationship between an item and the measured construct. It also indicates how well an item differentiates among people at different levels along the trait continuum (Reeve and Fayers 2005). A threshold parameter refers to an item's difficulty or location. The lower number of a threshold parameter indicates that the step from one category to the next is easier.

Let $m_i + 1 = K_i$ be equal to the number of item response categories within an item. Two stages are used to compute the category response probabilities in the GRM. The first step is to compute the probability of a respondent's raw item response falling in or above a given category threshold ($j = 1 \ldots m_i$) that is conditional on a trait level (θ). The probability is expressed as

$$P^{*}_{ij}(\theta) = \frac{\exp\left[Da_i\left(\theta - b_{ij}\right)\right]}{1 + \exp\left[Da_i\left(\theta - b_{ij}\right)\right]} \tag{1}$$

The categories within an item are ordered in terms of location, and the steps within an item are completed in order. For Eq. 1, D = 1.0 is used to place parameters on a logistic metric for simplicity (Chen et al. 2013). The second step is to estimate the actual category response probabilities based on the probabilities obtained from the first step. This category response probability can be estimated by subtracting adjacent cumulative probabilities conditional on θ such that

$$P_{ij}(\theta) = P^*_{ij}(\theta) - P^*_{i,j+1}(\theta) \tag{2}$$

The probability of choosing the lowest response category is calculated by subtracting $P^*_{i1}(\theta)$ from 1.0, and the probability of choosing the highest category is calculated by subtracting 0 from $P^*_{im}(\theta)$. Three basic assumptions are required for graded response model. Unidimensionality indicates that all the items measure the same construct. Local independence means that within the same group of respondents with the same value of θ, the distributions of the item responses are independent of each other. Finally, homogeneous discrimination of the items is required. Model fit, including graphical and empirical approaches, can be examined at both the item and individual level (Reeve and Fayers 2005). Item characteristics curves (ICCs), also known as category response curves (CRCs), can be used to reveal the relationship between a person's response to an item and the latent variable. The item information function, as shown by item information curves (IICs), can identify items that perform either well or poorly. While no standard indices are universally acceptable, most can measure deviations between the predicted and observed response frequencies. The MULTILOG 7.3 program that utilizes a marginal maximum likelihood estimation algorithm was employed to estimate the GRM parameters.

3 Descriptions of Sample and Happiness Items

Socio-demographic variables were first compared across groups to ensure that respondents in the six sub-samples were randomly assigned. These variables included gender, age, education, marital status, working status, and monthly income, as shown in Table 2. The sample sizes ranged from 587 to 634 respondents and there were more females than males across the six groups. In most of the groups, respondents who were 40–49 years old had the highest proportion, and the mean age of the respondents was around 48. The majority of the respondents were married and with a senior high school degree. With respect to working status, more than half of the respondents had a full-time job, followed by those retired, except for those in group A where homemakers had slightly higher proportion than the retired. Finally, around 30 % or more of the respondents had a monthly income of between NT$10,000 and NT$29,999 across the groups. The differences among the six groups were insignificant, as indicated by the results of the Chi square test, and evidenced the randomness of the respondents. The description of reversely coded self-rated health and perceived fairness are reported in the bottom of Table 2, with both variables treated as interval measures. The mean values of self-rated health ranged from 3.49 (group D) to 3.84 (group A) for 5-point scales and from 4.79 to 5.01 for 7-point scales, respectively. For perceived fairness, its mean values ranged from 3.06 (group D) to 3.34 (group A) for 5-point scales and from 4.21 to 4.44 for 7-point scales, respectively. These results suggested a general tendency that the respondents rated themselves healthy and felt that current living standards fairly reflected his or her efforts.

