The Impact of Pictures on Best-Worst Scaling in Web Surveys

Marinică BĂRBULESCU

Faculty of Sociology and Social Work, University of Bucharest

Alexandru CERNAT·

Institute for Social and economic Research, University of Essex

Abstract: Motivation and burden are two of the most important aspects that influence response rates and dropouts in online surveys. As a result, we focus our analyses on how pictures and Best Worst Scaling (BWS), two solutions for each problem, interact in the Web medium. We use an experimental design that compares a BWS with pictures, the experimental group, and BWS without pictures, the control group. Results show that pictures influence measurement of BWS in six out of 16 items. We also observe that Couper's (2001) conclusion that concordant text and images have an accentuation effect while a discordant relationship between the two has an interference impact is partly true in our data. Eight out of the 16 items are at least partially influenced by the concordant/discordant variable while four fully respect this model. We conclude by discussing the impact of our findings and its limitations.

Keywords: *web survey, best-worst scaling, pictures, measurement.*

Introduction

While Web surveys are becoming more and more popular, researchers are struggling to compensate for its limitations and enhance its advantages. Two major problems affect online research. The first one refers to coverage of the population using online samples and present researchers that use this medium with important limitations linked to external validity (Couper, 2000). The second major problem of Web research is low response rates and high dropouts compared to other modes of interviewing (Cook, Heath, & Thompson, 2000; Galesic, 2006).

Considering the latter issue Manfreda and Vehovar (2002) show using WebSM, a Web survey completed by 155 practitioners of online research, that the average response rate is around

·e-mail: cernat@essex.ac.uk.

40 percent for individual invitations and around 30 percent for general invitations while the mean dropout is 16 percent. When asked about the main reasons for dropout, as perceived by the researchers, the main factors were: the length of questionnaires (a mean of 3.15 on a scale from 'no problem at all' coded 1; to 'large problem' coded 5), respondents lost interest (mean 2.81), respondents got annoyed (2.29 mean), slow downloading (2.11 mean) and difficult questions (2.06 mean) (Manfreda & Vehovar, 2002). Similarly, Fan and Yan (2010) found that variables linked to burden, such as announced length of the questionnaire, and variables linked to interest, like topic of the survey or question display, influence both response rates and dropouts. Thus, we can conceptualize two main factors that have an influence on the percentage of responses to the invitations and on the completion rates: motivation and burden.

While in face to face or telephone surveys the interviewer has the opportunity to keep the respondent engaged and gives any extra information needed this is harder to achieve in the online medium. Thus, motivation is an essential issue with characteristics particular to the online medium that researchers must solve using innovative methods. One traditional approach is the use of financial incentives. These can come in different forms, from direct monetary compensation to vouchers or lotteries. This kind of motivational device has an impact both on response and on completion rates. Using 26 studies in a meta-analysis Göritz (2006) highlighted that financial incentives increase the odds ratio of finalizing the

online questionnaire by 1.27 (95% CI: 1.12 – 1.44). An alternative incentive, which is specific to online surveys and that influences mainly the completion rate, is the questionnaire design. It is expected that by improving the respondents' experience we will also increase the probability of finalizing the questionnaire. The online medium offers unlimited modes of design improvements, from fonts and background changes to movies, with low costs of implementation. Of particular interest to us are pictures as they are easy to implement and they pose relatively minor technical problems.

Another important aspect that influences both the response and completion rates is the burden. One such example, the announced length of the questionnaire, conceptualized here as the expected burden, has a significant effect on participation (Crawford, Couper, & Lamias, 2001). Other factors linked to burden, such as the time spent on the questionnaire or the presence of difficult questions, have a significant influence on dropout as well (Galesic, 2006; Kivu, 2010; Manfreda & Vehovar, 2002). Thus, in order to increase response and completion rates we need to decrease the burden on the respondents by asking fewer questions, e.g., using single items or smaller scales as proxy measures for the scales traditionally used in other modes, or by using alternative measurement instruments that consume less time and are easier to complete. One such instrument is Best Worst Scaling (BWS; Finn & Louviere, 1992). This is a type of discrete conjoint analysis that presents the respondents with subsets of

affirmations from the scale for which they are asked to choose their preferred and least preferred option. It has been shown that this method significantly decreases the time needed to complete the questionnaire and the burden on respondents while collecting valid and reliable data (e.g., Lee, Soutar, & Louviere, 2008).

