Kroh, Martin

**Working Paper**

# An Experimental Evaluation of Popular Well-Being Measures

DIW Discussion Papers, No. 546

**Provided in Cooperation with:**
German Institute for Economic Research (DIW Berlin)

*Suggested Citation:* Kroh, Martin (2006) : An Experimental Evaluation of Popular Well-Being Measures, DIW Discussion Papers, No. 546, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/18439

DIW Berlin

German Institute
for Economic Research

Discussion Papers

546

Martin Kroh

# An Experimental Evaluation of Popular Well-Being Measures

Berlin, January 2006

Opinions expressed in this paper are those of the author and do not necessarily reflect views of the Institute.

IMPRESSUM

DIW Berlin
German Institute for Economic Research
Königin-Luise-Str. 5
14195 Berlin
Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
www.diw.de

ISSN print edition 1433-0210
ISSN electronic edition 1619-4535

# An Experimental Evaluation of Popular Well-Being Measures

By:

*Martin Kroh*

DIW Berlin – German Institute for Economic Research
Socio-Economic Panel Study (SOEP)

*Abstract*

Drawing on data from two multitrait multimethod experiments carried out in the context of the German Socio-Economic Panel Study (SOEP), this paper identifies questionnaire designs that minimize measurement error in reports of subjective well-being. Among the survey instruments most often used to measure well-being, the analysis focuses on three response formats (11-point, 7-point and magnitude satisfaction scales) and three modes of data collection (self-administered paper-and-pencil questionnaires (SAQ), personal paper-and-pencil interviews (PAPI) and computer-assisted personal interviews (CAPI)). Results show that both the choice of a response format and the choice of a mode of data collection make a difference in terms of measurement error: The 11-point satisfaction scale and both CAPI and PAPI improve the quality of subjective well-being data. The paper also reports differences between response formats in terms of their ease of administration and illustrates that the choice of a survey instrument affects conclusions drawn from applied well-being research.

Contact:
Martin Kroh
DIW Berlin
Königin-Luise-Straße 5
D–14195 Berlin
Phone: +49 30 89789 678
Fax : +49 30 89789 109
Email: mkroh@diw.de

## 1. Subjective Well-Being in the Social Sciences

In recent years, there has been growing consensus among sociologists and economists that individuals' welfare cannot be described by their objective social situation alone (e.g., Easterlin 2002). The resulting more nuanced view of well-being reflects two developments: broader social trends placing higher value on the quality of life than on economic success (e.g., Inglehart 1990) and shifts in focus within the social sciences to recognize the limits of revealed preferences, i.e. measures of utility that draw on observable choices and their observable putative causes (for an overview of the debate in economics, see e.g. Frey and Stutzer 2002; for sociology see e.g. Glatzer and Zapf 1984; and for the political sciences, see e.g. Van der Eijk et al. forthcoming). These developments have caused the concept of subjective well-being to find its way from its origins in psychology into many other disciplines. The concept is used in the social sciences to investigate, for instance, the effects of unemployment (Clark and Oswald 1994; Winkelmann and Winkelmann 1998; Di Tella et al. 2001), marital status (Alwin 1987; Amato and Sobolewski 2001), gender (cf. Nolen-Hoeksema and Rusting 1999), race (Hughes and Thomas 1998), neighborhood composition (Fernandez and Kulik 1981) and institutional settings (Frey and Stutzer 2000) on well-being.[1]

Although well-being research has developed into a thriving branch of the social sciences, there remain some differences in opinion on the measurement of subjective well-being.[2] Among the various instruments that have been proposed to measure subjective well-being, satisfaction measures dominate in the applied empirical research (see Section 2 for a brief review). This paper does not attempt to contribute new measures of well-being but to investigate the relative performance of a set of accepted and frequently applied satisfaction measures from survey research. The analysis directly addresses the positivists' critique that subjective reports of well-being are prone to various forms of bias, and makes an effort to identify instruments that minimize such avoidable measurement error.

## 2. Measures of Subjective Well-Being

Among the survey-based instruments to measure subjective well-being,[3] two conflicting categories can be distinguished: single-item measures and multiple-item measures. But also within each category, alternative formats, wording, etc. cause considerable heterogeneity. Consider, for instance, the number of items which compose different multiple-item measures: While Diener (1984) uses a five-item satisfaction-with-life scale, Campbell et al. (1976) propose a well-being scale of eight semantic differentials. Bradburn's and Caplovitz's (1965) positive and negative affect scales draw on ten items, and Larsen uses as many as forty items to construct his affect intensity scale (cf. Larsen et al. 1985).
Within the group of single-item measures, one encounters notable differences in wording. Alternative formulations range from the delighted-terrible scale (Andrew and Withey 1976) to the self-anchoring ladder (Cantril 1965). Probably the most common wording of single-item well-being measures goes back to Gurin and his colleagues (1960). Many recent versions of the measure – here taken from the World Value Surveys – contain wording like the following:

> "All things considered, how satisfied are you with your life as a whole nowadays? Please answer using this card, where *a* means completely dissatisfied and *b* means completely satisfied."

Several studies investigate the quality of multiple and single-item measures of subjective well-being, i.e. scales composed of different numbers of items and drawing on different wordings (George and Bearon 1980; Larsen et al. 1985; Stones and Kozma 1985; Pavot et al. 1991; Pavot and Diener 1993; Lucas et al. 1996).[4] Largely unperturbed by these findings, the single-item measure by Gurin et al. (1960) became the accepted tool for many surveys. This probably has to do with survey expenditures (single-item measures require less questionnaire space than multiple-item measures) and the ease of administering alternative instruments (the measure by Gurin et al. (1960) probably fits better into the set of familiar survey questions than, e.g., the self-anchoring ladder proposed by Cantril (1965)). Not only many cross-sectional national surveys in the social sciences (e.g., the General Social Survey in the US), but also leading comparative data projects in the field (e.g., the World Value Surveys, the International Social Survey Programme, ISSP, the European Social Survey, ESS, and the Comparative Study of Electoral Systems, CSES) and popular longitudinal studies (e.g., the European Community Household Panel, ECHP, the British Household Panel Study, BHPS, and the German Socio-Economic Panel Study, SOEP), use wordings similar to the traditional satisfaction measure proposed by Gurin et al. (1960).

This similarity in wording, however, did not put an end to disputes over how to measure subjective well-being. An inspection of codebooks collected by data repositories like ICPSR, the Central Archive at the University Cologne and Steinmetz Archives reveals that probably most of the aforementioned surveys use their unique combination of response format and mode of data collection when administering the Gurin-like satisfaction question.

For instance, the General Social Surveys use the traditional 3-point scale suggested by Gurin et al. (1960). The Eurobarometer and the Comparative Study of Electoral System survey satisfaction data with a 4-point scale, the European Community Household Panel draws on a 5-point scale, the British Household Panel Study and the International Social Science Program use a 7-point scale.[5] The European Social Survey, the Household, Income and Labour Dynamics in Australia Survey and the German Socio-Economic Panel Study draw on 11-point scales.[6]

Researchers encounter a mixture of modes of data collection – even within single surveys. For instance, in the 2002 ISSP surveys, well-being data were collected through paper-and-pencil personal interviewing (PAPI) in Hungary, self-administered questionnaires (SAQ) in the Netherlands, computer-assisted personal interviewing (CAPI) in Switzerland and a combination of SAQ and computer-assisted telephone interviewing (CATI) in Denmark. Many long-running panels like SOEP and BHPS have switched over the course of time from paper-and-pencil interviews to computer-assisted interviewing. Even within a particular survey at a particular point in time, some respondents prefer, for instance, to use SAQ instead of CAPI or the other way around, and respondents are therefore supplied with both options.

Although it is recognized less often than the number or the wording of items as holding relevance for the quality of well-being data, the seemingly innocuous choice of a response format and a mode of data collection has proven to be of relevance for data quality in many instances (Schuman and Presser 1981; Alwin and Krosnick 1991; Scherpenzeel and Saris 1997; Tourangeau et al. 2000; Presser et al. 2004). Differences in data quality due to response formats and modes of data collection may loom large for comparative research in particular: Differences in the data collection process may interfere with the assumption of the equivalence of measures. The increasing number of comparative studies of subjective well-being (e.g., Inglehart 1990; Veenhoven 1993; Diener and Suh 2000) may benefit from established knowledge as to whether, for instance, data from the Swiss Household Panel, i.e. 11-point satisfaction scales surveyed by telephone interviews, are comparable with data from the Russia Longitudinal Monitoring Survey, i.e. a 4-point satisfaction scale surveyed by face-

to-face interviews. So far, very few studies discuss the consequences of alternative response formats and modes of data collection for measuring subjective well-being (e.g., Saris and Van Meurs 1990).

