PAPER ACCEPTED FOR PUBLICATION IN FIELD METHODS

SCHEDULED FOR VOL. 24, NO. 3, AUGUST 2012

Why Semantic Differentials in Web-Based Research Should be Made From Visual Analogue Scales and Not From 5-Point Scales

Frederik Funke

Kassel, Germany

Ulf-Dietrich Reips

Universidad de Deusto and IKERBASQUE, Basque Foundation for Science, Spain

Author Note

Parts of this study have been presented at the ISA RC33 7th International Conference on Social Science Methodology, Naples, Italy, September 1–5, 2008, under the title "Assessing semantic differentials with visual analogue scales in Web surveys". The authors would like to thank the anonymous reviewers for their helpful remarks and E.-D. Lantermann (University of Kassel) and his students who participated in the experiment.

Correspondence concerning this article should be addressed to Frederik Funke, Steinbruchweg 10, 34123 Kassel, Germany, URL: http://frederikfunke.net, email: email@frederikfunke.net

Abstract

In a Web experiment, participants were randomly assigned to 2 semantic differentials either made from discrete 5-point ordinal rating scales or from continuous visual analogue scales (VASs) with 250 gradations. Respondents adjusted their ratings with VASs more often to maximize the precision of answers, which had a beneficial effect on data quality. No side effects like differences in means, higher dropout, more nonresponse, or higher response times were observed. Overall, the combination of semantic differentials and VASs results in a number of advantages. Potential for further research is discussed.

Introduction

Researchers have an extensive methodological repertoire at their hands to design questionnaires that assist participants in giving accurate answers and maintaining their willingness to cooperate. Especially Web-based methods offer rich ways to alter questionnaire design. One convenient possibility to optimize questionnaire design is to alter the available response scales. However, changes in response scales can affect the question answer process, especially question understanding as well as the formatting of answers (e.g., Sudman, Bradburn, & Schwarz, 1996), and seriously impacting given ratings (see deLeeuw, Hox, & Dillman, 2008; Dillman, Smyth, & Christian, 2009; Funke, Reips, & Thomas, 2011; Groves, Fowler, Couper, Lepkowski, Singer, & Couper, 2004; Krosnick, 1999; Schwarz, 1999).

Semantic Differentials

A special way of presenting multiple, related survey items are semantic differentials where respondents are faced with different aspects of a single latent variable, e.g., when the present mood is described along bipolar items like *calm – excited*, *happy – sad*, and *depressed – cheerful*. C. E. Osgood is credited for introducing this way of presenting items in the 1950s (Osgood, 1952; Osgood, Suci, & Tannenbaum, 1957). By now it is an established measurement device used in many fields (e.g., psychology, sociology, and linguistics). A battery of contrasting, bipolar verbal anchors is presented in form of a matrix (see experimental section, Figure 2).

It is known that respondents are likely to see items not independently but in the context of one another (see Dillman et al., 2009; Schwarz & Oyserman, 2001). Thus, when measuring unrelated items it is advantageous to present each item on a separate page (see Reips, 2002) to minimize context, anchoring and contrast effects (e.g., Sudman, Bradburn, & Schwarz, 1996). In a semantic differential all aspects are intentionally presented on the same (Web) page to emphasize that all items are related to the topic at hand and that respondents are asked to give ratings regarding all aspects.

Research on the historic origin of semantic differentials revealed that this type of rating scale was initially made from continuous, not from discrete rating scales. McReynolds and Ludwig (1987) report that a device very similar to what we consider a semantic differential

nowadays – even though it did not contain contrary verbal labels on either side – was used as early as the beginning of the 19th century: "a metal plate […] had 10 scales, each marked off in 100 parts, and labeled […]. A system of sliding markers was provided so that a […] judged position on each scale could be graphically displayed" (p. 282). More than 180 years later we transfer this approach into cyberspace, substituting the metal plate with a Web browser's interface and the sliding markers with visual analogue scales.

Web-Based Data Collection

Web-based data collection has become a well-established instrument in the world of survey methodology (Best & Krueger, 2004; deLeeuw, Hox, & Dillman, 2008; Dillman & Bowker, 2001; Dillman et al., 2009; Joinson, McKenna, Postmes, & Reips, 2007) and science in general (Reips, 2008). In comparison to laboratory settings, Web-based research for example greatly extends the access to a large number of participants and to special populations (see Fuchs, 2008; Mangan & Reips, 2007), and it allows the presentation of various multimedia stimuli (e.g., Fuchs & Funke, 2007).

Nevertheless, there are some pitfalls. For example, respondents have to have a certain degree of computer literacy to be able to complete a Web-based questionnaire in a meaningful way. Another worry is that the setting in which respondents take part in a study cannot be controlled. On the one hand, these non-standardized situations can be a good reason why mode and other effects occur and Web experiments sometimes may not produce the same results as experiments conducted in laboratories (Reips, 2002, 2007). One the other hand, the validity of results obtained from testing persons in real-life environments without (involuntary) influence of the experimenter should be higher (see Honing & Reips, 2008).