With respect to general happiness, the distribution of the six designs for response scale is reported in Table 1. When comparing the distributions of balanced and unbalanced scales with the same number but different labels (A vs. B, C vs. D, and E vs. F), the differences were observable (results not shown). The proportions of unhappy categories in the unbalanced designs were collectively lower than those in the balanced designs. The skewness also indicated skew-to-the-happy-side distribution for all of the designs, especially the 7-point balanced scale that had the highest level of skewness. In contrast, unbalanced scale designs resulted in near-normal distributions as indicated by their lower skewness. Respondents who responded “don’t know” or “refusal” as “non-response” ranged from less than .3–1.1 % among the six happiness items. In addition, the proportions of the midpoint categories for balanced designs were less than 1 %, disregard the number of response scales, which may be attributed to the unfolding technique during the response process. Further analyses employed Samejima’s (1969) Graded Response Model to evaluate item and scale properties for scale improvement. Complete data in which nonresponses were deleted were included in the analyses for each of the response designs.

4 Results of Graded Response Model

The IRT parameter estimates for the six response scales are presented in Table 3, where *a* is the parameter for the discrimination ability of an item and *b* is the threshold parameter. As seen in Table 3, little variation in discrimination strength was found among all the items, except for group F. The items with the lowest discrimination ability were those in Group B and D (both were .75), while the measure in Group F had the highest (1.62). Overall, the single measure of general happiness in the six scales had reasonable discrimination abilities.

Due to the different numbers of categories in these response scales, the number of threshold parameters (the number of categories minus one) varied from four to six. The between-category-threshold parameters appeared to spread out fairly well for all of the items from the negative to the positive. The threshold parameters (e.g., b_2 and b_3 in groups A and C, and b_3 and b_4 in group F) between the middle and its adjacent categories did not change much, due to the fact that only a few respondents had chosen the middle categories as shown in Table 1.

Since the main purpose of this study is to compare different scale designs with respect to the measure of general happiness, it is important to evaluate the appropriateness of response categories. Information characteristics curves (ICCs) help determine how well an item performs and the appropriateness of the response categories. The ICC models the

Table 3 Measurement scale for general happiness: groups A to F

Item	A (balanced 5-point scale I)		B (unbalanced 5-point scale I)		C (balanced 5-point scale II)		D (unbalanced 5-point scale II)		E (unbalanced 7-point scale)		F (balanced 7-point scale)	
	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
a	.86	.09	.75	.08	.86	.09	.75	.08	.89	.08	1.62	.10
b_1	−1.88	.19	−2.88	.31	−2.65	.26	−3.05	.34	−3.28	.31	−2.30	.17
b_2	1.26	.17	−.02	.12	1.16	.16	−.21	.12	−1.63	.17	−.62	.08
b_3	1.30	.17	1.81	.23	1.27	.16	1.89	.24	.02	.10	1.15	.10
b_4	3.10	.34	3.77	.44	3.06	.34	3.62	.43	1.80	.19	1.16	.10
b_5									3.23	.33	2.20	.16
b_6									4.25	.45	2.83	.23