## 3. Expectations

The variety of survey instruments that are used to measure satisfaction suggests differences in opinion on respondents' ability to accurately report their level of well-being. Although it seems that, in many instances, choices of particular response formats and modes of data collection are the result of external decision-making not directly related to the expected quality of subjective well-being data, these decisions nonetheless speak, whether intentionally or not, to certain disputed aspects of questionnaire design. Based on theories of respondent behavior, this section discusses rivaling hypotheses as to why certain formats or modes of data collection may or may not be adequate to measure subjective well-being.

*Response 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 (for an overview, see Schuman and Presser 1981; Tourangeau 2000). Among these characteristics, the gradation of response options is particularly variable across popular well-being measures.

It is often assumed that more response options permit respondents to convey more information and thus increase the data quality (for a review see Alwin 1997). Preferably, respondents are not at all restricted by closed-ended response formats but are permitted to report all the ups and downs in terms of their experienced subjective well-being in an open-ended format. After all, the underlying concept is truly open-ended: Events in the course of one's life may range from infinite happiness to infinite sadness.[7] The review of well-being measures in the social sciences in the previous section suggests, however, that most surveys draw on traditional closed-ended rating formats. That is, response scales comprise a finite number of response options ranging from a minimum integer *a* (e.g., 0) to a maximum integer of well-being *b* (e.g., 2, 3, 4, 6, 10). The gradation of response formats measuring subjective well-being thus ranges from 3 to 11 closed-ended response categories and up to infinite choice options in open-ended formats. Although the reasoning for the choice of few response categories is often not spelled out explicitly, it probably goes back to the idea that few closed-ended response formats are easier to understand for respondents. Detailed answer categories will increase the cognitive demands of respondents and thus the tendency to shortcut answers by accepting the first response category that fits more or less well (Krosnick 1991). The advantage of detailed response options in terms of data quality may at the same time be its disadvantage as too many of them possibly overtax the motivation of many respondents.

*Modes of Data Collection*

While the choice of a mode of data collection is usually discussed in terms of survey expenditures and the ease of administration (cf. Couper et al. 1998), it may also affect respondents' reports of their well-being. The present paper considers the channel of presenting the questionnaire and the mode of responding to be of particular importance for the differences between the most popular well-being measures in CAPI, PAPI and SAQ.[8]

In the SOEP context as in many other surveys as well, CAPI interviews draw by and large on the auditory presentation of questions and oral responding. Showcards of the response scale

are the only visual image of the questionnaire.[9] PAPI also draws also predominantly on auditory presentation and oral responding but provides respondents more often than in CAPI settings with a visual image of questions: Interviewers read out the question but in the SOEP they are also instructed to provide respondents with a visual image of the questionnaire, i.e. the current question, their previous answers and prospective items. Note that the difference between CAPI and PAPI in terms of the channel of presentation is a gradual one. It cannot be excluded that some CAPI respondents get a visual image of questions and that some PAPI respondents only get a visual or auditory image of the questionnaire. In SAQ, on the other hand, respondents only have the visual image of the questionnaire and their only mode of responding is in a written form.

Since the visual and auditory channels of presenting the questionnaire each provide unique information and require special skills, they are known to affect the respondents' processing of questions (e.g., Krosnick and Alwin 1987). Auditory presentation and oral responding (as in CAPI) may improve the quality of data provided by respondents with literacy problems who would otherwise encounter difficulties with a solely visual image of the questionnaire (as in SAQ) (Tourangeau et al. 2000). However, auditory presentation has its drawbacks, too. First, the auditory image of complex questions may overtax the working memory capacities of respondents, diminishing the quality of their answers. The frequent choice of the last response option provided is a problem that is more relevant to auditory than to visual presentation (Schwarz et al. 1991). Visual images of the question, on the other hand, permit respondents to adapt their pace of reading to the complexity of the question. Moreover, the visual presentation of the questionnaire permits respondents to take into account and to edit previous answers. Consider, for instance, satisfaction items for several domains (income, housing, health, democracy, social contacts). Visual presentation of the questionnaire as in SAQ allows respondents to consider their answers on housing in their answers on social contacts and it permits respondents after having reported their satisfaction with their social contacts to revise their previous response on housing in the light of their answer on social contacts. This backtracking may increase data quality.[10]

## 4. Analysis and Data

The 11-point and 7-point scales are probably used most often to measure subjective well-being in surveys (see Section 2 for a review). Their key difference speaks to the issue of the number of scale points. Both scales are, however, limited in their number of answer categories (closed-ended formats). They do not permit respondents to precisely translate possibly continuous latent answers into survey responses (open-ended format). Magnitude scales that do permit such unrestricted responses are less frequently considered in survey research (Wegener 1982). These measures require that respondents express their level of satisfaction as a ratio of an externally defined anchor.[11] One thereby obtains log-interval data on subjective well-being (Saris 1988).

Of the most common modes of data collection, SAQ, PAPI and CAPI permit testing for differences in data quality between channels of presenting the questionnaire and modes of responding (visual presentation and written response in SAQ only, predominantly auditory presentation and oral response in CAPI and both auditory and visual presentation and oral response in PAPI).

The following sections thus investigates the data quality of the 7-point, 11-point and magnitude satisfaction scales as alternative response formats, and SAQ, PAPI and CAPI as alternative modes of data collection. Survey research utilizes various criteria for the evaluation of survey instruments (cf. Presser et al. 2004). This paper investigates two aspects:

measurement error, i.e. validity and reliability of survey responses, and problems of data administration, i.e. non-response, the elapsed time of interviews, respondents' willingness to provide answers and respondents' comprehension of their task.

*Split-Ballot Multitrait Multimethod Experiments*

For the estimation of measurement error, Saris et al. (2004) suggest a design that combines two of the classic approaches: an experimental design and statistical modeling. The multitrait multimethod (MTMM) approach was suggested first by Campbell and Fiske in 1959 and has since then attracted much attention in survey research (for an overview, see Wothke 1996). The basic idea of the MTMM approach is that by repeatedly observing single traits using different methods, the analyst can identify the amount of measurement error in survey instruments.[12] Figure 1 provides a simplified illustration of how data quality, i.e. validity and reliability, is defined in the MTMM context.

Suppose data are collected on respondents' life satisfaction using a 7- and a 11-point scale. Respondents' observed answers (in bounded boxes) are a function of (a) the 'true score' given the response format and (b) measurement error. The share of variance in the observed data that is attributable to the variance in the underlying 'true score' defines the reliability of the measurement instrument (e.g., Bohrnstedt 1983). Put differently, if one would repeat the same question using the same response format, one would expect exactly the same answers if reliability, $r$, equals 1.

<Figure 1>

The 'true score' of respondents' life satisfaction given a particular response format is a function of (a) the underlying 'life satisfaction factor', i.e. the latent answer, and (b) the method used, i.e. the 7- and the 11-point scales respectively. Validity, $v$, is the importance of the 'life satisfaction factor' in the 'true scores'.[13] For each response format, a unique method variance can be estimated, which is interpreted as a systematic error due to the response format.[14]

The identification of validity and reliability parameters in the classical MTMM approach requires observations on at least three traits which have to be measured with three different methods (Saris and Andrews 1991). In other words, respondents would have to provide answers to the same set of three items (e.g., satisfaction with domains $a$, $b$ and $c$) with some variation in the response format only (e.g., using survey instruments $x$, $y$ and $z$).

The repeated surveying of the same items in the classical MTMM context means not only a burden for respondents but also bears the risk of memory and order effects. The combination of the MTMM approach with a split-ballot design reduces the number of necessary repetitions. The advantage of randomly splitting the sample into groups which are presented with different formats of the questionnaire is that variation in response patterns between experimental groups is attributable to systematic differences between measurement instruments and random variation only (cf. Schumann and Presser 1981). Since each of the randomly drawn groups receives a different combination of two response formats, one requires only one instead of two repetitions of traits. For instance, a first group reports their level of satisfaction in domains $a$, $b$, and $c$ using the method $x$ at the beginning of the interview and does the same using method $y$ at the end of the interview. A second group may use method $y$ at the beginning of the interview and does the same using method $z$ at the end of the interview, etc. Even though not all combinations of traits and methods are observed for all respondents, validity and reliability parameters can nonetheless be identified by normal theory maximum likelihood in multiple groups assuming a common model, i.e. with equality

constraints of all parameters across random groups (for a validation of the estimation technique, see Saris et al. 2004).
Drawing on the split-ballot MTMM design reduces problems of repeated observations as compared to the classical MTMM approach. Memory effects are less likely to occur since a considerable time elapses between both observations of the same traits. Moreover, the design makes it possible to control for order effects by placing each method once at the beginning and once at the end of the interview.

*Experiment 1: Testing Alternative Response Formats in a Methodological Pretest of SOEP*
In the methodological pretest to *SOEP*, respondents were asked to report their satisfaction with several domains, among others with their lives in general, health, and household income. The total sample was divided into two random groups with variation in the response formats measuring subjective well-being as illustrated in Table 1. The first group used an 11-point scale at the beginning and a magnitude scale at the end of the interview, and the second group used 7-point well-being scales at the beginning and magnitude scales at the end of the interview. Traits were repeated on average 55 minutes after the first round of satisfaction items. None of the interviews had a time gap of less than 20 minutes between the two observations. Van Meurs and Saris (1990) show that 20 minutes are sufficient to obtain independent measures. Note that all interviews of the SOEP pretest were collected by means of computer-assisted personal interviewing (CAPI).