Web-based questionnaires are prone to involuntary changes in layout due to different client-sided software configurations. For example, poor HTML code can result in uneven spacing between radio buttons, which can shift ratings as Tourangeau, Couper, and Conrad (2004) demonstrated, and even small changes in layout and visual design can affect answers (e.g., Couper, Traugott, & Lamias, 2001; Dillman & Bowker, 2001; Reips, 2010; Smyth, Dillman, Christian, & Stern, 2006). New technologies tend to foster the development of new rating scales or measurement devices. However, it must be guaranteed that new methods do not systematically reduce data quality. Buchanan and Reips (2001) observed that education and

personality traits were confounded with the use of certain technologies. Funke, Reips, and Thomas (2011) found an interaction between formal education, technology, and dropout rate.

Overall, one should pay much attention to a straightforward implementation and control for technological background variables (e.g., operating system, browser, and availability of technologies). It seems wise to follow the low-tech principle, to use as many robust standard procedures as possible (e.g., HMTL instead of Flash or Java), and to keep technological requirements for participation as low as possible.

Visual Analogue Scales (VASs)

VASs are continuous graphical rating scales, first described by Hayes and Paterson (1921). The obvious advantage over discrete scales is that answers are not restricted to a certain number of response options but very fine gradations can be measured. In computerized data collection with VASs each pixel in length corresponds to a possible value (Reips & Funke, 2008). Figure 2 illustrates a considerable difference in reported intensity of the bottom three items with VASs that cannot be observed with 5-point scales. Additionally, data collected with VASs can be subjected to a greater number of statistical procedures than data collected with categorical scales (e.g., recoding by empirical quantiles) and goodness of fit tests are more powerful.

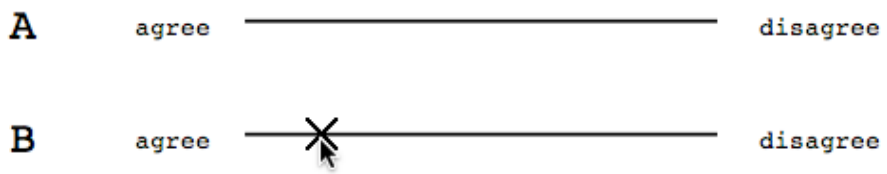

Figure 1. Agree-disagree VAS in Web-based questionnaire after loading of the Web page (A) and with marker after first click on the rating scale (B).

Many paper-based studies – especially in the medical sector, e.g., for assessment of subjective phenomena like fatigue or pain (Cork, Isaac, Elsharydah, Saleemi, Zaviska, & Alexander, 2004) – were not able to show differences between VASs and ordinal scales regarding mean ratings (also see Averbuch & Katzper, 2004; Flynn, van Schaik, & van Wersh, 2004). In a paper and pencil study on pain, Myles, Troedel, Boquest, and Reeves (1999) and Myles and Urquhart (2005) found that data from VASs are linear and equal changes in intensity

correspond to equal differences in length on VASs. Gerich (2007) found that mode differences between paper and Web are smaller with VASs than with 5-point scales. Hofmans and Theuns (2008) conclude that in a paper-based study "VASs can be considered as linear scales and that the type of end anchors used has no effect on the linearity of the VAS data" (p. 401). Reips and Funke (2008) found that even in Web surveys VASs fulfill many requirements of measurement on the level of an interval scale. With mentally well-represented constructs equal changes in intensity correspond to equal changes in ratings on VASs. So, differences between ratings on VASs can be interpreted in a meaningful way and the prerequisites for many statistical procedures are met.

In contrast to the historical origins, nowadays semantic differentials are rarely made of VASs. The most likely explanation is the effort associated with the manual readout in paper-based studies. The administration of VASs is burdensome and prone to error when distances are measured and read by eye and hand. However, this is no issue in computerized data collection where data are read out automatically.

Research Questions and Hypotheses

Making sound ratings on semantic differentials is a more complex task than just answering single items. Respondents have to consider not only a single item but also the relationships between the items presented to give sound overall judgments. Our research focuses on how the rating scale used with semantic differential influences the measurement process and data quality.

VASs allow giving finely nuanced ratings regarding many aspects of the construct being measured. However, this large number of possibilities may be too demanding. Ratings with *ordinal scales* pose the problem of shared ranks whenever the number of response options is smaller than the number of items. Thus, giving consistent ratings regarding the relation between all items is impossible. Ordinal rating scales always require decisions about compromises and thus may increase cognitive burden. Additionally, if continuous concepts are to be rated on ordinal scales, a transformation, a segmentation, and mapping to the nearest category have to be mastered by the respondent.

Web surveys allow unobtrusive data collection with nonreactive methods if response times or user actions like mouse clicks are recorded in the background without respondents taking notice (e.g., Heerwegh, 2003). Using JavaScript we implemented a measure of how long it takes to make ratings and how often ratings are modified. Thus, we made use of the enhanced technologies available in Internet-based research when investigating the manual answering process in remote participants (Stieger & Reips, 2010). Regarding the observable part of the question-answer process, we hypothesize that respondents take advantage of the possibilities of VASs. Thus, the greater number of response options with VASs should lead to deeper cognitive processing which in turn should increase response times in comparison to ordinal scales. Additionally, ratings should be changed more frequently with VASs when respondents make an effort to give a sound judgment regarding all items. Overall, measurement with VASs should lead to higher data quality.