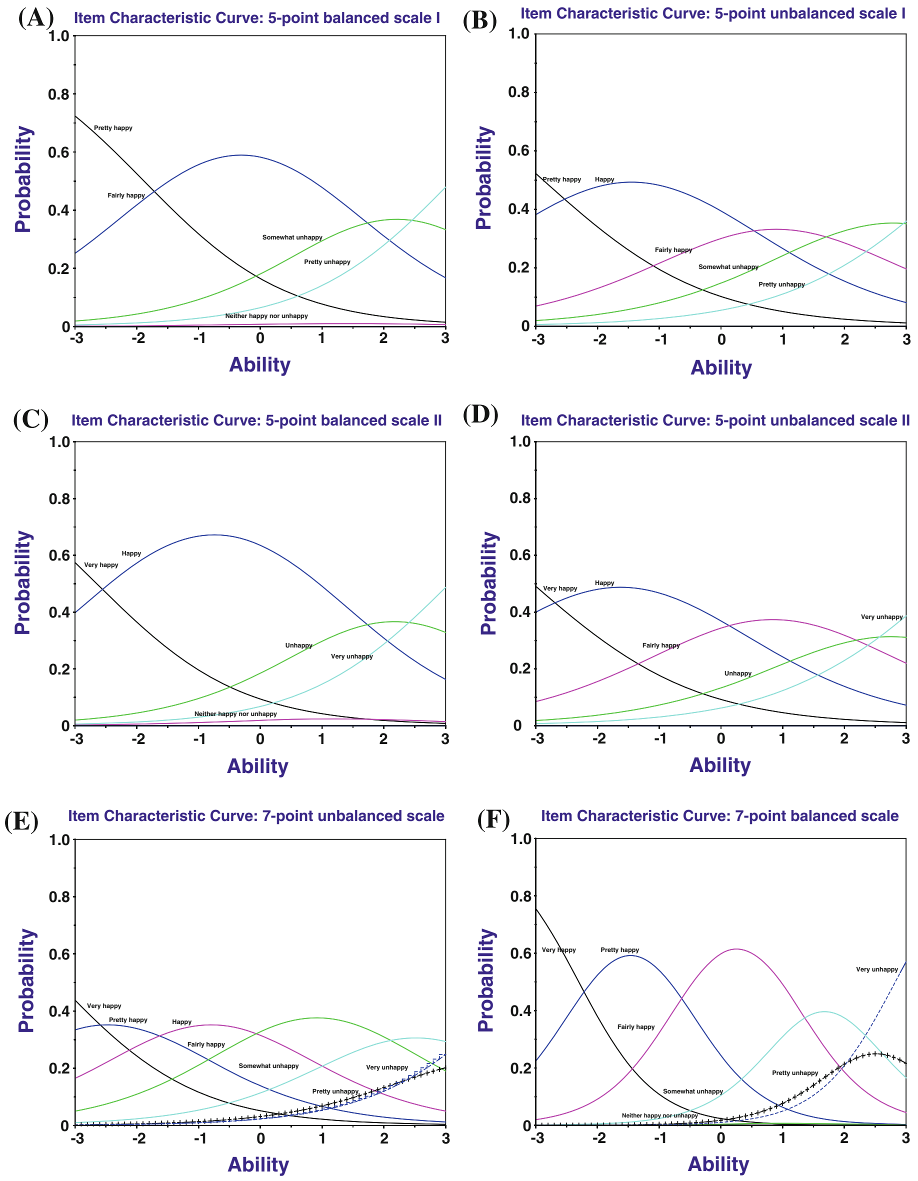

Fig. 1 ICCs for general happiness for the six designs of the response scale. **a** 5-point balanced scale I, **b** 5-point unbalanced scale I, **c** 5-point balanced scale II, **d** 5-point unbalanced scale II, **e** 7-point unbalanced scale, **f** 7-point balanced scale

probability of a person endorsing one of the response categories (for example, from “very happy” to “happy”), conditional on his or her general-happiness level. The ICCs of the six designs indicated that all of the items with different numbers of response categories functioned well and were spread out across response categories (see Fig. 1).

For both of the 5-point balanced scales (A and C), the ICCs showed that the second or less positive categories ("fairly happy" or "happy") were more likely to be chosen by the respondents by covering a relatively larger area under the latent construct. Similarly, the second response category in the 5-point unbalanced deigns also covered a large area of the θ-continuum, although the areas were more evenly divided. A large area under the θ-continuum indicated disproportional information of this dominated category in the 5-point scale when compared to other categories. Adding a category on the positive side may help divide the areas that are more evenly covered by each of the category. On the other hand, the patterns of unevenly divided areas on the positive side were less found in both of the 7-point scales. These findings indicated that three or more categories for the positive responses are desirable for the single-item measure of happiness.

