Presentation of a Single Item versus a Grid: Effects on the Vitality and Mental Health Scales of the SF-36v2 Health Survey

Mario Callegaro, Jeffrey Shand-Lubbers & J. Michael Dennis
Knowledge Networks, 1350 Willow Road, Suite 102, Menlo Park, CA 94025

Abstract

There is a lot of debate about whether questions should be presented on a grid or in a single item per screen. Operationally, grids take less time for respondents to complete. Use of grids should decrease response burden, although new research shows that respondents seem to prefer a single item per screen. From a measurement point of view, grids pose numerous issues: higher item non-response, higher item non-differentiation, and sometimes higher measurement error.

In this experiment, we are testing the Vitality (4 items) and Mental Health (5 items) scales of the SF-36v2® Health Survey. The SF-36v2 asks 36 questions to measure functional health and well-being from the patient's point of view. It is called a generic health survey, because it can be used across age (18 and older), disease, and treatment groups, as opposed to a disease-specific health survey which focuses on a particular condition or disease. Two of the four items of the vitality scale and two out of five items of the mental health scale are reversed in scoring.

A sample of 2,500 KnowledgePanel® respondents was randomly assigned to one of five experimental conditions: Group 1: Standard grid; Group 2: Shaded grid; Group 3: One item per screen with horizontal response options; Group 4: One item per screen with vertical response options; Group 5: One item per screen with vertical shaded response options. Approximately 360 respondents completed the survey per condition for a completion rate of 73.4%. The survey was optimized to be seen on a screen with minimum resolution of 800 by 600 pixels. During the study we collected the browser type for each respondent. This allowed us to exclude cases in which the survey was taken either on a MSNTV or on an iPhone/PDA because they could not properly see the grid items. The final sample used for the analysis, after exclusions, was of 1,419 cases for an average group size of about 280.

We hypothesized that items presented on a grid would lead to more measurement error as indicated by a higher rate of "inconsistencies" in the self-reports to grid questions and a lower rate of inconsistencies in the self-reports to the single-item questions. We speculated that presenting items on a single screen allows the respondent to bring more cognitive focus to each question and therefore be more consistent in their answers to questions. In contrast, when items are on a grid, it is easier for the respondent to get confused, especially when the meaning of some of the items is reversed. We computed an index of consistency by correlating the total sum of scores for the reversed items with the total sum of scores for the non-reversed items. If respondents are consistent in their answers the correlation between reversed and non-reversed should be higher. We calculated Cronbach's alpha scores to measure consistency in answers for each of the five experimental conditions.

The direction of the study findings were consistent with our hypotheses -- lower alpha level for the grid presentation and higher correlation for the single-item presentation -- although the differences among groups do not reach statistical significance.

Key words: grid, matrix, single item per screen, Cronbach's alpha, SF36v2, KnowledgePanel

1. Introduction

When moving from a self-administered paper and pencil questionnaire to a Web survey, researchers face design questions with their questionnaire layout. One issue is if the instrument should be presented with one question per screen or multiple questions per screen. A related issue is if they need to mimic as much as they can the paper design in the Web version.

The general goal is to assure equivalence of electronic measure and off-line measures (Roberts, 2007). When comparing an offline with an online instrument, we need to verify that the two populations are equivalent, so we can concentrate on the design features of the Web instrument.

2. Studies comparing one item per screen versus multiple items or grids

Usability guidelines suggest the minimum use of vertical scrolling and advice to use paging. Horizontal scrolling should be eliminated altogether (U.S. Department of Health and Human Services, 2006). Although horizontal scrolling can be easily eliminated with careful programming and testing, the question about placing one item per screen versus multiple items still applies.

In a paper questionnaire, respondents can look at the entire questionnaire and then answer the questions. They can also look ahead and possibly change some answers based on later questions. Early research done on paper-and-pencil questionnaires in comparison to a telephone interview (where by definition the questions are read one at a time) showed that later items influenced answers on previous answers (Schwarz & Hippler, 1995). These results were also reproduced in a Web survey where questions were placed either on a single page or on consecutive pages (Reips, 2002).