The Experiment

The Web experiment was conducted as part of a study on personal preferences, taste, and style preferences among a convenience sample of students of psychology at the University of Kassel, Germany, from January 7–30, 2008. Respondents provided a personal, but anonymous code to be able to get a feedback of their individual results in comparison to the overall mean as per capita incentive (see Göritz, 2006).

Questionnaire

The questionnaire consisted of 83 consecutive Web pages. To utilize advantages of the warm-up technique in Internet-based research (Reips, 2002) – i.e., the experimental manipulation is only introduced after a number of neutral items, so the dropout that naturally occurs at the beginning of Internet-based studies can not be attributed to the experimental manipulation – the experimental manipulation of the type of rating scales in the semantic differentials was placed on pages 81 and 82. On each of these two pages was one semantic differential consisting of 13 bipolar items, assessing different aspects of the respondents' style regarding furnishing (semantic differential 1 "My furnishing is:") and clothing (semantic differential 2 "Mostly my clothing

is:"). The labeling of the 13 items was the same on both pages (see Figure 2). The questionnaire contained no other semantic differentials.

Procedure

In a between-subjects design respondents were randomly assigned either to semantic differentials made from 5-point rating scales, implemented with HTML radio buttons, or to semantic differentials made from VASs with 250 pixels in length corresponding to 250 gradations. Randomized distribution of participants to the experimental conditions took place when a participant mouse-clicked the submit button on the first page of the questionnaire. The VASs were generated with the free Web service VAS Generator (maintained by the authors, located at http://vasgenerator.net, see Reips & Funke, 2008) and implemented using JavaScript. Type of rating scale was the same on both Web pages.

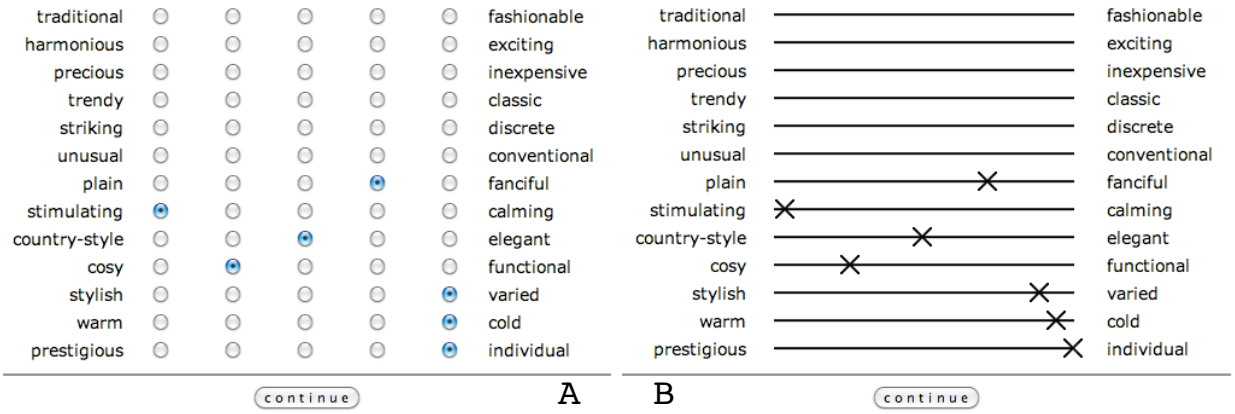

Figure 2. Rating scales (top seven items after load of the Web page, bottom six items with given ratings) varied between subjects: 5-point scale made from radio buttons (A) and VASs with 250 possible values (B).

Directly after loading the Web pages containing the semantic differentials, neither scale showed any marker (Figure 2, top six items). No special instruction on how to answer semantic differentials or how to use the respective rating scale was provided. The marker – a checked radio button with the 5-point scales and a cross with the VASs (see Figure 2, bottom items) – only appeared after clicking on the scales. Judgments with VASs were made in the following way (working examples can be seen at http://vasgenerator.net): Respondents clicked the blank

line at the appropriate point and the marker appeared at the very position. To modify the rating any other position on the VAS could be clicked. As with the radio button version, it was not possible to move the marker on the VAS by dragging it with the mouse. Every click on the VASs was recorded just as every click on a radio button. In both conditions respondents could adjust all ratings as often as they wanted and no rating was mandatory.

Results

Participants

The study the experiment was embedded in was advertised in an introductory psychology class. Overall, 413 participants started the questionnaire, i.e., they agreed to participate by clicking the submit button on the first page of the experiment where randomization to the experimental conditions took place and 278 participants reached the experimental section. At the beginning of the study, participants were asked for an anonymous code. Not providing a code indicated a low motivation to seriously participate in this study, so we excluded these seven cases. We excluded three additional cases (1.1%) of JavaScript not being enabled in the respondents' Web browsers. Overall, we thus obtained a sample of 268 cases, 134 in each experimental condition. Due to the warm-up technique no dropout occurred on the two pages of the experiment.