<Table 1>

The design of the pretest permits the estimation of validity and reliability parameters for alternative response formats measuring satisfaction. However, in order to fully investigate the performance of different survey instruments, the SOEP pretest provides four additional indicators of problems during the administration of interviews. A first indicator is the refusal to give well-being answers. Non-response is defined here as the refusal of the interviewed persons to provide answers on the satisfaction domains. This occurs in 2% of all cases.
The time necessary to conduct all satisfaction items operates as a second indicator for the ease of administration. On average, administrating all items takes 67 seconds.[15] The elapsed time between the introduction to the well-being items and the last satisfaction domain is not normally distributed. Analyzing the logarithm of the elapsed time in seconds instead of the raw data accounts for the skewed distribution.
Immediately after the administration of the satisfaction items, interviewers are asked to grade respondents' participation using a six-point school grading system. The third indicator of the ease of administration is interviewers' grade of respondents' willingness to provide answers on their level of well-being and the fourth indicator is interviewers' perception of respondents' comprehension of their task.

*Quasi-Experiment 2: Testing Alternative Modes of Data Collection in the SOEP*
Testing the quality of satisfaction data across alternative modes of data collection does not make it necessary to collect new experimental data as in the case of alternative response formats. For this purpose, readily available panel data can be used. Many surveys use a mix of different modes of data collection. If, for instance, respondents feel uncomfortable using a computer during the interview, interviewers may switch from computer-assisted interviewing to paper-and-pencil interviewing instead; or if respondents prefer an visual image of the questionnaire, they may read and fill in the questionnaire on their own. As a consequence of respondents' and interviewers' preferences, one obtains variation in the mode of data collection across several waves of a panel. The present paper draws on satisfaction data from the two consecutive waves in 2002 and 2003 of the Socio-Economic Panel (SOEP). Note that

SOEP always draws on the 11-point scale to survey subjective well-being. Table 3 illustrates the design of this (nonrandom) three-group split-ballot experiment: A first group used self-administration (SAQ) in wave 2002 and personal paper-and-pencil interviewing (PAPI) in 2003. Respondents from the second group switched between SAQ and computer-assisted personal interviewing (CAPI), the third group switched from PAPI to SAQ, etc. Hence, all combinations of change in methods are observed between 2002 and 2003, i.e. controlling for order effects.

<Table 2>

In contrast to the split-ballot experiment conducted in the SOEP pretest, findings of the quasi-experiment on basis of regular SOEP data may be plagued by two caveats. First, the process of assigning respondents to certain combinations of modes of data collection in the two waves of the panel may not be random. There is reason to believe that respondents and interviewers select themselves into certain modes of data collection. If characteristics of the self-selection process are related to respondents' ability to provide unbiased answers, the estimation of data quality across modes of data collection may produce misleading results. The second problem is one of true change between the two observations. The time lack between observations in the quasi-experiment is not one hour, as in the SOEP pretest, but a whole year. While it is plausible to assume that satisfaction with one's life, health or income does not change within an hour, this may very well be the case in one year's time. As a consequence, estimates of the quality of satisfaction data based on yearly observations may be downwardly biased due to the unspecified true change in satisfaction.

However, already Brickman and Campbell (1971) argue that individuals have an equilibrium level of life satisfaction. Positive or negative events lead to temporal changes only as individuals quickly adapt to their prior equilibrium. In that sense, estimating the data quality of subjective well-being measures on basis of yearly panel data may not interfere with true change in individuals objective situation as much as one would expect at first glance. Moreover, the biasing effects of true change on validity and reliability estimates would presumably affect estimates for all three modes of data collection uniformly. In other words, even if the *absolute* magnitude of reliability and validity estimates may be misspecified in the quasi-experiment, there is no obvious reason to believe that this affects the *relative* magnitude of validity and reliability estimates across modes of data collection. The latter is what this paper is primarily interested in.

To account for possible selectivity into certain modes of data collection, this paper employs a weighting strategy (e.g., Wooldridge 2002). That is, the estimation of reliability and validity across modes of data collection was performed once on the raw data and once weighted by the inverse probability of entering one of the groups describes in Table 2. The weights draw on interviewer characteristics (gender, age, education, experience with SOEP), respondent characteristics (gender, age, education, income, nationality, residence in East or West Germany, number of children under the age of six, experience with SOEP in years) and temporal changes in interviewers and respondents (change in interviewer, change in respondents' income and health). Moreover, the weights are multiplied by the cross-sectional and longitudinal weights provided by the SOEP team and thus control for panel attrition and selectivity of the initial sampling procedure (Kroh and Spiess 2005). However, both the weighted and the unweighted analysis of the data quality across modes of data collection lead to substantively the same conclusions. For lack of space, the derivation of weights is not reported in form of a table nor is the weighted estimation of reliability and validity. These information can be obtained from the author on request.

## 5. Findings

The primary criterion by which survey instruments are evaluated is their ability to measure respondents' views without random or even systematic error. This section reports the validity and reliability estimated by the two MTMM experiments described above. Moreover, the section provides a concrete example illustrating how, depending on the choice of a response format, alternative survey instruments affect substantive interpretations about the nature of subjective well-being. A secondary criterion for the evaluation of survey instruments are problems of survey administration. A subsequent empirical section reports the performance of response formats in terms of non-response, elapsed time of administration, respondents' motivation and comprehension.[16]

*Experiment 1: Validity and Reliability of Alternative Response Formats*

Given the split ballot design of the methodological pretest to SOEP (see Table 1), one obtains correlations between three traits (respondents' satisfaction with life, health and income) measured with three alternative instruments (11-point scale, 7-point scale and magnitude scale). Since the scales investigated differ in terms of the number of response categories, polychoric correlations are estimated. Whereas ordinary correlations assume continuous data, polychoric correlations are suited for data with different levels of measurement (Olsson 1979).

<Table 3>

Standardized parameters of Table 4 vary between 0 and 1. The squared validity and method parameters denote the share of variance in the true scores attributable to the well-being factor and the method factor respectively (see also Figure 1). For instance, 21% ($=.46^2$) of the variance in the true life satisfaction scores on the 11-point scale is due to the particular scale used and 79% ($= .89^2$) is due to the latent well-being factor. Similarly, the squared reliability parameters indicate the share of variance in the observed satisfaction data attributable to their true scores. For instance, 72% ($= .85^2$) of the variance in the observed life satisfaction data on the 11-point scale reflects true score variation; 28% of the variance reflects the unreliability of the measure.

Validity estimates of the 11-point scale hover around .89 for the 11-point scale, .80 for the 7-point scale and .70 for the magnitude scale indicating that the 11-point scale provides the highest levels of data quality in terms of validity. The low validity of the magnitude scale in particular suggests a biasing response behavior that leads to high correlations between items administered using this method. Respondents' affinity for exposed scores may explain this method effect (for problems of rounding, see e.g., Tourangeau et al. 2000). Although the open-ended magnitude scale permits respondents to select any number between zero and infinity, 38% of respondents who report their satisfaction with life in general using a magnitude scale select values which are multipliers of 50 (50, 100, 150, etc.) and 83% choose scores which are multipliers of 10 (10, 20, 30, etc.).

In terms of reliability, however, the magnitude scale performs better than the 11- and the 7-point scale, which fits with previous findings (Scherpenzeel and Saris 1997). On average, reliability equals .94 for the magnitude scale across traits, .83 for the 11-point scale and .79 for the 7-point scale across traits.

<Table 4>

A $\chi^2$ difference test provides a statistic of the significance of differences in parameter estimates across scales. The idea of the test is that in case of equal size of reliability and

validity estimates, a model that sets parameters to be equal across scales should perform about equally well to a model that allows these parameters to vary across scales. Table 4 reports the model fit in terms of $\chi^2$ values and degrees of freedom of such nested models relative to the basic model reported in Table 3. Note that the difference between $\chi^2$ test statistics of nested models is asymptotically independent of the test statistics themselves and that it is also $\chi^2$ distributed (Steiger et al. 1985). For instance, a model that constraints reliability estimates to be the same across all scales provides a significantly poorer fit to the data than the model reported in Table 3 that allows these parameters to vary ($\chi^2$ difference of 21.75 and a difference in degrees of freedom of 5). Hence, one would reject the null hypothesis of equality between estimates across scales. This also holds if one constraints reliability to be the same for only two out of three scales. Moreover all equality constraints of validity estimates across scales are rejected by $\chi^2$ difference tests.
In sum, the first MTMM-Experiment and the formal tests thereof suggest that the 11-point satisfaction scale ranks first in terms of validity and second in terms of reliability. The magnitude scale produces the highest reliability but the lowest validity while the 7-point satisfaction scale is the least reliable one of the response formats tested and ranks second in terms of validity.