So, while we know that motivation and burden are essential for response and completion rates in the online medium and that there are some solutions to these problems, such as design improvements for motivation or BWS for reducing burden, we have limited information about the effects of using them together. As a result, the present paper proposes an online experiment where two randomised groups, a control and an experimental group, will complete a questionnaire measuring values and lifestyles using BWS. While the control group will complete the questionnaire with a simple design, no images, the experimental group will complete a form that includes an associated image for each affirmation. We also aim to explain the observed difference between the two groups by the degree to which the picture corresponded to the associated affirmation in the eyes of the respondents.

In the next section we will present some of the findings in the literature regarding the effect of online questionnaire design on measurement. Then, we will present some of the research done about the impact of burden on participation and completion of online surveys and we show how BWS can decrease the respondents' burden while collecting quality data. This section will be followed by

descriptions of the experiment, the scale used and the samples. Finally, we will present the results and the conclusions.

Theory and hypotheses

Impact of design on and measurement

While Web research brings many advantages in the form of a large array of design opportunities, this also has effects on data quality and measurement error. Small differences in design, such as the number of items or columns presented on a page or the type of boxes used, have important effects on the degree of item missing and measurement (Dillman & Smyth, 2007). For example, Mahon-Haft and Dillman (2010) hypothesise that general appeal of the questionnaire has an overall effect on data quality. A randomised experiment that compares a visual pleasing design and a 'visceral' one has shown mixed results. On the one hand the questionnaire that was more pleasing aesthetically had more favourable options, more characters and more thorough responses in open ended questions, tended to have fewer selections at the extremes of the scales (primacy) and had shorter response times. On the other hand, the displeasing aesthetic design did not have a significant effect regarding response rates, early terminations or perceived burden. A different research showed that using a 'fancy' design, that includes a sophisticated page layout with different colours to guide the respondent, leads to lower response rates, more dropouts and a smaller probability of fully completing the

questionnaire as opposed to the plain design (Dillman, Tortora, Conradt, & Bowker, 1998).

Pictures in online questionnaires

The use of pictures enters in this larger scheme of Web design. They are some of the most attractive design features for the online researchers due to the technical ease of implementation and relatively fast download times. But while this may be true, and images continue to be extensively used in the online medium, there is still inconclusive evidence regarding their impact on motivation, dropouts or measurement errors. Although theoretically it is expected that improving the design of the questionnaire through the use of pictures will increase interest in the survey and, thus, reduce the number of breakoffs this idea still lacks substantial empirical evidence.

In one of the few studies that tackled these issues Couper et al. (2004) did not find any relationship between the presence of the pictures in the questionnaire and satisfaction or dropouts. In contrast, using 9762 surveys applied between 2007 and 2010 by Ipsos Interactive Services in the U.S. and Canada, Kivu (2010) found that the presence of pictures had a negative influence on dropouts in short questionnaires and a positive effect in long ones. Toepoel and Couper (2010) also found limited proof that the inclusion of images leads to a more positive evaluation of the questionnaire. In contrast, the use of visuals depicting a male or female as an interface in Web surveys does not have a significant influence on social desirability, impression management or

sensitive topics, but it has an effect on gender attitudes. Thus, when presented with male pictures the mean of the pro-feminist attitudes was significantly lower than when respondents were presented with female ones (Sproull, Subramani, Kiesler, Walker, & Waters, 1996).