The majority of experiments comparing single versus multiple questions per screen use grids or matrices for the latter. Although grids are commonly used in Web instruments, many authors caution against them. Dillman, Smith and Christian (2009, p. 179), for example, advise to minimize the use of matrices for four main reasons: a) they require respondents to match information in rows with questions in columns (or vice versa), a quite complex task; b) the instructions to fill out a grid are generally difficult to understand; c) the structure of the matrix leaves it up to the respondent as to whether to navigate it and fill in either rows or columns or a combination of both; d) there is an increased likelihood to miss items. One issue in comparing grids versus single question per screen is how to visually lay out the response options. In a grid environment generally the response options are organized by columns, while the questions are organized by row. In a one question per screen format, response options can be organized vertically or horizontally.

In one of the first experiments comparing grids versus single questions per screen in an online survey, Couper, Traugott and Lamias (2001) contrasted a five-item knowledge measure administered with a fully labeled five point approve-disapprove scale in a grid format versus five single pages with response options laid out vertically. They also used an 11 item measure with a five-point Likert answer scale. In one condition, each item was presented on a single screen (11 screens), while in the other condition the items were split into three screens with four, four, and three items per grid per page. Although the correlation between items presented in the grid (for both knowledge and Likert scale) was slightly higher, the difference among the two groups did not reach statistical significance. The study was conducted with of 8,747 University of Michigan students.

In another experiment, Tourangeau, Couper and Conrad (2004, experiment 6) randomized 2,568 opt-in non-probability panel respondents into three groups. In the first group, four questions about diet (attitude towards healthy diet) and four questions about eating habits (health diet behavior) were presented on a single screen in a grid format. In the second group the questions were presented on two screens with a four item grid per screen. In the last group the questions were presented one per screen. The response option was a seven point agree – disagree scale with endpoint-only labels. The responses presented on a single grid obtained a higher Cronbach's alpha value (.621) then when presented in two screens (.562) or in one single screen (.511). The differences among groups were statistically significant. However, the responses in the single grid were less differentiated. In other words, the score of item non-differentiation (proportion of items that obtain the same response) was significantly higher for the single grid (mean = .436) in comparison to the second (mean= .422) and third group (mean= .412). In a new analysis of the same dataset, Peytchev (2006) used structural equation modeling (SEM) to test the hypothesis that a single question per page yields better measurement than the same questions in a grid. Better measurement should allow better identification of the two latent constructs (attitude towards healthy diet and health diet behavior). This was in fact the case: the standardized path coefficient (validity coefficient) of the SEM was found to be consistently higher as questions were separated on different screens increasing from 0.82 to 0.94.

Yan (2005) compared three different versions of six items about various risky behaviors on a sample of 2,587 opt-in panel members coming from Survey Spot and American Online Opinion Place (experiment 1). In one condition the items were presented on single screens, in another the items were presented on a single screen but separated (not in a grid) in a scrolling design. In the last condition the items were presented on a grid. She did find an increase in Cronbach's alpha going from a single item per screen to the grid although the difference among groups was marginally significant (p=0.98).

In paper-and-pencil tests with 30,000 residential addresses of two forms of the American Community Survey (Chesnut, 2008), the Census Bureau tested two different layouts to collect demographic information: a sequential layout verses a grid layout. In the first case demographic information for each household member is organized sequentially, with two columns per page, one column for each member. In the grid format information for each household member is organized per rows, with the row continuing in the next page (Appendix B and C). The sequential presentation yielded a response rate 1.5 percentage points higher by also lowering the percent of item nonresponse.

Toepel, Das, and van Soest (2009) experimented with 40 items of the Arousal Seeking Tendency scale, an instrument designed to measure involvement, in the Dutch probability

based online panel CentERpanel with a final response from 2,565 panel members. The response options were ranging from "totally disagree" to "totally agree" on a five point scale laid out horizontally and fully labeled. Respondents were randomly placed into one of four conditions: one item per screen, four items per screen in a grid format (with no grid borders), 10 items per screen in a grid format, and 40 items per screen in a grid format. They did find a modest increase of Cronbach's alpha when items were placed together. They also found a higher amount of item nonresponse the more questions were placed on a screen.

Thorndike and colleagues (2009) had 710 participants in Sweden seeking self-help treatment on the Internet to complete an online questionnaire composed of four set of items (Beck Depression Inventory, Beck Anxiety Inventory, Quality of Life Index and Montgomery-Asberg Depression Rating Scale). The questionnaire was either presented with one item per screen, or with multiple items per screen. The same participants were asked to repeat the instrument no sooner than one hour and no later than four hours. The order of two questionnaires was counterbalanced. SEM analysis showed that the scales were functioning to measure the same constructs regarding presentation format or order. The majority of participants, however, when asked which format they preferred, chose the single item per screen even if that format required more time.