*Quasi-Experiment 2: Validity and Reliability of Alternative Modes of Data Collection*
The second MTMM experiment on the basis of regular SOEP data indicates that differences in validity and reliability are less pronounced between modes of data collection. The average validity estimate for compute-assisted interviewing (CAPI) is .95, for the personal paper-and-pencil interviewing (PAPI) .93 and for the self-administered questionnaires .86. Reliability estimates hover around .84 for CAPI, .82 for PAPI and .80 for SAQ.

<Table 5>

Note the similarity between the results of the pretest on the response format 11-point scale (given CAPI as mode of data collection) and the results of the regular SOEP data on CAPI as the mode of data collection (given the 11-point scale as response format) in terms of reliability (.83 in the first and .84 in the second case). However, estimated validity is with .95 surprisingly higher in the quasi-experiment that draws on data repeated with a one-year gap than in the second experiment that draws on reports of subjective well-being repeated with an one-hour gap with .89. This difference may be explained by the level of survey experience of respondents in the SOEP pretest and the SOEP main survey: Respondents of the pretest most probably reported their levels of well-being for the first time in a survey while respondents of the long running SOEP main survey in 2002 reported their levels of subjective well-being on average for the seventh time.

<Table 6>

The $\chi^2$ difference test statistics reported in Table 6 suggest that all reliability and validity estimates are significantly different from each other across modes of data collection with the exception of the pairwise comparison between CAPI and PAPI both in terms of reliability and validity and the pairwise comparison of SAQ and PAPI in terms of reliability. Models that set these parameters to be equal across CAPI and PAPI, respectively SAQ and PAPI in terms of reliability, produce about the same fit to the data than the more general model reported in Table 5.
In sum, the second MTMM-Experiment and the formal tests thereof suggest that well-being data collected in auditory interview settings (CAPI and PAPI) produce less measurement error than the visual presentation of well-being questions (SAQ).

*Consequences for Applied Research*

Previous analyses show that alternative methods surveying subjective well-being perform differently well in terms of data quality. As a practitioner in the social sciences, one may nonetheless ask whether the choice of a certain survey instrument makes a notable difference for applied research. To better understand the costs of different instruments, consider the relationship between objective income and subjective satisfaction with income. Already early studies of subjective well-being reported a surprisingly low explanatory power of objective data for subjective well-being (Easterlin 1974; Andrews and Whitey 1976; Campbell et al. 1976). This finding – repeatedly confirmed by later studies – became part of the established knowledge of subjective well-being research and sparked quite a number of new theoretical approaches (cf. Argyle 1999; Van Praag and Frijters 1999; Easterlin 2001). An explanation that received less attention is the straightforward argument that the low correlation between objective income and subjective satisfaction (with income) is the consequence of inadequate measures of both income and satisfaction (Saris 2001).

Consider, for instance, the estimation of the effect of log equivalent household income on satisfaction with household income in empirical data collected in the SOEP pretest. An OLS regression[17] of satisfaction with income reported in Table 8 shows that objective income affects respondents' satisfaction with income, however, with significant variation across response formats. Note that all satisfaction data are rescaled to the same length to permit comparisons of parameter estimates.

<Table 7>

As one would expect, income satisfaction increases with log-income, however, as indicated by the interaction terms with significantly different magnitude across response formats. The estimated effect of respondents' income on satisfaction with income is $b = .15$ for the 7-point scale, $b = .15 + .10 = .25$ for the 11-point scale and $b = .15 - .05 = .10$ for the magnitude scale.

<Figure 2>

Figure 2 illustrates these differences in the effect magnitude between different response formats for the predicted satisfaction with income and objective income. The curve for the 11-point scale is much steeper than the curve for the 7-point and magnitude scales. Different response formats produce notably different predictions about the nature of respondents' levels of satisfaction. One can of course not judge from Figure 2 which scale reveals the 'true' relationship between objective income and satisfaction with income. However, the higher validity of the 11-point scale as compared to the 7-point and the magnitude scales documented in Table 3 may be interpreted as indicative of the presumption that the 11-point scale produces the more valid picture.

*The Ease of Administering Different Well-Being Scales*

While it is difficult to study the problems of survey administration across different modes of data collection basically because external information on, for instance, the duration of interviews, misunderstandings and the motivation of respondents are unavailable for self-administered questionnaires, the methodological pretest of SOEP was designed to provide such information across different response formats. Four indicators of the ease of data administration enable us to better evaluate the performance of alternative response formats in terms of non-response, elapsed time of surveying subjective well-being items, respondents' motivation to provide answers and their comprehension of the survey task. Information on the

latter two points are obtained by way of behavioral coding, i.e. interviewers report their perception of respondents' motivation and comprehension when using subjective well-being measures.
Since respondents received the well-being items repeatedly during the interview, each of these indicators is available twice for each respondent. Moreover, each interviewer conducted several personal interviews, on average 6. This leads to a hierarchical data structure across three levels in which, for instance, the time necessary to complete the subjective well-being questions may depend on characteristics of interviewers, characteristics of respondents and characteristics of the particular setting in the beginning and the end of each personal interview. This hierarchical data structure necessitates error terms for each level of the data that can be achieved by multilevel modeling (e.g., Snijders and Bosker 1999).

The hierarchical regression models reported in Table 8 test the effect of the response format on indicators of problems of administration. In case of non-response, a binary probit model is used; in case of the log-transformed elapsed time in seconds, a least squares regression is used; and in case of grading by interviewers, as is the case in Models 3 and 4, ordinal probit regression is used. Previous analyses show that respondents' cognitive skills and memory capacities may moderate the performance of different survey instruments (Tourangeau et al. 2000, Chapter 3). To control for such intervening factors, the regression models in Table 8 stratify the effect of response formats according to respondents' educational level and age. The inclusion of age as an intervening factor rests on the assumption that the poorer memory of older people may affect the administration of interviews. To control for the location of satisfaction items in the questionnaire an additional binary variable is included, distinguishing between responses in the beginning and the end of the interview.

<Table 8>

Inspecting the effect of the questionnaire design, administering satisfaction items at the end of the interview takes less time but increases respondents' reluctance to provide answers. Response formats affect the elapsed time of interviews, respondents' reluctance and misunderstandings but not non-response. The open-ended magnitude scale is associated with more reluctance and misunderstandings and takes more time to administer. Differences within the two closed-ended formats occur in terms of reluctance only with the 11-point scale being associated with lower levels of motivation.
As one would expect, age positively affects the elapsed time of the interview, respondents' reluctance, and misunderstandings; while education decreases non-response and misunderstandings. These effects of respondents' characteristics are associated with the response format used, i.e. memory effects seem more or less pronounced for certain response formats.[18] However, there seems to be no clear pattern linking one or the other format to differences in cognitive skills.

## 6. Conclusions

The present paper aims at identifying survey instruments that maximize the quality of well-being data. Among the high number of alternative measures, the analysis focuses on the most common single-item measures in applied survey research. Two multitrait multimethod experiments test for differences in terms of validity and reliability between three response formats (7-point scale, 11-point scale and magnitude scale) and three modes of data collection (SAQ, PAPI and CAPI). Moreover, an analysis of the ease of administration tests for

differences between response formats (non-response, time of interviewing, respondents' reluctance and misunderstandings).
In terms of validity, the 11-point scale and auditory presentation of the questionnaire (PAPI and CAPI) provide the highest data quality. In terms of reliability, the open-ended magnitude scale and auditory presentation perform better than the closed-ended scales and SAQ. In terms of the ease of administrating the three response formats, the magnitude scale produces somewhat more problems than the closed-ended 7-point and 11-point scales. On balance, the combination of the 11-point satisfaction scale and CAPI or PAPI comparatively high levels of data quality.[19]

What explains these differences? Section 3 formulates rivaling hypotheses as to why certain survey instruments measuring subjective well-being may outperform others in terms of their derived data quality. With regard to response formats, one may expect differences in data quality across satisfaction scales with different gradation. Supportive of the hypothesis that respondents' latent satisfaction reports are elaborate in nature, the more detailed 11-point response scale produces somewhat better data than the shorter 7-point satisfaction scale.

However, in line with the expectation that closed-ended formats (7- and 11-point scales) as compared to open-ended formats (magnitude scale) are easier to understand and consequently to apply by respondents in mapping their latent responses, validity is higher for the closed-ended formats. The magnitude scale applied here, leading to a method effect in well-being data, seems to be limited in accurately measuring the infinite degrees of satisfaction. The problems of the open-ended format are also reflected in indicators of the ease of administration: interviews with magnitude scales take more time to conduct and respondents seem less motivated to use the format.

Section 3 also formulates expectations with regard to the data quality of alternative modes of data collection measuring satisfaction. The argument in favor of self-administration is that this mode of data collection is more flexible to respondents, enabling them to change their minds and revise previous answers (e.g., satisfaction with housing, job) in the light of answers made later on (e.g., satisfaction with standard of living, income). The higher quality of answers obtained by auditory presented questionnaires, however, suggests that the tangibility of an oral interview is more important to respondents' reports of their well-being than the flexibility of a written interview.