As highlighted by Couper et al. (2004) images can have three roles in Web survey research. The first one is an 'essential role' where viewing and interpreting the picture is part of the task, such as brand recognition. In the second situation, images supplement the information given by the text and can act as a clarification and/or a motivational device. The last category includes images that are 'incidental' and act as background and part of the overall design of the questionnaire. Although all three situations are relatively common in Web research, it is the second one that is the most problematic as the image and text may interact in unexpected ways to create measurement bias. While it may be clear to the researcher what is the role of the picture in the questionnaire it may be less so to the respondent that must assume a relationship in the absence of any explicit reference to the visual presence in most of the cases.

When the picture is perceived as supplemental to the text by the respondent, two types of interactions are possible. In the first case, when the picture and the text are congruent, the expected effect is that of accentuation of the message. On the other hand, if the text and images are incongruent, an interference effect is most likely (Couper, 2001, 4). Thus, either a process of assimilation (the meaning of the contextual artefact is embedded

in the text) or one of contrast (in which case the conceptualization of the respondent moves away from the contextual factor) are possible (Couper, Conrad, & Tourangeau, 2007).

The assimilation process has been highlighted in the online environment by Couper et al. (2004). They use a series of experiments to see how pictures interact with questions in a Web questionnaire when respondents must report the frequency of a number of activities. It has been found that the presence of the pictures modifies the responses given. The authors' hypothesis is that the pictures not only activate certain memories linked with the questions but also help construct the way respondents understood the questions as a whole. Couper et al. (2007) also present supporting evidence regarding the contrast effect when text and pictures interact. This was highlighted by using two different images when respondents were asked about their health in an experimental design. Thus, a group of respondents was shown a healthy person jogging, while the other group was shown a person in a hospital bed. Significant difference between the two groups was found, with those seeing the picture of the sick woman reporting higher mean health. This shows that respondents tend to choose their health level by contrasting their state with that presented by the picture.

Current evidence also underlines that while respondents tend to use pictures to complement the information given by the text, they also give more weight to the text when the picture contradicts the written information. In addition, it is possible to minimize the effect of contradicting verbal and visual cues by using a more precise language (Toepoel & Couper, 2010).

Burden and BWS

Burden is another aspect that has an important impact on response and dropouts rates. Here an important distinction can be made between expected burden and experienced burden. Through the former we refer to what the respondent expects to do in order to complete the questionnaire and how difficult they perceive those tasks. This type of burden influences especially response rates. For example, the announced duration of the questionnaire is an important predictor for response rates (e.g., Crawford, Couper, & Lamias, 2001; Galesic, 2006). In addition to the expected burden, the respondent takes into consideration the experienced burden, this latter type affecting mostly the dropouts. Thus, the time spent completing the questionnaire was found to influence the dropout rates (Kivu, 2010; Manfreda & Vehovar, 2002). Similarly, the number of the questions and how difficult they are to complete are indicators of burden that have statistically significant influences on dropouts (Galesic, 2006; Manfreda & Vehovar, 2002). In this context BWS can have a significant effect both on the expected burden, the researcher announcing a shorter questionnaire, and on the experienced burden by the shorter time of completion and by the fact that the task of completing BWS are simpler than in traditional scales (Lee et al., 2008).

Best-worst scaling

Best worst scaling, or Maximum Difference Scaling, is a type of conjoint analysis that uses choice data gathered by using experiments. Through choice data we understand the selection of one option as 'the best' or 'preferred' one (Louviere, 1988, 94). The method through which the respondent is asked to choose the best and the worst option in an experiment that presents a subset of the scale affirmations was first proposed by Finn and Louviere (1992). Through this selection process the BWS method maximizes the utility difference between the best and the worst choices and gives a score of utility for each item.

One of the important advantages of BWS is that it can use a large number of items with minimal impact on the respondent's effort. This is mainly due to the fact that by presenting a subset of items the respondent reduces the total number of combinations they need to analyse. Thus, from a subset of four items, A, B, C, D, when the respondent says that he prefers A the most and D the least he is giving information about 5 of the 6 comparisons possible. Through these two choices we know that A is more preferable than B, C and D, and that D is less preferable than A, B and C. The only relationship that we do not have any information about is the comparison of B with C.