Lastly, Garland (2009) assigned SurveySpot panelists to three conditions: a questionnaire with grid format, a questionnaire with multiple items per screen and a questionnaire with single item per screen. Satisfaction by version did not change, while the distribution of answers significantly changed across conditions. In order to measure data quality, factor analysis was run on a subset of items showing that the single item per screen version yielded double the variance explained in comparison to the two other versions and also yielded the exact number of factors that the questionnaire was supposed to measure. Interestingly and contrary to the previous studies, Cronbach's alpha was higher for the single item per screen in comparison to the grid versions.

2.1 Specific studies on SF-36 and SF-12

In a study conducted by Bell, Mangione & Kahn (2001), 4,876 consenting visitors to a health Web survey were randomly assigned to the SF-36 Health Survey in either a version where the questions were presented as multiple items per screen with response option laid out vertically, or to a version where questions sharing the same response options were presented in a grid with response options organized per columns. The presentation of multiple items per page did not yield lower Cronbach's alpha scores for the eight SF-36 scales in comparison to the other version.

In a paper and pencil study, Iglesias, Birks and Torgerson (2001) tested two versions of the SF-12v2 in sample of 422 women 70 years old or older. In the first version, most items were organized in grids with response options organized per columns. In their modified version, items were presented one at a time with response options presented horizontally for each question. The grid version had a significantly higher number of item missing (non response to any item), 26.6% in comparison to 8.5%.

2.2 Summary of single versus multiple items per screen

From all the studies reviewed, it appears that in most cases placing multiple items per screen increases inter-item correlations. Grids seem to increase item nonresponse, and respondents at least in two studies do not particularly like them even if filling them out goes faster than when questions are on one page per screen. We have then a measurement

issue: respondents view items on grid as belonging together but that does not mean that the answers are more valid. It also appears that more measurement error is found in grids versus single items per page.

3. Hypotheses

Based on previous findings we hypothesized in this study the following:

- Grid conditions should obtain a lower Cronbach's alpha level due to the items reversed in meaning (description later) which direction might be more likely to be confused in a grid in comparison to a single screen;
- Grid answers should be "less consistent" than single-item answers because of the reversed direction of four items;
    - An example of an "inconsistent" respondent: Reports that he or she feels "full of life all of the time," yet answers another question by reporting a feeling of "tired all of the time."
- Grid conditions should elicit higher item nonresponse than single-item presentation; and
- Single-item conditions should obtain higher satisfaction than grid presentations.

4. Experimental design

A sample of 2,500 KnowledgePanel® respondents was randomly assigned to one of five experimental conditions:

- Group 1: Standard grid
- Group 2: Shaded grid
- Group 3: One item per screen with horizontal response options
- Group 4: One item per screen with vertical response options
- Group 5: One item per screen with vertical shaded response options

In this experiment, we employed the Vitality and Mental Health scales of the SF-36v2® Health Survey (Ware, et al., 2007; Ware & Sherbourne, 1992). The SF-36v2 asks 36 questions that measure functional health and well-being from the patient's point of view. It is called a generic health survey because it can be used across age (18 and older), disease, and treatment groups, as opposed to a disease-specific health survey which focuses on a particular condition or disease. Figure 1 shows the measurement model of the items used in the experiment. The appendix reports the exact question wording and direction of the scales.

The questionnaire also asked about difficulty and enjoyment in completing the questionnaire (two more questions). In Figure 2 we show the actual Vitality and Mental Health items as they appeared on the screen. As we can see, two items in each scale have the meaning reversed in comparison to the general direction of the scale. In the Vitality scale, two items measure fatigue, while two items measure energy and are reverse scored. In the Mental Health scale, three items measure negative affect, while two items measure positive affect and are reverse scored.

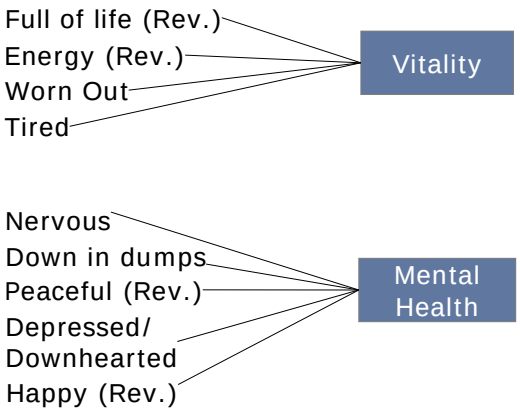


Figure 1: Measurement model for the vitally and mental health scale

Note: this is only a portion of the SF-36v2 model. SF-36v2®is a trademark of the Medical Outcomes Trust. Items indicated as "Rev." are reverse scored.