These differences in data quality – explicable by certain properties of survey instruments and associated response behaviors as discussed before – have implications for substantive research on well-being. As illustrated by means of the relationship between income and satisfaction with income, different instruments do not produce equivalent measures. That is, alternative survey instruments interfere with interpretations of the nature of subjective well-being. Comparative research on well-being in particular (e.g., Inglehart 1990; Veenhoven 1993; Diener and Suh 2000; Frey and Stutzer 2000), which often draws on existing survey data, may therefore control for different survey instruments when analyzing pooled well-being data. Future research would greatly benefit from general consensus on a single instrument for surveying subjective well-being. The analysis presented in this paper suggests use of the 11-point scale and auditory presentation of the questionnaire, a combination which ensures the highest data quality of single-item measures.

## 7. References


Agryle, Michael. 1999. Causes and Correlates of Happiness. In *Well-Being: The Foundations of Hedonic Psychology*. Daniel Kahnemann, Ed Diener and Norbert Schwarz (eds.). New York: Russell Sage Foundation: 353-373.

Alwin, Duane F. 1987. Distributive Justice and Satisfaction with Material Well-Being. *American Sociological Review* 52: 83-95.

Alwin, Duane F. and Krosnick, Jon A. 1991. The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes. *Sociological Methods and Research* 20: 139-181.

Alwin, Duane F. 1997. Feeling Thermometers Versus 7-Point Scales. Which are Better? *Sociological Methods and Research* 25: 318-340.

Amato, Paul R. and Sobolewski, Juliana M. 2001. The Effects of Divorce and Marital Discord on Adult Children's Psychological Well-Being. *American Sociological Review* 66: 900-921.

Andrews, Frank M. and Withley, Stephen B. 1976. *Social Indicators of Well-Being*. New York: Plenum.

Bohrnstedt, George W. 1983. Measurement. In *Handbook of Survey Research*. Peter H. Rossi, James D. Wright and Andy B. Anderson (eds.). New York: Academic Press: 70-121.

Bradburn, Norman M. 1969. *The Structure of Psychological Well-Being*. Chicago: Aldine.

Bradburn, Norman M. and Caplovitz, David. 1965. *Reports of Happiness*. Chicago: Aldine.

Brickman, Philip and Campbell, Donald T. 1971. Hedonic Relativism and Planning the Good Society. In *Adaptation Level Theory: A Symposium*. Mortimer H. Appley (ed.). New York: Academic Press, 287-301.

Browne, Michael W. 1984. The Decomposition of Multitrait-Multimethod Matrices. *British Journal of Mathematical and Statistical Psychology* 37: 1-21.

Campbell, Angus, Converse Philip E. and Rodgers, Willard L. 1976. *The Quality of American Life*. New York: Sage.

Campbell, Donald T. and Fiske, Donald W. 1959. Convergent and Discriminant Validation by the Multitrait Multimethod Matrices. *Psychological Bulletin* 56: 81-105.

Cantril, Hadley. 1965. *The Patterns of Human Concern*. New Brunswick: Rutgers University Press.

Clark, Andrew E. and Oswald, Andrew J. 1994. Unhappiness and Unemployment. *Economic Journal* 104: 648-659.

Couper, Mick P., Baker, Reginald P., Bethlehem, Jelke , Clark, Cynthia Z.F., Martin, Jean Nicholls, William L. II, O'Reilly, James M. (eds). 1998. *Computer Assisted Survey Information Collection*. New York: John Wiley.

Di Tella, Rafael, MacCulloch, Robert J. and Oswald, Andrew J. 2001. Preferences over Inflation and Unemployment: Evidence from Surveys of Happiness. *American Economic Review* 91: 335-341.

Diener, Ed. 1984. Subjective Well-Being. *Psychological Bulletin* 95: 542-575.

Diener, Ed, Suh, Eunkook M., Lucas, Richard E. and Smith, Heidi L. 1999. Subjective Well-Being: Three Decades of Success. *Psychological Bulletin* 125: 276-302.

Diener, Ed, and Oishi, Shigehiro. 2000. Money and happiness: Income and subjective well-being across nations. In *Subjective well-being across cultures*. Ed Diener and Eunkook M. Suh (eds.). Cambridge: MIT Press: 185-218.

Diener, Ed and Suh, Eunkook M. (eds.). 2000. *Culture and Subjective Well-Being*. Cambridge: MIT Press.

Easterlin, Richard A. 1974. Does Economic Growth Improve the Human Lot? Some Empirical Evidence. In *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz*. Paul A. David and Melvin W. Reder (eds.). New York: Academic Press: 89-125.
Easterlin, Richard A. 2001. Income and Happiness: Towards a Unified Theory. *Economic Journal* 111: 465-484.
Easterlin, Richard A. (ed.). 2002. *Happiness in Economics*. Edward Elgar: Cheltenham.
Fernandez, Roberto M. and Kulik, Jane C. 1981. A Multilevel Model of Life Satisfaction: Effects of Individual Characteristics and Neighborhood Composition. *American Sociological Review* 46: 840-850.
Frey, Bruno S. and Stutzer, Alois. 2000. Happiness, Economy and Institutions. *Economic Journal* 11: 456-484.
Frey, Bruno S. and Stutzer, Alois. 2002. What Can Economists Learn from Happiness Research? *Journal of Economic Literature* 40: 402-435.
George, Linda K. and Bearon, Lucille B. 1980. *The Quality of Life in Older Persons: Meaning and Measurement*. New York: Human Sciences Press.
Glatzer, Wolfgang. 1984. Lebenszufriedenheit und alternative Maße subjektiven Wohlbefindens. In *Lebensqualität in der Bundesrepublik: objektive Lebensbedingungen und subjektives Wohlbefinden*. Wolfgang Glatzer and Wolfgang Zapf (eds.). Frankfurt: Campus: 177-191.
Glatzer, Wolfgang and Zapf, Wolfgang (eds.). 1984. *Lebensqualität in der Bundesrepublik: objektive Lebensbedingungen und subjektives Wohlbefinden*. Frankfurt: Campus.
Gurin, Gerald, Veroff, Joseph and Feld, Sheila. 1960. *Americans View of Their Mental Health*. New York: Basic.
Hall, John, Lord, David, Marsh, Cathie and Ring, James. 1973. *Quality of Life Survey (Urban Britain: 1973)*. London: SSRC Users Manual.
Hughes, Michael and Thomas, Melvin E. 1998. The Continuing Significance of Race Revisited: A Study of Race, Class, and Quality of Life in America, 1972 to 1996. *American Sociological Review* 63: 785-795.
Inglehart, Ronald. 1990. *Cultural Shift in Advanced Industrial Society*. Princeton: Princeton University Press.
Kahnemann, Daniel, Diener, Ed. and Schwarz, Norbert (eds.). 1999. *Well-Being: The Foundations of Hedonic Psychology*. New York: Russell Sage Foundation.
Kenny, Charles. 1999. Does Growth Cause Happiness, or Does Happiness Cause Growth? *Kyklos* 52: 3-26.
Kroh, Martin. 2005. *Surveying the Left-Right Dimension: The Choice of a Response Format*. DIW Discussion Paper 491.
Kroh, Martin and Spieß, Martin. 2005. *Documentation of Sample Sizes and Panel Attrition in the German Socio Economic Panel (SOEP) 1984–2004*. DIW Data Documentation 6.
Krosnick, Jon A. 1991. Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology* 5: 213-236.
Krosnick, Jon A. and Alwin, Duane F. 1987. An Evaluation of Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opinion Quarterly* 51: 201-219.
Larsen, Randy J., Larsen, Ed and Emmons, Robert A. 1985. An Evaluation of Subjective Well-Being Measures. *Social Indicators Research* 17: 1-17.
Larsen, Randy J. and Fredrickson, Barabara L. 1999. Measurement Issues in Emotion Research. In *Well-Being: The Foundations of Hedonic Psychology*. Daniel Kahnemann, Ed Diener and Norbert Schwarz (eds.). New York: Russell Sage Foundation: 40-60.