The decrease in burden is most evident in the time needed to complete the equivalent of a traditional sociological scale. When the Schwartz Values Scale was applied both with BWS and with ten point questions the mean response duration was 12 minutes smaller in the former (Lee et al., 2008).

The BWS has a set of other advantages as well. Firstly, the burden for the respondents is decreased because the choices mimic closer real life situations (Louviere, 1988, 102) and it is easier for people to choose at the extremes of a scale (Marley & Louviere, 2005, 464). Also, reliability and cross-cultural comparability are bigger as selecting the best and the worst options have smaller interpretation errors than other scales (Lee et al., 2008). Thirdly, due to the fact that it is an indirect method, BWS avoids response styles like that of social desirability or primacy (Tavares, Cardoso, & Dias, 2010). Finally, the BWS offers the possibilities to use a wider variety of statistical measures as it does not assume ordinal measures, like some other types of scales regularly used (Lee et al., 2008).

Hypotheses

Considering the current results and practice, we believe that future research in the online medium will see an increase in usage of both pictures and BWS. In such an environment we are interested in how the two interact and if using them together changes in a significant way the final measurements. Thus, we aim to see if conditional on the usage of BWS the implementation of pictures in web surveys will impact measurement. Current practice shows that pictures are used in online research as a means to keep the respondents interested and motivated and, as a result, decreases dropouts and poor quality data. On the other hand, current scientific results show that images have less of an impact on dropouts and

more on the measurement, resulting in biased results. Thus our first hypothesis is:

H$_1$: The design with pictures does significantly influence the measurement (item utility) without a clear direction

In order to test this hypothesis we will compare the sample utility for each item of our scale in the control group, design without pictures, and the experimental group, design with pictures. Due to randomization any differences found between the two groups will be due to the presence of the pictures.

Our second research question tries to explain the measurement differences between the two groups, if these indeed exist. Using the assimilation and contrast perspectives (Couper et al., 2007, 2004) we expect to see a mean utility larger for items that have concordant pictures associated than those that have discordant images. In order to answer this question we have asked respondents in the experimental group to rate the degree to which the picture and the affirmation

were concordant at the end of the questionnaire.

H$_2$: Concordance between image and text increases mean utility while discordance decreases it

Data and methods

The data collection took place in June 2011 with an online convenience sample of 215 respondents. At the starting of the questionnaire the respondents were asked about their age (four categories) and sex. The membership to the two groups, the experimental and the control, was randomised inside the age and sex categories. This was done in order to guarantee equivalent samples on these two variables. Approximately half of the sample, 109 respondents, were randomly chosen to answer the questionnaire that included a picture associated with each affirmation (experimental group) while the rest of the sample, 106 respondents, completed the questionnaire that included only text (control group).

Figure 1. *Example of the pictures used in the experimental design*

In order to test our hypotheses we have used a subset of 16 items out of the Values, Attitudes and Lifestyle (VALS) scale. The scale was first proposed by Arnold Mitchell at SRI International and included approximately 34 questions that produced nine distinct lifestyles. The typology went to become one of the most influential lifestyles typology used in marketing to date (Kahle, Beatty, & Homer, 1986, 405–406). Later, the methodology was updated to VALS 2 which included new questions and segmentation scheme (Winters, 1989). Our preference for a selection of these items is its widespread use in marketing research, where also BWS and Web research are very popular. Also, the application on topics like values and lifestyle will be useful to methodological research as already showed in the application of Schwartz Values Scale (Lee et al., 2008).

In order to apply BWS respondents were shown 10 combinations of 4 affirmations and asked to select the affirmation that they prefer the most and the one that they prefer the least. Each affirmation was used 500 times and combined approximately 100 with each of the other items. The position of the affirmation in the group of four was completely randomised. The design was identical for the two groups, the only difference being the presence of images in the experimental group (e.g., **Figure 1**). Respondents in the latter group were also asked at the end of the questionnaire to rate how close was the fit between the affirmation and the associated picture for each item evaluated.