In Figure 2 we show the actual screens that were seen by the respondents in each group. The script did not prompt for any item nonresponse.

The survey completion rate for the KnowledgePanel study (Callegaro & DiSogra, 2008) was of 73.4%. The survey was optimized to be seen on a screen with minimum resolution of 800 by 600 pixels. During the study we collected the type of Internet browser used by each respondent to complete the instrument. Respondents who completed the study either on a MSNTV or on an iPhone/PDA were excluded because they could not completely see the Group 1 or 2 versions of the screen without scrolling. The final sample used for the analysis, after browser type exclusions and a very limited number of cases with missing values, was of 1,419 cases for an average group size of about 280.

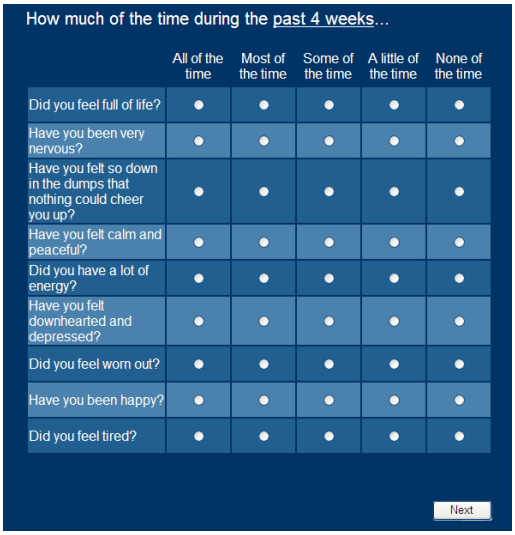
The survey took about 45 seconds for the grid groups and about 70 seconds for the single item per screen groups as shown in Table 1.

Table 1. Median length of the 9 items of the Vitality and Mental Health scale by condition (in seconds)

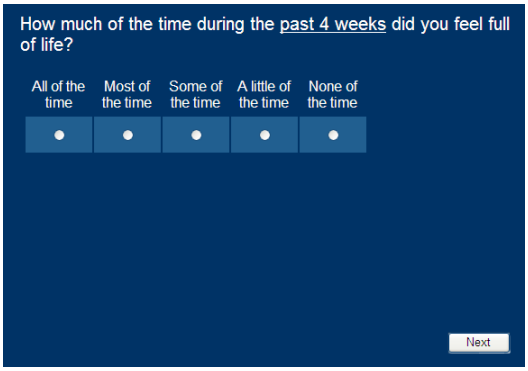
	Grid		**Single item per screen**		
	Group 1	**Group 2**	**Group 3**	**Group 4**	**Group 5**
Median (in seconds)	44	46	69	69	72

Group 1

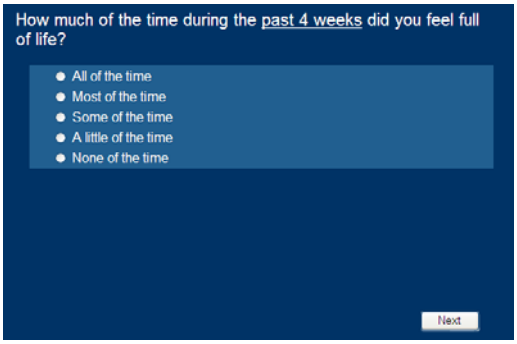
Group 2



Group 3



Group 4



Group 5

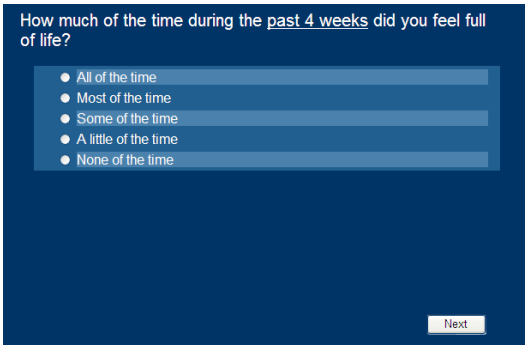


Figure 2: The five experimental group conditions.