In addition, some of the response categories in the scales could be dropped. In both of the 7-point scales, the category "unhappy" (the second from the right in Fig. 1) was overshadowed by the last category "very unhappy," indicating that this response category was not as likely to be chosen as its adjacent categories. When the category "very unhappy" was used as an end point category in Groups C and D, the area was overshadowed by "unhappy" as well, although the overlapped areas were smaller here. In both of the 7-point designs, the number of the category on negative or unhappy side of the scale was both three. It is thus appropriate to declare that two categories on the negative or unhappy side can work sufficiently well.

As it was not voluntarily provided, the middle response was not utilized by the subsamples. This suggests that including such a choice provides only limited information, as indicated by the nearly flat lines for Groups A, C, and F in Fig. 1. This finding may not contribute much to the use of unbalanced response design, however, because of the possible contamination of unfolding technique during the interviewing process.

With respect to response labels, item information curves (IICs) were compared between Groups A and C, as well as between Groups B and D. IICs are used to evaluate the overall performance of an item. When two items use the same wording, the information it provides may differentiate how well different response scales perform. Figure 2 presents IICs for the scales with the same number of categories but different response labels. The information functions within either of the comparisons, however, were not observably different, as indicated by the discrimination estimates in Table 3. Similar results can be found in Fig. 1 where ICCs for the positive side of happiness scale differed between Groups A and C but no considerable difference was observed between Groups B and D; neither in the comparisons for the negative side of scales. Such results suggest that the number of response categories is more relevant than the label in scale design to reveal various attitudes among respondents.

5 Ordered Logit Regression on Happiness

The effects of demographic variables, self-rated health, and perceived fairness on the different response scales of happiness were examined. All of the response designs of happiness were reversely coded with higher values indicating more happiness. The equation of general happiness can be estimated as:

$$HAPPY = \chi'\beta + \varepsilon \tag{3}$$

where HAPPY is general happiness; and χ' is the vector of the explanatory variables. Since the happiness measures were either 5- or 7-point scales, in which the highest

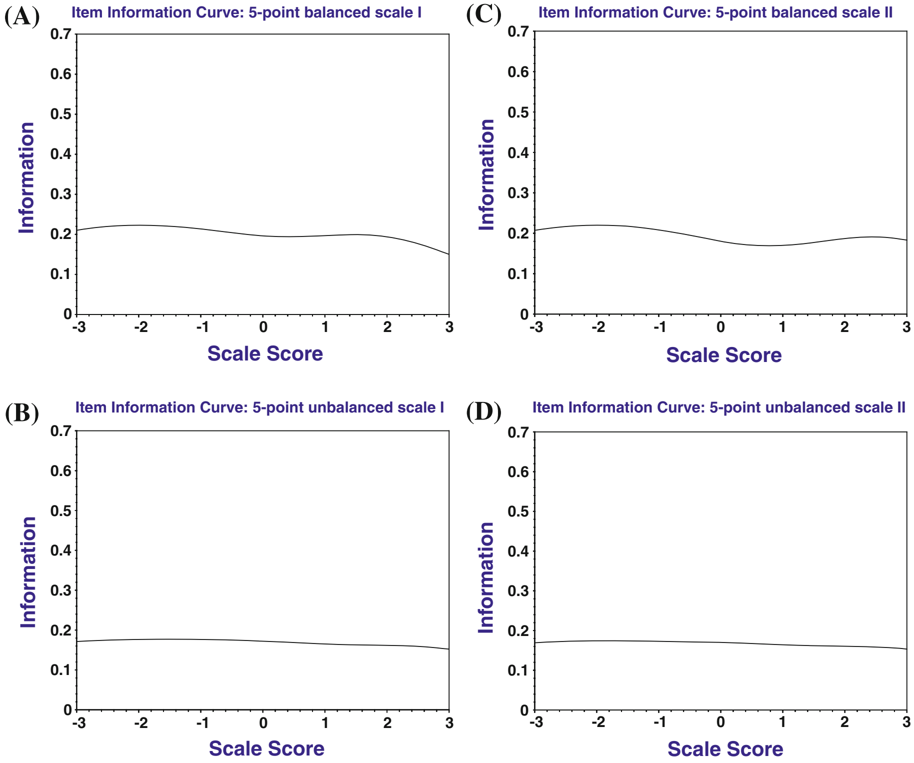

Fig. 2 IICs for general happiness for the 5-point scales. **a** 5-point balanced scale I, **b** 5-point un balanced scale I, **c** 5-point balanced scale II, **d** 5-point unbalanced scale II

value indicates the happiest, an ordered logit regression model is employed to estimate Eq. (3).