Nonresponse

In both conditions 7.1% of the data set (19 cases with each scale) were incomplete, showing at least one missing value during the experiment. Following the common definition of *lurking* (e.g., Bosnjak, 2001), we take lurkers as respondents who do not provide even a single answer within a study. We detected two lurkers, one per condition. Additionally, we found two cases in the condition with the radio button scale type that did not provide a single value for either semantic differential (one-page lurker). We excluded these four cases, resulting in 264 cases (133 with VASs and 131 with 5-point scale) for analysis of item nonresponse.

Overall, we found no statistically significant difference in item nonresponse between VASs and the 5-point scales. However, there was a trend for less nonresponse with VASs that is consistent across both semantic differentials: Semantic differential 1: M(VAS) = 0.28 (SD =

1.34), *M*(5-point) = 0.42 (*SD* = 1.63); semantic differential 2: *M*(VAS) = 0.40 (*SD* = 1.72); *M*(5-point) = 0.50 (*SD* = 1.89).

For the analysis of frequency of changing answers, means, correlations, and response time, we excluded all cases with at least one missing item, resulting in 230 cases, 115 with each type of scale. All in all, data from 17.3% of respondents did not meet the strict criteria for inclusion of data we set for the analyses that follow.

The proportion of females among the 230 cases for the following analyses was 76.5% with VASs and 71.3% with 5-point scales. The mean age was *M* = 23.1 years (*SD* = 2.2) with VASs and *M* = 23.7 (*SD* = 3.8) years with 5-point scales. None of these demographic differences between the experimental conditions was statistically significant (all $p > .10$).

Adjusting Responses

To make inferences on the process of decision making we analyzed how often ratings were modified. The effort to give accurate responses is reflected in the number of changes. To be more conservative, we disregarded data from one respondent with an extremely high frequency of changes (53 changes with VASs and the second semantic differential).

Overall, with the first semantic differential 66.1% of the respondents changed at least one rating with VASs and 57.4% with the 5-point scale. Fewer changes were made with the second semantic differential: 54.8% of the respondents changed at least one answer with VASs versus 50.4% of the respondents with the 5-point scale. The maximum number of changes was 12 for all conditions, except for the first semantic differential with VASs where one respondent made 23 changes.

On the first page of the experiment, the mean number of clicks needed to complete all 13 items was *M* = 16.6 (*SD* = 4.9) with the VASs and *M* = 14.5 (*SD* = 2.1) with the 5-point scales, $F(1, 228) = 18.42$, $p < .001$, $\eta^2 = .075$. In other words: on average, 3.7 changes occurred with VASs and only 1.5 with the 5-point scales. Thus, ratings with VASs were modified more than twice as often, see Figure 3. On page two, answering the semantic differential with VASs, *M* = 14.9 (*SD* = 2.8), again took more clicks than answering with the 5-point scales, *M* = 14.0 (*SD* = 1.7), $F(1, 228) = 8.19$, $p = .005$, $\eta^2 = .035$. This corresponds to 1.9 changes with VASs and 1.0 change with the 5-point scales. Again, ratings with VASs were modified about twice as often, see Figure 3.

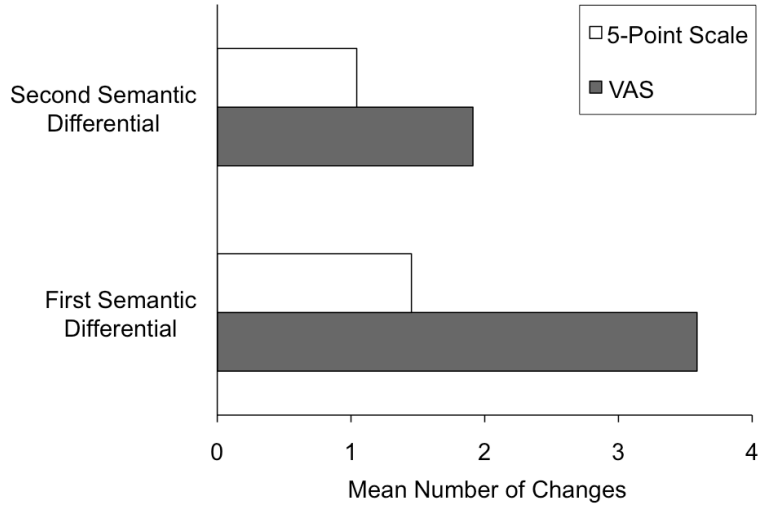

Figure 3. Means number of changes for answering semantic differentials made up with 5-point scales and made up with VASs.