Lucas, Richard E., Diener, Ed and Suh, Eunkook M. 1996. Discriminant Validity of Well-Being Measures. *Journal of Personality and Social Psychology* 71: 616-628.
Marsh Herbert W. 1989. Confirmatory Factor Analysis of Multitrait-Multimethod Data: Many Problems and Few Solutions. *Applied Psychological Measurement* 13: 335-361.
Nolen-Hoeksema, Susan and Rusting, Cheryl L. 1999. Gender Differences in Well-Being. In *Well-Being: The Foundations of Hedonic Psychology*. Daniel Kahnemann, Ed Diener and Norbert Schwarz (eds.). New York: Russell Sage Foundation: 330-350.
Olsson, Ulf. 1979. Maximum Likelihood Estimation of the Polychoric Correlation Coefficient. *Psychometrika* 44: 443-460.
Paulhus, Delroy L. 1991: Measurement and control of response bias. In *Measures of Personality and Social Psychological Attitudes*. John P. Robinson, Phillip R. Shaver and Lawrence S. Wrightman (eds.). San Diego: Academic Press: 17-59.
Pavot, William, Diener, Ed, Colvin, C. Randall and Sandvik, Ed. 1991. Further Validation of the Satisfaction With Life Scale: Evidence for the Cross-Method Convergence of Well-Being Measures. *Journal of Personality Assessment* 57: 149-161.
Pavot, William, and Diener, Ed. 1993. Review of the Satisfaction with Life Scale. *Psychological Assessment* 5: 164-172.
Presser, Stanley, Rothgeb, Jennifer M., Couper, Mick P., Lessler, Judith T., Martin, Elizabeth, Martin, Jean and Singer, Eleanor. (eds.). 2004. *Methods for Testing and Evaluating Survey Questions*. New York: Wiley.
Saris, Willem E. 1988. A measurement model for psychophysical scaling. *Quality and Quantity* 22: 417-433.
Saris, Willem E. 2001. The Strength of the Causal Relationship Between Living Conditions and Satisfaction. *Sociological Methods and Research* 30: 11-34.
Saris, Willem E. and Andrews, Frank M. 1991. Evaluation of Measurement Instruments Using a Structural Modeling Approach. In *Measurement Errors in Surveys*. Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz and Seymour Sudman 8eds.). New York: Wiley: 575-599.
Saris, Willem E., Satorra, Albert and Coenders, Germà. 2004. A New Approach to Evaluating the Quality of Measurement Instruments: The Split-ballot MTMM Design. *Sociological Methodology* 34: 311-347.
Saris, Willem E. and Van Meurs, E. (eds.). 1990. *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait-Multimethod Studies*. Amsterdam: North Holland.
Scherpenzeel, Annette and Saris, Willem E. 1997. The Validity and Reliability of Survey Questions: A Meta-Analysis of MTMM studies. *Sociological Methods and Research* 25: 341-383.
Schuman, Howard and Presser, Stanley. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Order and Context*. New York: Academic Press.
Schwarz, Norbert, Strack, Fritz and Mai, Hans-Peter. 1991. Assimilation and Contrast Effects in Part-Whole Question Sequences: A Conversational Logic Analysis. *Public Opinion Quarterly* 55: 3-23.
Siara, Christian S. 1980. *Komponenten der Wohlfahrt: Materialien zu Lebensbedingungen und Lebensqualität in der Bundesrepublik Deutschland*. Frankfurt: Campus.
Snijders, Tom and Bosker, Roel. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage Publications.
Steiger, James H., Shapiro, Alexander and Browne, Micheal W. 1985. On Multivariate Asymptotic Distribution of Sequential Chi-square Statistics. *Psychometrika* 50: 253-264.
Stones, Michael J. and Kozma, Albert. 1985. Structural Relationships among Happiness Scales: A Second Order Factorial Study. *Social Indicators Research* 17: 19-28.

Tourangeau, Roger, Rasinski, Kenneth A. and Bradburn, Norman. 1991. Measuring Happiness in Surveys: A Test of the Subtraction Hypothesis. *Public Opinion Quarterly* 55: 255-266.

Tourangeau, Roger, Rasinski, Kenneth A., Jobe, J.B., Smith, T.W. and Pratt, W. 1997. Sources of Error in a Survey of Sexual Behavior. *Journal of Official Statistics* 13: 341-365.

Tourangeau, Roger, Rips, Lance J. and Rasinski, Kenneth. 2000. *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Van der Eijk, Cees, Van der Brug, Wouter, Kroh, Martin and Franklin, Mark N. forthcoming. Rethinking the Dependent Variable in Electoral Behavior — On the Measurement and Analysis of Utilities. *Electoral Studies*.

Van Meurs, A., and Willem E. Saris. 1990. Memory Effects in MTMM Studies. In *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait-Multimethod Studies.* Willem E. Saris and A. van Meurs (eds.). Amsterdam: North Holland, 134-146.

Van Praag, Bernard M. and Frijters, Paul. 1999. The Measurement of Welfare and Well-Being: The Leyden Approach. In *Well-Being: The Foundations of Hedonic Psychology*. Daniel Kahnemann, Ed Diener and Norbert Schwarz (eds.). New York: Russell Sage Foundation: 413-433.

Veenhoven, Ruut. 1993. *Happiness in Nations: Subjective Appreciation of Life in 56 Nations 1946-1992*. Rotterdem: Erasmus University Press.

Wegener, Bernd (ed.). 1982. *Social Attitudes and Psychophysical Measurement*. Hillsdale: Erlbaum.

Winkelmann, Liliana and Winkelmann, Rainer. 1998. Why are the Unemployed so Unhappy? Evidence from Panel Data. *Economica* 65: 1-15.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.

Wothke, Werner. 1996. Models for Multitrait-Multimethod Matrix Analysis. In *Advanced Structural Equation Modeling: Issues and Techniques*. George C. Marcoulides and Randall E. Schumacker (eds.). Mahwah: Lawrence Erlbaum: 7-56.

**Table 1** Design of the MTMM Experiment (SOEP Pretest 2004)

| | Beginning of Interview | End of Interview | n |
|---|---|---|---|
| Group 1 | 11 Point Scale | Magnitude Scale | 248 |
| Group 2 | 7 Point Scale | Magnitude Scale | 251 |

**Table 2** Design of the 6-Group Split-Ballot MTMM Experiment (SOEP 2002, 2003).

| | Wave 2002 | Wave 2003 | n |
|---|---|---|---|
| Group 1 | SAQ | PAPI | 282 |
| Group 2 | SAQ | CAPI | 284 |
| Group 3 | PAPI | SAQ | 513 |
| Group 4 | PAPI | CAPI | 511 |
| Group 5 | CAPI | SAQ | 338 |
| Group 6 | CAPI | PAPI | 321 |

**Table 3** Data Quality of Alternative Response Formats Measuring Satisfaction.

| | | **Validity** | | | **Method Effect** | | | **Reliability** |
|---|---|---|---|---|---|---|---|---|
| | | Life | Health | Income | 11-Point Scale | 7-Point Scale | Magnitude Scale | |
| 11-Point Scale | Life | 0.89 | | | 0.46 | | | 0.85 |
| | Health | | 0.88 | | 0.47 | | | 0.81 |
| | Income | | | 0.89 | 0.46 | | | 0.84 |
| 7-Point Scale | Life | 0.82 | | | | 0.58 | | 0.83 |
| | Health | | 0.82 | | | 0.58 | | 0.84 |
| | Income | | | 0.75 | | 0.67 | | 0.69 |
| Magnitude Scale | Life | 0.65 | | | | | 0.76 | 0.91 |
| | Health | | 0.71 | | | | 0.70 | *0.97* |
| | Income | | | 0.73 | | | 0.68 | 0.95 |

*Note*. All estimates significant at $p<.05$; $N=496$; $\chi^2 = 94.52$ with 67 df. *Data Source*. SOEP Pretest 2004.

**Table 4** Chi-Square Difference Tests Between Nested MTMM-Models (with Equality Constraints) and the Basic Model (without Equality Constraints).

| | Model Fit | | Difference in Model Fit | | |
|---|---|---|---|---|---|
| | $\chi^2$ | df | $\Delta\chi^2$ | $\Delta$df | p($\Delta$) |
| **Basic Model**[†] | 94.52 | 67 | - | - | - |
| | | | | | |
| **Equality-Constraints on Reliability Estimates** | | | | | |
| Between All Scales | 116.27 | 72 | 21.75 | 5 | *** |
| Between 11 and 7-Point Scale | 104.64 | 70 | 10.12 | 3 | ** |
| Between 11-Point and Magnitude Scale | 102.82 | 69 | 8.30 | 2 | *** |
| Between 7-Point and Magnitude Scale | 107.06 | 69 | 12.54 | 2 | *** |
| **Equality-Constraints on Validity Estimates** | | | | | |
| Between All Scales | 116.82 | 74 | 22.30 | 7 | *** |
| Between 11 and 7-Point Scale | 103.02 | 71 | 8.50 | 4 | * |
| Between 11-Point and Magnitude Scale | 103.44 | 70 | 8.92 | 3 | ** |
| Between 7-Point and Magnitude Scale | 108.32 | 70 | 13.80 | 3 | *** |

*Note.* [†] Model allows for variation in reliability and validity estimates across all traits and methods. See Table 4 for respective parameter estimates. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$. *Data Source.* SOEP-Pretest 2004.