Sample sizes of the six groups ranged from 527 to 559 respondents were used for the empirical estimation of the ordered logit model for general happiness. Table 4 presents the maximum likelihood estimation results of the model. The reference category of gender, marital status, working status, and income were female, divorce/widow, not working full-time nor part-time (other), and NT9,999 or less, respectively. As expected, the significant effect of age on happiness was negative while that of age-square was positive across the six groups, meaning that there exists a noteworthy U-shaped relationship between age and happiness. In other words, the younger respondents and the elderly in Taiwan were happier than the middle-aged. Other socio-demographic characteristics had significant effects on happiness in some of the groups, where males and the better educated were less happy than their counterparts. Also, those who were married or living with a partner, working full-time, those who had higher income were happier than their counterparts, while being single had a mixed effect on happiness.

The contribution of health to happiness was, again, demonstrated in this study, regardless of the labels and the number of response categories. Respondents who rated themselves healthier were happier than those who rated less healthy. In addition, those who felt that current living standards fairly reflected his or her efforts than those who did not feel so. Model fit indicated by −2 log likelihood suggested that the 7-point unbalanced

Table 4 Ordinal logit regression on happiness

Variables	Group					
	A	B	C	D	E	F
Age	−.123**	−.131**	−.083*	−.134**	−.124**	−.111**
Age square	.001**	.001**	.001*	.002**	.001**	.001**
Education (year)	−.087**	−.000	−.078*	−.044	−.034	−.034
Gender						
Male	−.398*	−.132	−.114	−.405*	−.587***	−.143
Female[a]	–	–	–	–	–	–
Marital status						
Single and never married	.725†	−.805*	1.172**	.525	.112	−.892*
Married or living w/partner	.675*	−.349	1.098***	.237	.220	−.597†
Divorced/widow/other[a]	–	–	–	–	–	–
Working status						
Full-time	−.266	.203	.554*	.320	.353	.140
Part-time	−.360	−.070	.008	−.218	−.096	.323
Other[a]	–	–	–	–	–	–
Income						
NT9,999 or less[a]	–	–	–	–	–	–
NT10,000–NT29,999	.688 *	.613**	.274	.392	.051	−.027
NT30,000–NT49,999	.917*	.575*	−.323	.400	−.039	.409
NT50,000–NT69,999	.948*	.464	−.077	.711†	.201	.689†
NT70,000 and more	.905*	.870*	−.505	.957*	.045	.468
Health	.536***	.438***	.446***	.576***	.490***	.391***
Fairness	.858***	1.086***	1.058***	1.143***	.752***	.824***
Constant 1	−1.97	−1.32	.12	−.46	−2.33*	−2.37
Constant 2	.18	.38	2.19*	1.10	−1.21	−1.36
Constant 3	.20	2.82*	2.30*	3.38**	.34	.50
Constant 4	4.92**	4.97***	6.16***	6.03**	2.38*	.57
Constant 5					4.14***	3.51**
Constant 6					5.72***	8.81***
−2 log likelihood	983.761	1,267.435	1,019.171	1,210.747	1,563.638	1,315.872
n	557	538	555	527	555	559

† $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$

[a] Reference category

scale (group E) had the highest goodness of fit among the six response designs. Similar results were obtained for models on 5-point scales, where unbalanced scales had better model fit than the balanced.