Ratings

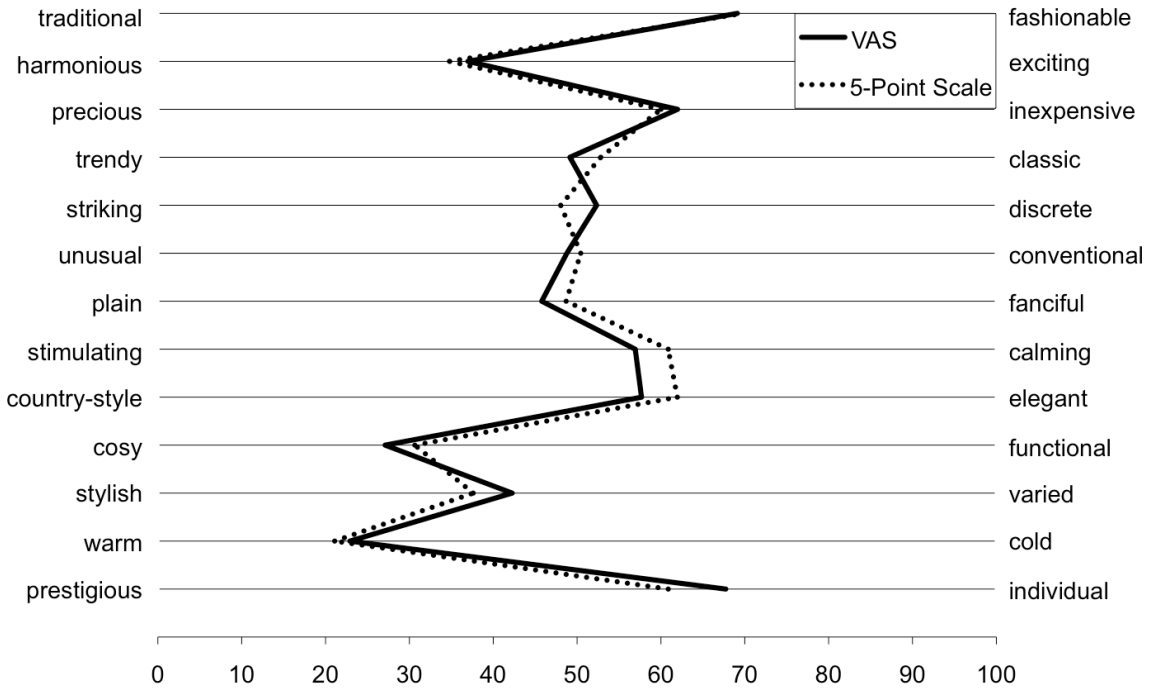

Figure 4. Mean ratings semantic differential 1.

For better comparability, data from both scales were recoded ranging from 0 to 100. Thus, a difference of one unit can be interpreted as one percentage point difference between ratings. Overall, both rating scales lead to similar mean ratings. After Bonferroni correction no difference in mean ratings was detected, neither for the first semantic differential (see Figure 4) nor for the second semantic differential (see Figure 5).

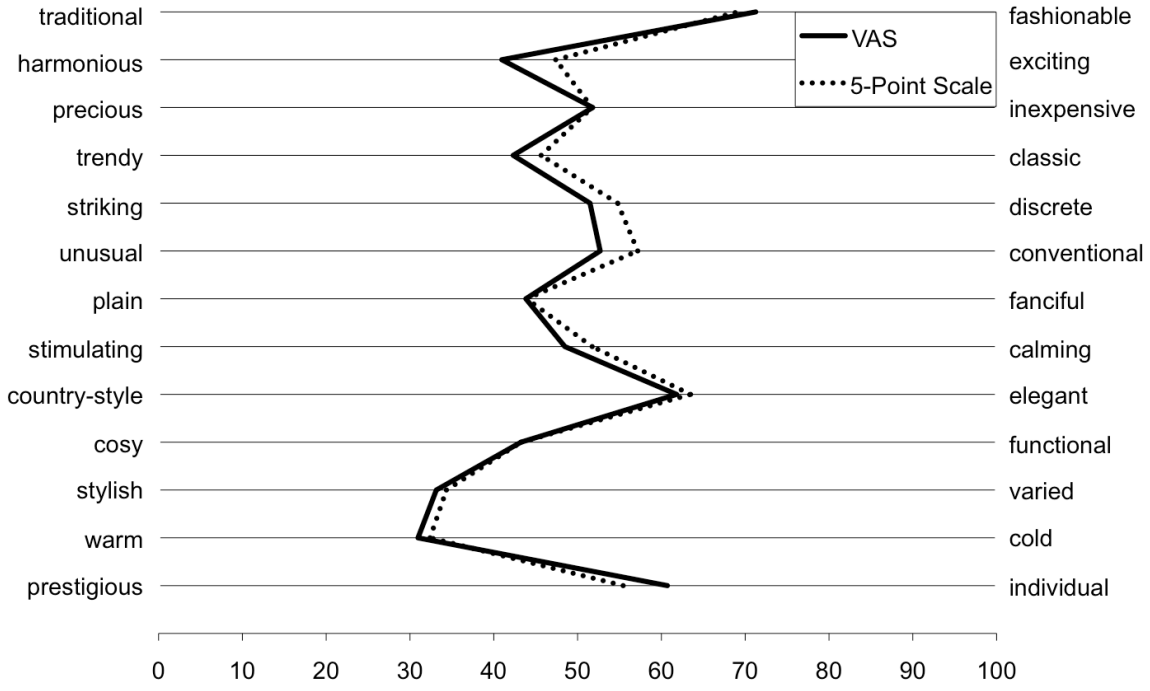

Figure 5. Mean ratings semantic differential 2.