**Table 5** Data Quality of Alternative Modes of Data Collection Measuring Satisfaction.

| | | Validity | | | Method Effect | | | Reliability |
|---|---|---|---|---|---|---|---|---|
| | | Life | Health | Income | SAQ | PAPI | CAPI | |
| SAQ | Life | 0.85 | | | 0.53 | | | 0.79 |
| | Health | | 0.86 | | 0.52 | | | 0.79 |
| | Income | | | 0.87 | 0.50 | | | 0.82 |
| PAPI | Life | 0.93 | | | | 0.38 | | 0.80 |
| | Health | | 0.93 | | | 0.36 | | 0.83 |
| | Income | | | 0.93 | | 0.36 | | 0.84 |
| CAPI | Life | 0.94 | | | | | 0.33 | 0.80 |
| | Health | | 0.95 | | | | 0.30 | 0.88 |
| | Income | | | 0.95 | | | 0.31 | 0.85 |

*Note*. All estimates significant at $p<.05$; $N$=2,249; $\chi^2 = 173.50$ with 246 df. *Data Source*. SOEP 2002 and 2003.

**Table 6** Chi-Square Difference Tests Between Nested MTMM-Models (with Equality Constraints) and the Basic Model (without Equality Constraints).

| | Model Fit | | Difference in Model Fit | | |
|---|---|---|---|---|---|
| | $\chi^2$ | df | $\Delta\chi^2$ | $\Delta$df | p($\Delta$) |
| **Basic Model**[†] | 173.50 | 246 | - | - | - |
| | | | | | |
| **Equality-Constraints on Reliability Estimates** | | | | | |
| Between All Modes of Data Collection | 188.07 | 252 | 14.57 | 6 | ** |
| Between SAQ and PAPI | 176.98 | 249 | 3.48 | 3 | |
| Between SAQ and CAPI | 187.78 | 249 | 14.28 | 3 | *** |
| Between PAPI and CAPI | 178.13 | 249 | 4.63 | 3 | |
| **Equality-Constraints on Validity Estimates** | | | | | |
| Between All Modes of Data Collection | 199.20 | 254 | 25.70 | 8 | *** |
| Between SAQ and PAPI | 186.68 | 250 | 13.18 | 4 | *** |
| Between SAQ and CAPI | 198.21 | 250 | 24.71 | 4 | *** |
| Between PAPI and CAPI | 174.85 | 250 | 1.35 | 4 | |

*Note.* [†] Model allows for variation in reliability and validity estimates across all traits and methods. See Table 6 for respective parameter estimates. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$. *Data Source.* SOEP 2002 and 2003.

**Table 7** The Effect of Income on the Satisfaction with Income Across Response Formats (OLS Model).

| | |
|---|---|
| Intercept | - 0.45** (0.20) |
| Response Format | |
| *7-Point Scale* | - |
| 11-Point Scale | - 0.72***(0.25) |
| Magnitude Scale | 0.48** (0.23) |
| Household Income (ln) | 0.15***(0.03) |
| Response Format x Household Income | |
| *7-Point Scale x HH-Income* | - |
| 11-Point Scale x HH-Income | 0.10*** (0.04) |
| Magnitude Scale x HH-Income | - 0.05 (0.03) |
| Model Fit | |
| N | 547 |
| $R^2$ | 0.20 |

*Note.* *** p < 0.01; ** p < 0.05; * p < 0.10; standard errors in parentheses.
*Data Source.* SOEP-Pretest 2004.

**Table 8** Hierarchical Regression Models of Problems During Data Administration.

| | Model 1 Non-Response | | Model 2 Elapsed Time | | Model 3 Reluctance | | Model 4 Misunderstanding | |
|---|---|---|---|---|---|---|---|---|
| Intercept 1 | - 2.44** | (0.98) | 4.35*** | (0.10) | 1.15*** | (0.35) | 0.62* | (0.32) |
| Intercept 2 | - | | - | | 2.96*** | (0.37) | 2.31*** | (0.33) |
| Intercept 3 | - | | - | | 3.92*** | (0.38) | 3.49*** | (0.35) |
| Intercept 4 | - | | - | | 4.48*** | (0.40) | 4.20*** | (0.36) |
| Intercept 5 | - | | - | | 5.13*** | (0.42) | 4.88*** | (0.40) |
| **Position in Questionnaire** | | | | | | | | |
| *Beginning* | - | | - | | - | | - | |
| End | 0.61 | (0.51) | - 0.61*** | (0.05) | 0.34** | (0.14) | 0.08 | (0.13) |
| **Response Format** | | | | | | | | |
| *7-Point Scale* | - | | - | | - | | - | |
| 11-Point Scale | - 0.53 | (1.13) | - 0.14 | (0.13) | 0.82** | (0.38) | - 0.12 | (0.36) |
| Magnitude Scale | - 0.16 | (0.97) | 0.43*** | (0.13) | 1.01*** | (0.36) | 0.98*** | (0.33) |
| **Respondent Characteristics** | | | | | | | | |
| Age | - 0.02 | (0.01) | 0.00** | (0.00) | 0.02*** | (0.00) | 0.01*** | (0.00) |
| Education | - 0.34** | (0.16) | - 0.02 | (0.01) | - 0.05 | (0.04) | - 0.08** | (0.03) |
| **Respondent Characteristics x Response Format** | | | | | | | | |
| *Age* x *7-Point Scale* | - | | - | | - | | - | |
| *x* 11-Point Scale | 0.00 | (0.02) | 0.00 | (0.00) | - 0.01* | (0.01) | - 0.00 | (0.00) |
| *x* Magnitude Scale | 0.00 | (0.01) | - 0.00* | (0.00) | - 0.01** | (0.00) | - 0.01 | (0.00) |
| *Education* x *7-Point Scale* | - | | - | | - | | - | |
| *x* 11-Point Scale | 0.14 | (0.19) | - 0.01 | (0.02) | - 0.10** | (0.05) | - 0.01 | (0.05) |
| *x* Magnitude Scale | - 0.04 | (0.17) | 0.03* | (0.02) | - 0.06 | (0.04) | - 0.03 | (0.04) |
| **Random Effects** | | | | | | | | |
| Variances | | | | | | | | |
| Level 1, Observation | 1.00 | | 0.27*** | (0.02) | 1.00 | | 1.00 | |
| Level 2, Respondent | 0.96 | (0.77) | 0.01 | (0.01) | 0.41*** | (0.11) | 0.31*** | (0.10) |
| Level 3, Interviewer | 2.40* | (1.43) | 0.11*** | (0.02) | 2.70*** | (0.51) | 2.45*** | (0.43) |
| **Model Fit** | | | | | | | | |
| – Log Likelihood | - 122.25 | | - 1194.55 | | - 1247.14 | | - 1310.27 | |
| N | | | | | | | | |
| Level 1, Administration | 1512 | | 1390 | | 1514 | | 1514 | |
| Level 2, Respondent | 757 | | 749 | | 757 | | 757 | |
| Level 3, Interviewer | 142 | | 142 | | 142 | | 142 | |

*Note.* *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; standard errors in parentheses. *Data Source.* SOEP-Pretest 2004.