5. Results

In order to test the first hypothesis that a single item per screen should obtain a higher Cronbach's alpha level, we computed alpha and standard errors (Duhachek & Iacobucci, 2004) on each scale for each group and used formula 4 and SPSS code to compare the alpha level of two groups. Cronbach's alpha is an index of consistency in the self-reports of survey respondents to questions with related meanings. If respondents answering questions in a grid are more likely to "miss the meaning" of an item, the alpha should be lower; a higher score indicates more consistency in survey responses for a respondent.

Table 2. Cronbach's alpha level for each experimental group and combined (grid vs. single item) on the mental health and vitality scale

			VITALITY				
	Grid		**Single item per screen**				
	Group 1	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**GRID**	**SINGLE**
Alpha	0.842	0.803	0.867	0.855	0.831	0.823	0.851^{ns}
St. Error	0.016	0.019	0.013	0.015	0.017	0.012	0.009
N	284	287	283	268	297	571	848
			MENTAL HEALTH				
	Grid		**Single item per screen**				
	Group 1	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**GRID**	**SINGLE**
Alpha	0.840	0.855	0.883	0.867	0.854	0.849	0.868^{ns}
St. Error	0.015	0.014	0.011	0.013	0.013	0.010	0.007
N	284	287	283	268	297	571	848

To avoid all the possible combinations of comparisons (10) we first tested if there were statistically significant differences among the grid versions or among the single item versions. We did not find any, and therefore we compared the combined grids groups versus the combined single item per screen groups. Although the trend is going in the expected direction, that is, higher alpha level for single items in comparison to the grid, this group difference did not reach statistical significance.

In order to further test the hypothesis that grids introduce more measurement error, we took advantage of the fact that some items in each scale were reversed in meaning. If the grid is making it more difficult for the respondents to comprehend the meaning of the items, answers in a grid should be more "inconsistent" than answers from single screens. We operationalized this concept by computing the correlation between the sum of the "reversed" items in each scale and the sum of the other items. If respondents are inconsistent in their answers, the correlation should be lower. For example, we hypothesize that respondents answering a question in a single-item format will be less likely to report that they have a lot of energy "all of the time" and simultaneously report that they are worn out "all of the time."

Table 3. Spearman's correlations between reversed and straight items for each experimental group and combined (grids vs. single item) on the mental health and vitality scale

				MENTAL HEALTH			
	Grid			**Single item per screen**			
	Group 1	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**GRID**	**SINGLE**
Spearman	0.607	0.617	0.649	0.608	0.547	0.535	0.593[ns]
N	284	287	283	268	297	571	848
				VITALITY			
	Grid			**Single item per screen**			
	Group 1	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**GRID**	**SINGLE**
Spearman	0.602	0.467[a]	0.624	0.690	0.613	0.614	0.652[ns]
N	284	287	283	268	297	571	848

[a] Z=2.259 p<0.05

Because of the nature of the response options (ordinal scale) we used the Spearman's rank order correlation coefficient. In order to test if the two groups of variables (reversed and straight) are equally correlated among the different groups, a Fisher's Z transformation was applied to the Spearman's correlation coefficients. As with the previous table we first tested if there were statistically significant differences among the grid versions or among the single item versions. Only for the vitality scale, group 1 was significantly different than group 2 (grid vs. shaded grid), and all the other comparisons did not reach statistical significance.

To test the third hypothesis (higher number of item nonresponse in the grid presentations) we combined group 1 and 2 and group 3, 4, and 5 due to the very low number of item nonresponse. We did find an extremely low number of item nonresponse in every group.

Lastly, we computed perceived difficulty and "enjoyment" scores for each condition. We found no difference across group administration.

6. Discussion

Items on a grid appear to increase measurement error, although the differences did not reach statistical significance. Presenting items one by one may allow respondents to focus more on the question and the response options than when presented together as a grid.

It appears that on a grid there are more chances to "miss" the meaning of the items, resulting in more inconsistencies than when evaluating one item per screen. In fact the correlations across reversed and straight items did go in the expected direction of the hypotheses, but we did not have enough statistical power to detect these differences. We are looking forward to new studies with a similar design and larger sample sizes to provide further evidence.

Lastly, we did not find any higher scores in terms of difficulty of completion or enjoyment. However, it should be noted that that this is a very short instrument.

In sum, the results of the study are consistent with the findings documented in the research literature on this topic and with the hypotheses. Still, we did not have adequate sample size to detect these differences at a statistical level. Additionally, this test was

done only with very few items and the entire instrument took only a few minutes to complete.

Acknowledgements

We want to thank you Barbara Gandek for the assistance in designing the experiment and Yongwei Yang in providing statistical advice on how to handle Cronbach's alpha standard errors.