Table 1. *Socio-demographic distribution of samples with Chi-square tests*

		Experimental group	Control group	Total
Sex χ^2 **(df: 1) 0.118, p. 0.731**	**Male**	50.5%	48.1%	49.3%
	Female	49.5%	51.9%	50.7%
Age χ^2 **(df: 3) 1.298, p. 0.730**	**Under 18**	9.2%	9.4%	9.3%
	Between 18 and 34	51.4%	47.2%	49.3%
	Between 35 and 54	36.7%	37.7%	37.2%
	Older than 55	2.8%	5.7%	4.2%
Education χ^2 **(df: 3) 3.879, p. 0.275**	**Primary school**	14.7%	10.4%	12.6%
	High school	28.4%	34.0%	31.2%
	Bachelor	45.0%	36.8%	40.9%
	Graduate	11.9%	18.9%	15.3%
Income χ^2 **(df: 4) 4.033, p. 0.402**	**Bellow 500 RON**	33.0%	31.1%	32.1%
	Between 501 and 1500 RON	44.0%	37.7%	40.9%
	Between 1501 and 2500 RON	16.5%	23.6%	20.0%
	Between 2501 and 4500 RON	6.4%	5.7%	6.0%
	More than 4500 RON	.0%	1.9%	.9%
Total sample size		109	106	215

The rate of participation (The American Association for Public Opinion Research, 2011) was 59.5 percent: out of 600 invitations sent 357 replied. Out of these 9 ended the questionnaire completion prematurely while 134 could not participate because the quotas associated with their sex and age groups were full. These rather high response rates and low number of dropouts are mainly due to the use of incentives and of a professional online research firm to conduct the survey. **Table 1** presents the distribution of the two samples on sex, age, education and income and the chi-square tests for the two groups.

The analysis of the BWS data is based on the Hierarchical Bayes which, in its turn, is rooted in the multinomial logistic regression (Orme, 2000). The advantage of this method is that it can estimate individual scores even though the respondents did not answer all the questions. It is also robust and gives results at least as good as ten point scales (Moore, Jason Gray-Lee, & Louviere, 1998; Orme, 2000, 3).

Results

Testing H_1

As mentioned before we used the Hierarchical Bayesian estimates to calculate individual utility based on the group membership. The models for the two groups show some difference in the model fit (Root Likelihood), 0.637 for the control group as opposed to 0.598 for the experimental group. Due to the fact that the only difference between the two groups is the presence of the images we can deduce that these

have reduced the model fit for the regression model somewhat. Also, out of the nine dropouts three were from the control group and six were from the experimental group, although the difference between the two is not statistically significant.

The item mean utility for each group is presented in **Table 2** along with significance tests for the difference between them. We can observe that six out of 16 questions have significant different scores between the two groups although the sample size is relatively small. We also observe that in half of these the mean preference is bigger in the control group (**work, skills, compliance**) while for the other half the mean scores are higher for the experimental group (**socialize, tradition, universe**). While the results show that images change the measurement for an important part of the affirmations no overall trend is present regarding the direction of the influence.

Testing H_2

In the second part of the analysis we try to explain the differences between the groups by using the perception of the respondents in regard to the fit between each affirmation and the associated image. In order to do this, the respondents from the experimental group answered for each affirmation a four category question about the link between text and image: 'How appropriate is the picture to the associated image in your opinion?' The respondents were rescaled in two groups. Those who chose 'very inappropriate' or 'more inappropriate