6 Conclusion and Discussion

The response design for the single-item measure of general happiness has been a focus mainly due to its widespread use in surveys (Cummins 2003; Cummins and Gullone 2000), as well as comparison difficulty caused by different versions of response scales (Bjørnskov 2010; Lim 2008). Given the negatively skewed distribution of response to general happiness worldwide (Cummins 2003), it is meaningful to examine the possibility of constructing an unbalanced design for such a measure with more positive categories than negative ones in the response scale. Instead of replacing existing measures for general happiness or SQoL, this study examines alternative response scales that may result in near-normal distributions, corresponding to the assumption of commonly used statistical approaches. With the consideration of the number of response categories and the corresponding labels, survey data collected by CATI were used to compare balanced and unbalanced response designs for general happiness.

In deciding the number of response categories, the results of the Graded Response Model indicates that a higher number of response categories are needed on the positive side than those on the negative side to better reveal general happiness with a single measurement. It is suggested that a scale with at least three happy categories and no more than two unhappy categories can function sufficiently well. On the other hand, a response scale with 4 positive categories may also apply to the case given its highest goodness-of-fit in ordered logit regression models when explained by socio-demographic characteristics, self-rated health, and perceived fairness (Table 4). Although previous studies suggested using a balanced design for the response scale to minimize measurement error (Frisbie and Brandenburg 1979; Lam and Klockars 1982; Saris and Gallhofer 2007), this study demonstrates the usefulness of an unbalanced design for the response scale when applied to a positive-oriented construct such as general happiness (Cummins 2003; Veenhoven 1993).

The suggested number of response categories for an unbalanced design should be limited to the use of a Likert-type rating scale and the method of telephone interviews where verbal labels are read out to the respondents. Although more number of categories can be used in face-to-face interviews, it is often difficult for telephone respondents to choose from many alternatives (Schaeffer and Presser 2003). Previous research has suggested a higher number of response categories for better reliability differentiation among respondents' attitudes (Alwin 1997; Cummins 2003), such as numeric rating scales with endpoints labeled. However, respondents often have different interpretations for those unlabeled categories (Schaeffer and Presser 2003). One may obtain happiness scores from numeric scales but the affection feelings of happiness (see Michalos et al. 2000; Veenhoven 1996) may not be revealed. Instead, Likert-type rating scales provide an opportunity for respondents to disclose how they feel.

With regard to the response labels, this study indicates that labels with different levels of intensity provide similar information, regardless of the un/balanced design of response scales. This finding is not consistent with previous studies which suggested significant effects of response labels, even if there is only a subtle difference among the intermediate categories (Lam and Klockars 1982; Schaeffer and Charng 1991; Wildt and Mazis 1978). However, results from previous studies were obtained from data either on different topics, or drawn from non-representative samples. Since the effect of label intensity was decided by sensitivity of the content and intensity of questions, the results may not to be concluded.

While the findings are generally in favor of unbalanced scale design, the balanced response designs may have their advantages during the interview process. Respondents may get used to the common practice pretty soon, particularly when most response scales of other questions are in the same form, and then reduce their cognitive burden for answering survey questions (Schaeffer and Presser 2003). In addition, balanced response designs indeed collectively provide more information than the unbalanced ones as indicated by the IICs in the GRM results (figure not shown), although the differences are relatively small. Researchers need to thoroughly consider the pros and cons of different designs of response scales before choosing an appropriate one for their studies.

As it was also shown by the GRM results, a midpoint category provided only limited information when using the unfolding approach to collect survey data. Among the three designs where a middle category was included, only a few respondents had chosen such an answer, which may be the contaminating effect of unfolding technique. This finding is consistent with previous studies where a midpoint category was omitted (Bishop 1987; Moors 2007) and it also supports the use of unbalanced design without a midpoint category. However, it may raise the question why a midpoint category is included if its function is not anticipated. When a midpoint category is offered or being recognized, it certainly provides information about respondents' opinions to some extent.