Correlations

Both concepts we observed with the semantic differentials (style of furnishing and style of clothing) should be highly correlated, as we expect a general factor "sense of style" to influence both variables. We used the correlations between identical aspects (e.g., *plain – fanciful* for furniture style and *plain – fanciful* for clothing style) to infer on measurement error (see Groves et al, 2009). This concept addresses the difference between an ideal measurement process and the actually obtained response. A desired measurement with lower measurement error would show high correlations between variables measuring the same aspect and low correlations – indicating high discrimination – between different aspects (e.g., *plain – fanciful* and *trendy – classic*). We

Table 1. *Correlations Between Different Aspects of Furnishing and Clothing Measured With VASs and 5-Point Scales*

Aspect	1	2	3	4	5	6	7	8	9	10	11	12	13
	VASs (n = 115)												
1	**.47****	.07	-.23*	-.20*	-.31**	-.20*	.18	-.15	.25**	.08	-.23**	-.11	-.05
2		**.43****	.18	.13	-.19*	-.10	.06	-.21*	-.29**	.02	.46**	.16	.04
3			**.63****	.10	.36**	.18	-.30**	.26**	-.35**	-.25**	.35**	-.09	.09
4				**.46****	.39**	.38**	-.32**	.26**	-.02	-.10	.19*	.20*	-.31**
5					**.46****	.36**	-.36**	.33**	.19*	-.01	-.17	.14	-.04
6						**.54****	-.36**	.24**	.18	-.14	-.21*	.01	-.18
7							**.60****	-.18*	-.11	-.09	.11	-.14	.19*
8								**.36****	.07	.01	-.15	.05	.08
9									**.56****	.00	-.46**	-.20*	.00
10										**.34****	.02	.38**	-.10
11											**.64****	.15	.02
12												**.40****	-.01
13													**.39****
	5-point scales (n = 115)												
1	**.46****	.25**	-.30**	-.37**	-.27**	-.11	.20*	-.11	.22*	.13	-.15	.00	-.04
2		**.35****	.01	-.11	-.01	-.15	-.07	.00	-.16	.20*	.20*	.38**	.08
3			**.44****	.26**	.09	-.03	-.16	.05	-.34**	-.06	.32**	.05	.11
4				**.44****	.18	.14	-.32**	.13	-.18	-.02	.09	-.08	.05
5					**.33****	.24*	-.35**	.21*	-.02	-.15	.07	.01	-.05
6						**.46****	-.24**	.16	.10	-.13	-.15	-.10	-.21*
7							**.48****	-.21*	.06	-.08	-.03	-.12	-.01
8								**.17****	.04	-.25**	-.01	.02	-.05
9									**.58****	.12	-.26**	-.06	-.14
10										**.33****	-.08	.28**	.00
11											**.53****	.18	.14
12												**.39****	.17
13													**.32****

Note. Correlations between corresponding aspects are in bold type.

*Correlation is significant at the .05 level (two-tailed). **Correlation is significant at the .01 level (two-tailed).

computed Pearson's correlation coefficients between all 13 variables for comparing measurement with VASs to measurement with 5-point scales, see Table 1.

Correlations between corresponding aspects. For this analysis we looked at the correlations between the 13 corresponding aspects (see Table 1, principal diagonal printed in bold type). Mean correlations measured with VASs (M = .48, SD = .10) were higher than correlations with the 5-point scales (M = .41, SD = .11). This difference is marginally significant and has a large effect size, $F(1, 25) = 3.41$, $p = .077$, $\eta^2 = .125$.

Correlations between distinct aspects. If VASs produced higher correlations in general, not only with corresponding aspects but also with unrelated aspects, this would be an undesired effect (e.g., always just clicking in the middle of a rating scale would have produced high correlations between all aspects). Thus, we computed the mean correlations for all unrelated aspects (see Table 1, printed in regular type). Mean correlations between unrelated aspects of the construct *style* were zero measured with VASs (M = .00, SD = .21) and nearly zero with the 5-point scales (M = -.01, SD = .17).

Response Times

In contrast to a laboratory setting, there can be multiple sources of participant distraction in Web surveys not related to the study (e.g., pet distracting respondent, incoming phone calls). We decided to remove all unreasonably high response times from analyses and identified outlier as proposed by Tukey (1977): Within every experimental condition, all response times lower than 25th percentile minus one and a half times the interquartile range were omitted as well as times higher than 75th percentile plus one and a half times the interquartile range. For further analyses only respondents with moderate response times on both pages were considered, resulting in 33 respondents (15 with VASs and 18 with 5-point scales) that were excluded from analyses of response time.