Table 2. *Item utility for experimental and control groups*

	Experimental group	Control group
I do everything I can to make sure my **family** is well	13.52	13.13
I prefer to **socialize** at home than go out	**2.15***	**1.08**
I don't like the **rules**, prohibitions and situations where I'm told what to do	4.66	5.57
To succeed in life you have to **work** hard	**9.29**	**12.23**
Financial **security** is very important to me	11.45	12.27
I always try to find new ways of **leisure**	6.67	6
Tradition is very important to me	**5.05**	**3.51**
I like **unusual** people and things	4.19	4.65
I like a group to take into **account** my opinion	7.33	7.24
I follow the latest trends and I'm passionate about **fashion**	1.85	1.64
I have more **skills** than most people	**1.4**	**2.65**
Community and social **compliance** is very important	**6.55**	**7.9**
Not risking is better because it is possible to lose also the things you already have	3.98	3.17
I'm uncomfortable in crowds and bustle of **city life**	2.96	2.72
It is important not to forget to enjoy life's little **pleasures**	10.83	9.92
I would like to understand better how the **universe**/things work	**8.12**	**6.3**

* bold text and gray highlighting indicate significant difference in t-test at 0.05 level

than appropriate' were grouped together, while those who answered 'more appropriate than appropriate' or 'very appropriate' formed the second group. A fifth category was 'I don't know/I can't decide' and was selected by a minority of the respondents (mean selection bellow 10 percent). As suggested by Couper (2001) we expect that the mean utility will be higher for individuals that perceived as appropriate the association between picture and text and lower for those that saw the link as inappropriate. Also, we expect that the mean utility of the control group to be placed between the appropriate and inappropriate groups.

Comparing the mean item utility in the two experimental groups, inappropriate and appropriate, we observe that in 14 out of the 16 affirmations the mean utility is higher in the appropriate group than in the inappropriate one. Eight of these differences are statistically significant despite de small sample[1]. **Table 3** adds to this comparison the item mean utility from the control group. The most often observed pattern in this case, nine times out of 16, is a trend where the appropriate group has the largest mean utility while the inappropriate has the smallest one, the control group being somewhere between these two. In four cases (**socialize**, **tradition**, **not smoking**, **universe**) all the differences between the three groups are statistically significant while in another four at least one difference is statistically

Table 3. Item utility for inappropriate, appropriate and control groups

	Inappropriate	Control group	Appropriate
I do everything I can to make sure my **family** is well	**8,06***	13,13	14
I prefer to **socialize** at home than go out	**0,99**	**1,08**	**2,66**
I don't like the **rules**, prohibitions and situations where I'm told what to do	3,34	5,57	5,3
To succeed in life you have to **work** hard	8,45	**12,23**	**9,65**
Financial **security** is very important to me	11,63	12,27	11,29
I always try to find new ways of **leisure**	6,54	6	6,52
Tradition is very important to me	**2,18**	**3,51**	**5,84**
I like **unusual** people and things	**2,92**	**4,65**	5,31
I like a group to take into **account** my opinion	**4,23**	**7,24**	8,25
I follow the latest trends and I'm passionate about **fashion**	1,3	1,64	2,23
I have more **skills** than most people	0,93	2,65	2,04
Community and social **compliance** is very important	5,27	7,9	7,07
Not risking is better because it is possible to lose also the things you already have	**1,79**	**3,17**	**5**
I'm uncomfortable in crowds and bustle of **city life**	**1,33**	**2,72**	3,65
It is important not to forget to enjoy life's little **pleasures**	10,86	9,92	10,99
I would like to understand better how the **universe**/things work	**5,87**	**6,3**	**9,43**

* bold text and gray highlighting indicate significant difference in t-test at 0.05 level

significant (**family**, **unusual**, **account**, **city life**). In conclusion, our hypothesis is only partially verified. Most of the time the appropriateness of the fit between the picture and the statement increases the mean utility while the inappropriateness decreases it, but only four times out of 16 was this entire pattern statistically significant.