**Figure 1** Validity and Reliability in the MTMM Context

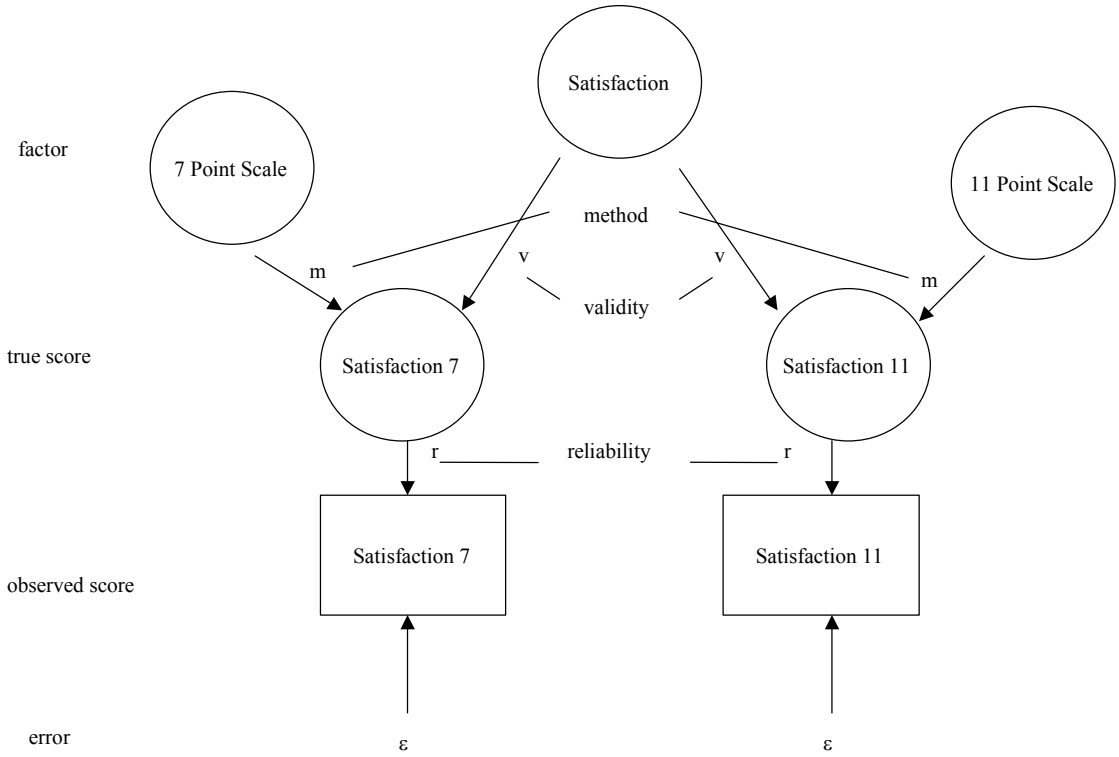
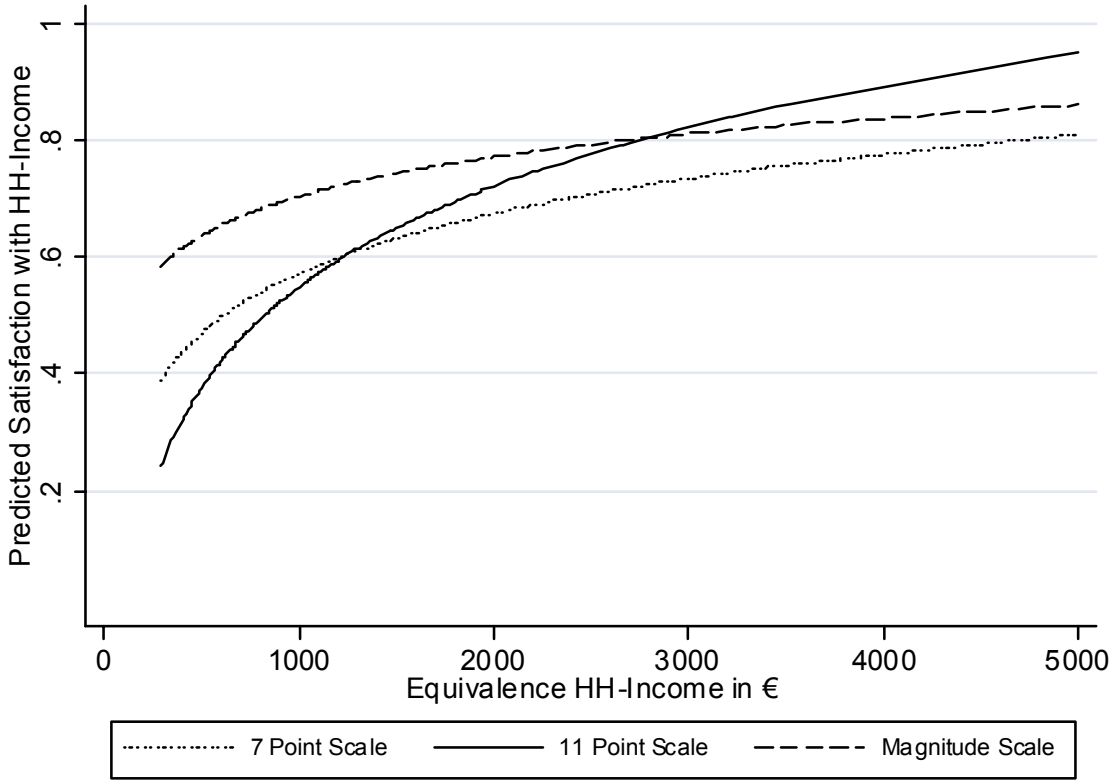


**Figure 2** Observed Functional Relationship Between Income and Satisfaction with Income



[1] A comprehensive review of well-being research is beyond the scope of this paper and can be obtained from Kahnemann et al. (1999) and Diener et al. (1999).

[2] One common view holds that subjective well-being refers to an individual's evaluation of experienced affect, happiness or satisfaction (e.g., Bradbury and Caplovitz 1965). Many researchers divide subjective well-being into an affective component (positive affect, negative affect, happiness) and a cognitive component (satisfaction) (e.g., Lucas et al. 1996). Whether this or alternative conceptualizations find empirical support is the subject of intense debate, however (e.g., Bradburn 1969; Andrews and Withey 1976; Stones and Kozma 1985; Glatzer 1984; Larsen et al. 1985).

[3] This paper focuses on the data quality of alternative questionnaire-based self-reports of subjective well-being alone. For an overview of non-reactive measures of well-being based on, for instance, brain electrical activity see e.g. Larsen and Fredrickson (1999).

[4] Another set of methodological studies on measuring subjective-well being focus on order effects. Schwarz et al. (1991) and Tourangeau et al. (1991) demonstrate that the order of surveying satisfaction with different aspects of life (e.g., health, income, housing, life in general) affects respondents' answers notably.

[5] Campbell et al. (1976) already used a 7-point satisfaction scale.

[6] Early applications of the 11-point satisfaction scale can be found in the British Quality of Life Survey (Hall et al. 1973) and the German Welfare Survey in 1978 (Siara 1980).

[7] The inadequacy of the closed-ended formats may become obvious in longitudinal studies of aggregate levels of well-being: In the last decades, objective living conditions have improved vastly in many western societies. At the same time, average life satisfaction remains rather stable (e.g., Kenny 1999; Diener and Oishi 2000; Easterlin 2001). The divergence between objective living conditions and subjective satisfaction with life on the aggregate level may be attributed to the limitations of closed-ended response formats measuring subjective well-being.

[8] Note that the analysis compares in-person modes of data collection only, i.e. even the analyzed self-administered questionnaires were filled in while an interviewer was present. Hence, in all three cases, interviewers may have clarified misunderstandings.

[9] CAPI protocols often include edit checks and controls for the interviewer. For instance, warning messages may appear on the screen that certain answers are not credible or that certain answers differ in implausible ways from information obtained in previous interviews. SOEP-Interviewers in CAPI settings are therefore hesitant to allow respondents to have a detailed look at the screen.

[10] Another reason for the improved data quality of visual presentation in SAQ as compared to auditory presentation in PAPI and CAPI may be the absence of interviewer effects, i.e. the tendency of respondents to react to characteristics of interviewers when providing their answers. Respondents may, as a form of impression management, report higher levels of satisfaction if they perceive the interviewer as being happy and optimistic (e.g., Paulhus 1991). In the SOEP context, where even in self-administered questionnaires interviewers personally contact respondents and deliver the questionnaire, one cannot exclude the possibility of interviewer effects even if they are not present when respondents fill in the questionnaire. However, results by Tourangeau et al. (1997) indicate that in such a setting, interviewer effects can be ignored.

[11] The wording of such a scale may be as follows: "Please assume that - in terms of satisfaction - the score 50 describes a situation in which dissatisfaction and satisfaction are in perfect balance. Please express your personal satisfaction with a number that relates to the score 50: Numbers smaller 50 mean dissatisfaction, numbers greater 50 mean satisfaction. For instance, if your satisfaction is only half of the balanced situation, please assign the score 25. If your satisfaction is three times as high as in the balanced situation, please assign the score 150."

[12] There is much debate on the choice of a formal model for the analysis of multitrait multimethod matrices. Many parameterizations have been suggested in the literature, such as the correlated uniqueness model (Marsh 1989), the direct product model (Browne 1984) or the true score model applied here (Saris and Andrews 1991). A full discussion of the merits and caveats of these different models is beyond the scope of this study. Note that this study draws on the ´true score model` proposed by Saris and Andrews (1991) as it represents one of the accepted and frequently applied parameterizations.

[13] Bohrnstedt (1983) specifies the validity in the MTMM context as construct validity.

[14] Note that the squared (standardized) validity coefficient $v^2$ represents the validity of the measure and the squared (standardized) method coefficient $m^2$ represents the method effect. Since $m^2 = 1 - v^2$, the method effect is equal to the invalidity due to the method used.

[15] In about 8 % of all cases, administrating these questions took more than five minutes. Since it appears likely that this indicates an interruption of the interview, these cases were exclude from the analysis.

[16] Since comparable information is partly unavailable for the difference between modes of data collection (e.g., there is no external information in the elapsed time of self-administered questionnaires), the ease of administration is discussed for alternative response formats only.

[17] Note that alternative model specifications (e.g., ordered probit) lead to essentially the same results and are therefore not reported in the main text or tables.

[18] The main effect of age, which is associated with the 7-point scale, is smaller for the magnitude scale with respect to the elapsed time of the interview and respondents' reluctance. The latter also holds for the 11-point scale. However, for the 11-point scale, Model 3 predicts a negative effect of education, i.e. the more educated the respondents, the less reluctant they are. Finally, the magnitude scale entails an effect of education, i.e. the more educated a respondent, the more time she needs for the magnitude scale.

[19] This does of course not imply that the 11-point scale and personal interviewing are optimal measuring attitudes under all circumstances. Rather, the conclusions drawn on basis of the analysis of this paper are limited to subjective well-being only.