With the first semantic differential completion with VASs (M = 66.4 seconds, SD = 19.4) did not take statistically longer than completion with 5-point scales (M = 62.7 seconds, SD = 20.0). On the second page, the absolute difference was even lower: M(VAS) = 48.9 seconds (SD = 13.9), M(5-point) = 47.8 seconds (SD = 14.1).

Discussion

VASs allow respondents to communicate subjective values more exactly than radio button scales. Moreover, the number of response categories communicates how elaborated the expected answer should be. A small number of response options implicitly conveys the message that roughly estimated answers are sufficient, whereas a large number of response options can be understood as an instruction to maximize cognitive efforts (see also Schwarz, 1999). These advantages of VASs should be especially valuable with semantic differentials, the prime method for assessing multiple aspects of one construct on a single (Web) page. Aim of this study was to test if the theoretical advantages of semantic differentials made from VASs with a range of 250 values in comparison to semantic differentials made from 5-point scales hold in a Web experiment.

Decision Making Processes

We considered response times and frequency of adjusting responses to infer on the decision making processes. VASs had a clear influence on the number of changes. Ratings with VASs were modified around twice as often as with 5-point scales. In line with our hypothesis, respondents indeed made use of VASs' fine gradations. Are more adjustments with VASs an indicator for deeper cognitive processing? In contrast to our expectations, we found no difference in response times. Regardless of the available rating scale respondents seem to be willing to invest a certain amount of time for dealing with the task of answering the items in a semantic differential. With discrete 5-point scales' limited number of response options and the problem of shared ranks this time is likely to be used for formatting the answer, i.e., to make a *compromise* in given responses. In contrast, with continuous VASs respondents do not have to bother about restrictions of the rating scale. Instead they use the time to *maximize the precision* of the given answers. This process of maximizing efforts should find its expression in data quality.

The indicator for data quality we used was the correlation between style of furnishing and style of clothing. Our reasoning is that both domains of style should be highly correlated as they are influenced by the same general factor. A superior measurement should lead to high correlations between corresponding aspects of style and to no correlations between distinct aspects. Indeed, correlations between corresponding aspects were significantly higher with VASs

in comparison to 5-point scales. Correlations between distinct aspects were around zero for both scales. We take these findings as an indicator that measurement with VASs has a beneficial influence on data quality.

Mean ratings and non-response

Overall, mean ratings are hardly affected by the rating scales we tested, the absolute difference was very small. This is in line with a large body of literature on VASs yielding the same mean score as ordinal scales (e.g., Averbuch & Katzper, 2004; Flynn et al., 2004; for Web-based research see Couper et al., 2006; Funke & Reips, 2010).

Regarding general indicators of data quality – dropout, lurking and item nonresponse – our results show no statistically significant differences between rating scales. There is – in contrast to Couper et al. (2006) – a tendency for even less item nonresponse with VASs, which may be owed to the low-tech implementation of VASs in the study presented.

Conclusion

For the first time, the present study compared semantic differentials made up with VASs to those made up with ordinal scales. VASs are suited to (literally) draw an accurate picture of all aspects of the construct being measured. In contrast, ratings with 5-point scales are troubled by shared ranks and the low degree of differentiation between aspects on the individual level. The inherent qualities of VASs (e.g., detection of small differences and far more possibilities for data analyses) are convincing arguments for the use of this type of rating scale. Overall, we consider the results reported as encouraging for the use of VASs with semantic differentials. Further research with eye tracking (e.g., Galesic, Tourangeau, Couper, & Conrad, 2008) or the User Action Tracer technique (Stieger & Reips, 2010) could help in understanding the exact way ratings are made with semantic differentials.

Our recommendation is to go back to the historical roots, transfer the original ideas of semantic differentials made from VASs, take it to the next level, and use this combination for Web-based data collection.

References

Averbuch, M., & Katzper, M. (2004). Assessment of visual analog versus categorical scale for measurement of osteoarthritis pain. *Journal of Clinical Pharmacology, 44,* 368–372.

Best, S. J., & Krueger, B. S. (2004). *Internet data collection.* Thousand Okaks, CA: Sage.

Bosnjak, M. (2001). Participation in non-restricted Web-surveys: A typology and explanatory model for item-nonresponse. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 193–207). Lengerich: Pabst Science Publishers.

Buchanan, T., & Reips, U.-D. (2001). Platform-dependent biases in online research: Do Mac users really think different? In K. J. Jonas, P. Breuer, B. Schauenburg & M. Boos (Eds.), *Perspectives on Internet research: Concepts and methods*. Retrieved January 6, 2009, from http://www.psych.uni-goettingen.de/congress/gor-2001/contrib/buchanan-tom

Cork, R. C., Isaac, I., Elsharydah, A., Saleemi, S., Zavisca, F., & Lori A. (2004). A comparison of the verbal rating scale and the visual analog scale for pain assessment. *The Internet Journal of Anesthesiology, 8*(1). Retrieved January 15, 2009, from http://www.ispub.com/ostia/index.php?xmlFilePath=journals/ija/vol8n1/vrs.xml

Couper, M. P., Tourangeau, R., Conrad, F. G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A Web experiment. *Social Science Computer Review, 24,* 227–245.