To further understand the relationship between the utility and degree of fit of the picture to the affirmation, we have created indicators for the percentage of times each item was considered appropriate, inappropriate or when the respondent could not decide regarding the fit between the image and the text. **Table 4**

presents the correlation between these three indicators and the mean utility. As expected utility is highly correlated with the degree of appropriateness (r: 0.662) and inappropriateness (r: -0.847). It appears that the degree of inappropriateness has a bigger impact on the utility than the degree of appropriateness. Also, as expected, a high negative correlation can be observed between appropriateness and inappropriateness (r: -0.909).

We also hypothesized an extended theoretical model. We expected that the effect of the concordance on the mean utility was caused by the extra burden due to the conflicting message of image and text. This would mean

Table 4. Pearson correlation between mean utility and appropriateness indicators

	Utility			
Appropriate	r: 0.662, p: 0.005*	**Appropriate**		
Inappropriate	r: - 0.847, p: 0.000	r: -0.909, p: 0.000	**Inappropriate**	
Undecided	r: 0.280, p: 0.293	r: -0.393, p: 0.132	r: -0.026, p: 0.923	**Undecided**

* 16 df. for all significance tests

that as a result of the extra task needed to be completed by people in order to solve the discordance between image and text they would tend to avoid these situations. We tested this hypothesis by explaining the mean appropriateness through the time needed for the completion of the questionnaire - a proxy for burden - education, sex and age. The OLS regression did not show any statistically significant relationship, putting in doubt our theoretical expectations.

Overall our experiment partially supports our second hypothesis. Nine out of sixteen times a pattern of inappropriate < control group < appropriate regarding mean utility appeared. Eight out of these have at least one of the relationships statistically significant. Also, Tabel 4 presented the correlation pattern between the utility and appropriateness/ inappropriateness. Once again strong relationships are evident with the negative relationship between utility and inappropriateness being somewhat higher. A link between burden, time to complete the questionnaire, and overall appropriateness for all the questions did not show a statistically significant effect.

Conclusions

Although pictures are constantly used in online surveys to attract respondents and keep them interested, current research has showed that their impact is less on interest and dropouts and more on measurement. Our results support this. No significant differences were found in dropouts between the questionnaire with and without pictures. On the other hand the pictures had an important effect on measurement, influencing the mean of the utility function of eight out of 16 items. Couper's (2001) theoretical assumption of an accentuation effect when text and visual input are in concordance and a contrast one when they are in discordance has received partial support from our experiment. In nine out of 16 items we have seen a pattern of inappropriate < control group < appropriate in the mean utility. Eight of these situations have at least one relationship statistically significant although only four have the entire relationship significantly different from zero. Thus, we reiterate a recommendation that is now becoming common in the methodological research of the online medium: avoid pictures if possible. If researchers decide to use visual cues then the questionnaire should be thoroughly pretested. When using BWS we have shown that of outmost importance is the similarity

in the level of concordance between pictures and text. Differences in this aspect will have important effects on measurement.

Our analysis has covered only a small aspect of the design issue in the online environment. Some limitations of our research have an impact on its external validity. Firstly, the sample and the language used were Romanian, different patterns may appear in other cultural contexts. Furthermore, the incentives and sampling may lead to different results. Additionally, the samples are relatively small and, as a result, some of the non-significant effects may be due to this cause. Another limitation is linked to the absence of information regarding the respondent's interest, satisfaction and perceived burden. This would have helped us to see to what degree the pictures improved the respondents' experience. Finally, we do not know if similar effects are presents for Likert or ten point scales. Future research on the impact of design on the quality of online data should focus more on theoretical models that can be tested in order to understand more thoroughly the causal mechanisms involved.

Note

[1] The table with this information can be accessed by contacting one of the authors.

References

Cook, C., Heath, F., & Thompson, R. L. (2000). A Meta-Analysis of Response Rates in Web-or Internet-Based Surveys. *Educational and Psychological Measurement*, *60*(6), 821–836. doi:10.1177/00131640021970934

Couper, M. P. (2000). Review: Web surveys: A review of issues and approaches. *The Public Opinion Quarterly*, *64*(4), 464-494.