With regard to the effect of "usual suspect" (Frijters and Beatton 2012) as well as self-rated health and perceived fairness, on happiness, the results were consistent with previous studies. Age, age square, self-rated health and perceived fairness were significantly associated with general happiness, regardless of the design of response scale (Argyle 1997; Frijters and Beatton 2012; Liao et al. 2005; Michalos et al. 2000; Tsou and Liu 2001). It is evidenced that using an unbalanced response design for the single measure of happiness can achieve representative reliability and criterion validity to some extent. Again, more empirical evidence is in need to obtain stable results.

Acknowledgments This study is sponsored by the National Science Council, Taiwan (NSC 98-2628-H-001-010-SS2). The author thanks Willem E. Saris and anonymous reviewers for their helpful comments on an earlier version. Please direct all correspondence to the author.

References

Alwin, D. F. (1992). Information transmission in the survey interview: Number of response categories and the reliability of attitude measurement. *Sociological Methodology, 22*, 83–118.

Alwin, D. F. (1997). Feeling thermometers versus 7-point scales: Which are better? *Sociological Methods & Research, 25*, 318–340. doi:10.1177/0049124197025003003.

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods & Research, 20*, 139–181. doi:10.1177/0049124191020001005.

Argyle, M. (1997). Is happiness a cause of health? *Psychology and Health, 12*, 769–781.

Bishop, G. F. (1987). Experiments with the middle response alternative in survey questions. *Public Opinion Quarterly, 51*, 220–232. doi:10.1086/269030.

Bjørnskov, C. (2010). How comparable are the Gallup world poll life satisfaction data? *Journal of Happiness Studies, 11*, 41–60. doi:10.1007/s10902-008-9121-6.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement, 18*, 205–215.

Chen, S. K., Hwang, F. M., & Lin, S. S. J. (2013). Satisfaction ratings of QOLPAV: Psychometric properties based on the graded response model. *Social Indicators Research, 110*, 367–383. doi:10.1007/s11205-011-9935-1.

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology, 56*, 754–761. doi:10.1037/0022-006X.56.5.754.

Cummins, R. A. (2003). Normative life satisfaction: Measurement issues and a homeostatic model. *Social Indicators Research, 64*, 225–240. doi:10.1023/A:1024712527648.

Cummins, R. A., & Gullone, E. (2000). Why we should not use 5-point Likert scales: The case for subjective quality of life measurement. In *Proceedings, second international conference on quality of life in cities* (pp. 74–93). Singapore: National University of Singapore.

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research, 50*, 61–77.

Diener, E. (1984). Subjective well-being. *Psychological Bulletin, 95*, 542–575.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ferrer-i-Carbonell, A., & Frijters, P. (2004). How important is methodology for the estimates of the determinants of happiness? *The Economic Journal, 114*, 641–659.

Fordyce, M. W. (1988). A review of research on the happiness measures: A sixty second index of happiness and mental health. *Social Indicators Research, 20*, 355–381.

Fox, M. T., Sidani, S., & Streiner, D. (2007). Using standardized survey items with older adults hospitalized for chronic illness. *Research in Nursing & Health, 3*, 468–481. doi:10.1002/nur.20201.

Francis, J. D., & Busch, L. (1975). What we know about "I don't knows. *Public Opinion Quarterly, 39*, 207–218.

Frijters, P., & Beatton, T. (2012). The mystery of the U-shaped relationship between happiness and age. *Journal of Economic Behavior & Organization, 82*, 525–542. doi:10.1016/j.jebo.2012.03.008.

Frisbie, D. A., & Brandenburg, D. C. (1979). Equivalence of questionnaire items with varying response formats. *Journal of Educational Measurement, 16*, 43–48. doi:10.1111/j.1745-3984.1979.tb00085.x.

Holbrook, A., Cho, Y. K., & Johnson, T. (2006). The impact of question and respondent characteristics on comprehension and mapping difficulties. *Public Opinion Quarterly, 70*, 565–595. doi:10.1093/poq/nfl027.