We are also grateful to the numerous colleagues who stopped by our poster presentation at the 64th conference of the American Association for Public Opinion Research and provided comments and suggestions.

References

Bell, D. S., Mangione, C. M., & Khan, C. E. (2001). Randomized testing of alternative survey formats using anonymous volunteers on the world wide web. *Journal of the American Medical Informatics Association, 8*, 616-620.

Callegaro, M., & DiSogra, C. (2008). Computing response metrics for online panels. *Public Opinion Quarterly, 72*, 1008-1032.

Chesnut, J. (2008). Effects of using a grid versus a sequential form of the ACS basic demographic data. Retrieved from http://www.census.gov/acs/www/Downloads/ACS-MP-09_Grid-Sequential_Test_Final_Report.pdf

Couper, M. P., Trougott, M. W., & Lamias, M. J. (2001). Web survey design and administration. *Public Opinion Quarterly, 65*, 230-253.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail and mixed-mode surveys. The tailored design method* (Third ed.). Hoboken: John Wiley & Sons, Inc.

Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology, 89*, 792-808.

Garland, P. (2009). Alternative question designs outperform traditional grids. Retrieved from http://www.surveysampling.com/en/news/ssi-research-finds-alternative-question-designs-outperform-traditional-grids

Iglesias, C. P., Birks, Y. F., & Torgerson, D. J. (2001). Improving the measurement of quality of life in older people: the York SF-12. *QJM, 94*, 695-698.

Peytchev, A. (2006). Web survey design: Effect of layout on measurement error. In C. van Dijkum, J. Blasius & C. Durand (Eds.), *Recent developments and applications in social research methodology. Proceedings of the RC33 Sixth International Conference on Social Science Methodology, Amsterdam 2004 [CD-ROM]*. Opladen & Farmington Hills: Barbara Budrich.

Reips, U.-D. (2002). Context effects in web surveys. In B. Batanic, U.-D. Reips & M. Bosnjak (Eds.), *Online social sciences* (pp. 69-79). Seattle: Hogrefe & Huber.

Roberts, L. D. (2007). Equivalence of electronic and off-line measures. In R. A. Reynolds, R. Woods & J. D. Baker (Eds.), *Handbook of research on electronic surveys and measurements* (pp. 97-103). Hershey: Idea Group Reference.

Schwarz, N., & Hippler, H.-J. (1995). Subsequent questions may influence answers to preceding questions in mail surveys. *Public Opinion Quarterly, 59*, 93-97.

Thorndike, F. P., Calbring, P., Smyth, F. L., Magee, J. C., Gonder-Frederick, L., Ost, L.-G., et al. (2009). Web-based measurement: Effect of completing single or multiple items per webpage. *Computers in Human Behavior, 25*, 393-401.

Toepoel, V., Das, M., & van Soest, A. (2009). Design of web questionnaires: A test for number of items per screen. *Field Methods, 21*, 200-213.

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2004). Spacing, position, and order interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly, 68*, 368-393.

U.S. Department of Health and Human Services. (2006). *Research-based Web design & usability guidelines*. Washington D.C.: U.S. Government printing office.

Ware, J. E., Kosinski, M., Bjorner, J. B., Turner-Bowker, D. M., Gandek, B., & Maruish, M. E. (2007). *User's manual for the SF-36v2® Health Survey* (2nd ed.). Lincoln, RI: QualityMetric Incorporated.

Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care, 30*, 473-483.

Yan, T. (2005). *Gricean effects in self-administered survey. Ph.D. Dissertation*. College Park, MD: University of Maryland.

Appendix: SF-36v2 Vitality and Mental Health Scales

Items

How much of the time during the past 4 weeks did you feel full of life? (*Reversed*)

How much of the time during the past 4 weeks have you been very nervous?

How much of the time during the past 4 weeks have you felt so down in the dumps that nothing could cheer you up?

How much of the time during the past 4 weeks have you felt calm and peaceful? (*Reversed*)

How much of the time during the past 4 weeks did you have a lot of energy? (*Reversed*)

How much of the time during the past 4 weeks have you felt downhearted and depressed?

How much of the time during the past 4 weeks did you feel worn out?

How much of the time during the past 4 weeks have you been happy? (*Reversed*)

How much of the time during the past 4 weeks did you feel tired?

Response options

All of the time
Most of the time
Some of the time
A little of the time
None of the time