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

de Leeuw, E. D., Hox, J. J., & Dillman, D. A. (Eds.). (2008*). International handbook of survey methodology.* New York: Lawrence Erlbaum Associates.

Dillman, D. A., & Bowker, D. K. (2001). The Web questionnaire challenge to survey methodologists. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 159-178). Lengerich: Pabst.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys.* Hoboken, NJ: Wiley.

Flynn, D., van Schaik, P., & van Wersh, A. (2004). A comparison of multi-item Likert and visual analogue scales for the assessment of transactionally defined coping function. *European Journal of Psychological Assessment, 20,* 49–58.

Fuchs, M. (2008). Mobile Web survey: A preliminary discussion of methodological implications. In F. Conrad & M. Schober (Eds.), *Envisioning the future of survey interviews* (pp. 77–94). New York: Wiley.

Fuchs, M., & Funke, F. (2007). Multimedia Web surveys: Results from a field experiment on the use of audio and video clips in Web surveys. In M. Trotman et al. (Eds.), *The challenges of a changing world: Proceedings of the fifth international conference of the association for survey computing* (pp. 63–80). Berkeley: ASC.

Funke, F., & Reips, U.-D. (2010). *Making small effects observable: Reducing error by using visual analogue scales.* Manuscript submitted for publication.

Funke, F., Reips, U.-D., & Thomas, R. K. (2011). Sliders for the smart: Type of rating scale on the Web interacts with educational level. *Social Science Computer Review, 29,* 221–231.

Galesic, M., Tourangeau, R., Couper, M., & Conrad, F. (2008). Eye-tracking data: New insights on response order effects and other signs of cognitive shortcuts in survey responding. *Public Opinion Quarterly, 72,* 892–913.

Gerich, J. (2007). Visual analogue scales for mode-independent measurement in self-administered questionnaires. *Behavior Research Methods, 39,* 985–992.

Göritz, A. S. (2006). Incentives in Web studies: Methodological issues and a review. *International Journal of Internet Science, 1*, 58-70.

Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology* (2nd ed.). New York: Wiley.

Hayes, M. H. S., & Patterson, D. G. (1921). Experimental development of the graphic rating method. *Psychological Bulletin, 18,* 98–99.

Heerwegh, D. (2003). Explaining response latencies and changing answers using client side paradata from a Web survey. *Social Science Computer Review, 21,* 360–373.

Hofmans, J., & Theuns, P. (2008). On the linearity of predefined and self-anchoring visual analogue scales. *British Journal of Mathematical and Statistical Psychology, 61,* 401–413.

Honing, H., & Reips, U.-D. (2008). Web-based versus lab-based studies: A response to Kendall (2008). *Empirical Musicology Review, 3*(2), 73–77.

Joinson, A., McKenna, K., Postmes, T., & Reips, U.-D. (Eds.). (2007*). The Oxford handbook of Internet psychology*. Oxford: Oxford University Press.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5,* 213–236.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50,* 537–567.

Mangan, M. A., & Reips, U.-D. (2007). Sleep, sex, and the Web: Surveying the difficult-to-reach clinical population suffering from sexsomnia. *Behavior Research Methods, 39,* 233–236.

McReynold, P., & Ludwig, K. (1987). On the history of rating scales. *Personality and Individual Differences, 8,* 281–283.

Myles P. S., & Urquhart N. (2005). The linearity of the visual analogue scale in patients with severe acute pain. *Anaesthesia and Intensive Care, 33,* 54–58.

Myles, P. S., Troedel, S., Boquest, M., & Reeves, M. (1999). The pain visual analog scale: Is it linear or nonlinear? *Anesthesia & Analgesia, 89,* 1517–1520.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin, 49,* 197–237.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning.* Urbama, IL: University of Illinois Press.

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology, 49,* 243–256.

Reips, U.-D. (2007). The methodology of Internet-based experiments. In A. A. Joinson, K. McKenna, T. Postmes, & U.-D. Reips (Eds.), *The Oxford handbook of Internet psychology* (pp. 373–390). Oxford: Oxford University Press.

Reips, U.-D. (2008). How Internet-mediated research changes science. In A. Barak (Ed.), *Psychological aspects of cyberspace: Theory, research, applications* (pp. 268-294). Cambridge University Press. URL http://gsb.haifa.ac.il/~sheizaf/cyberpsych/12-Reips.pdf

Reips, U.-D. (2010). Design and formatting in Internet-based research. In S. Gosling & J. Johnson, *Advanced Internet methods in the behavioral sciences* (pp. 29–43). Washington, DC: American Psychological Association.

Reips, U.-D., & Funke, F. (2008). Interval level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behavior Research Methods, 40,* 699–704.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54,* 93–105.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Effects of using visual design principles to group response options in Web surveys. *International Journal of Internet Science, 1*(1), 6–16.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation, 22,* 127–160.

Stieger, S., & Reips, U.-D. (2010). What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior, 26*, 1488–1495.

Sudman, S., Bradburn, N., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

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