Hung, Y.-T. (2001). □□□□□□□ *(Respondent selection within household)*. Taipei: Wu Nan.

Kalmijn, W. M., Arends, L. R., & Veenhoven, R. (2011). Happiness scale interval study: Methodological considerations. *Social Indicators Research, 102*, 497–515. doi:10.1007/s11205-010-9688-2.

Kroh, M. (2007). Measuring left–right political orientation: The choice of response format. *Public Opinion Quarterly, 71*, 204–220. doi:10.1093/poq/nfm009.

Lam, T. C. M., & Klockars, A. J. (1982). Anchor point effects on the equivalence of questionnaire items. *Journal of Educational Measurement, 19*, 317–322.

Liao, P.-S., Fu, Y.-C., & Yi, C.-C. (2005). Perceived quality of life in Taiwan and Hong Kong: An intra-culture comparison. *Journal of Happiness Studies, 6*(1), 43–67. doi:10.1007/s10902-004-1753-6.

Lim, H. E. (2008). The use of different happiness rating scales: Bias and comparison problem? *Social Indicators Research, 87*, 259–267.

Masters, J. R. (1974). The relationship between number of response categories and reliability of Likert-type questionnaires. *Journal of Educational Measurement, 11*, 49–53. doi:10.1111/j.1745-3984.1974.tb00970.

Mazaheri, M., & Theuns, P. (2009a). Effects of varying response formats on self-ratings of life-satisfaction. *Social Indicators Research, 90*, 381–395.

Mazaheri, M., & Theuns, P. (2009b). Structural equation modeling (SEM) for satisfaction and dissatisfaction ratings; multiple group invariance analysis across scales with different response format. *Social Indicators Research, 91*, 203–221.

Michalos, A. C., Zumbo, B. D., & Hubley, A. (2000). Health and the quality of life. *Social Indicators Research, 51*, 245–286.

Moors, G. (2007). Exploring the effect of a middle response category on response style in attitude measurement. *Quality & Quantity, 42*, 779–794. doi:10.1007/s11135-006-9067-x.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1–15. doi:10.1016/s0001-6918(99)00050-5.

Raaijmakers, Q. A. W., van Hoof, A., Hart, H., Verbogt, T. F. M. A., & Vollebergh, W. A. M. (2000). Adolescents' midpoint responses on Likert-type scale items: Neutral or missing values. *International Journal of Public Opinion Research, 12*, 208–216.

Reeve, B. B., & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire item and scale properties. In P. Fayers & R. Hays (Eds.), *Assessing quality of life in clinical trials* (2nd ed., pp. 55–73). Oxford: Oxford University Press.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph,* No. 17.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). NY: Springer.

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

Schaeffer, N. C., & Charng, H.-W. (1991). Two experiments in simplifying response categories: Intensity ratings and behavioral frequencies. *Sociological Perspectives, 34*, 165–182.

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology, 29*, 65–88. doi:10.1146/annurev.soc.29.110702.110112.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.

Tourangeau, R., Raslnski, K. A., & Bradburn, N. (1991). Measuring happiness in surveys: A test of the subtraction hypothesis. *Public Opinion Quarterly, 55*, 255–266.

Tsou, M., & Liu, J. (2001). Happiness and domain satisfaction in Taiwan. *Journal of Happiness Studies, 2*, 269–288.

Veenhoven, R. (1993). *Happiness in nations*. Rotterdam: Risbo.

Veenhoven, R. (1996). Developments in satisfaction research. *Social Indicators Research, 37*, 1–45.

Wedell, D. H., & Parducci, A. (1988). The category effect in social judgment: Experimental ratings of happiness. *Journal of Personality and Social Psychology, 55*, 341–356.

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test–retest reliability. *Educational and Psychological Measurement, 60*, 956–972. doi:10.1177/0013164404268674.

Wildt, A. R., & Mazis, M. B. (1978). Determinants of scale response: Label versus position. *Journal of Marketing Research, 15*, 261–267.