Couper, M. P. (2001). Web surveys: The questionnaire design challenge. *Proceedings of the 53rd session of the ISI*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.99.3436&rep=rep1&type=pdf

Couper, M. P., Conrad, F. G., & Tourangeau, R. (2007). Visual Context Effects in Web Surveys. *Public Opinion Quarterly*, *71*(4), 623–634. doi:10.1093/poq/nfm044

Couper, M. P., Tourangeau, R., & Kenyon, K. (2004). Picture This!: Exploring Visual Effects in Web Surveys. *Public Opinion Quarterly*, *68*(2), 255-266. doi:10.1093/poq/nfh013

Crawford, S. D., Couper, M. P., & Lamias, M. J. (2001). Web Surveys: Perceptions of Burden. *Social Science Computer Review*, *19*(2), 146–162. doi:10.1177/089443930101900202

Dillman, D. A., & Smyth, J. D. (2007). Design Effects in the Transition to Web-Based Surveys. *American Journal of Preventive Medicine*, *32*(5), S90-S96. doi:10.1016/j.amepre.2007.03.008

Dillman, D. A., Tortora, R. D., Conradt, J., & Bowker, D. (1998). Influence of Plain Vs. Fancy Design on Response Rates for Web Surveys. *Proceedings of the Survey*

Research Methods Section, American Statistical Association, 1998.

Fan, W., & Yan, Z. (2010). Factors Affecting Response Rates of the Web Survey: A Systematic Review. *Computers in Human Behavior*, *26*(2), 132–139. doi:10.1016/j.chb.2009.10.015

Finn, A., & Louviere, J. J. (1992). Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety. *Journal of Public Policy & Marketing*, *11*(2), 12-25.

Galesic, M. (2006). Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey. *Journal of Official Statistics*, *22*(2), 313.

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

Kahle, L. R., Beatty, S. E., & Homer, P. (1986). Alternative measurement approaches to consumer values: the list of values (LOV) and values and life style (VALS). *Journal of consumer research*, *13*(3), 405-409.

Kivu, M. (2010). *Long Questionnaires: Impact on Abandon Rate* (1-5). Bucharest: IPSOS - Romania.

Lee, J. A., Soutar, G., & Louviere, J. (2008). The Best–Worst Scaling Approach: An Alternative to Schwartz's Values Survey. *Journal of Personality Assessment*, *90*(4), 335-347. doi:10.1080/00223890802107925

Louviere, J. J. (1988). Conjoint Analysis Modelling of Stated Preferences: A Review of Theory, Methods, Recent Developments and External Validity. *Journal of Transport Economics and Policy*, 93-119.

Mahon-Haft, T. A., & Dillman, D. A. (2010). Does Visual Appeal Matter? Effects of Web Survey Aesthetics on Survey Quality. *Survey Research Methods*, *4*, 43-59.

Manfreda, K., & Vehovar, V. (2002). Survey design features influencing response rates in web surveys. *The International Conference on Improving Surveys Proceedings*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.515&rep=rep1&type=pdf

Marley, A. A. J., & Louviere, J. J. (2005). Some Probabilistic Models of Best, Worst, and Best–Worst Choices. *Journal of Mathematical Psychology*, *49*(6), 464-480. doi:10.1016/j.jmp.2005.05.003

Moore, W. L., Jason Gray-Lee, & Louviere, J. J. (1998). A Cross-Validity Comparison of Conjoint Analysis and Choice Models at Different Levels of Aggregation. *Marketing Letters*, *9*(2), 195-207.

Orme, B. (2000). Hierarchical Bayes: Why all the Attention? *Sawtooth Software Research Paper Series*, 1-7.

Sproull, L., Subramani, M., Kiesler, S., Walker, J. H., & Waters, K. (1996). When the Interface is a Face. *Human-Computer Interaction*, *11*(2), 97-124.

Tavares, S., Cardoso, M., & Dias, J. G. (2010). The Heterogeneous Best-Worst Choice Method in Market Research. *International Journal of Market Research*, *52*(4), 533. doi:10.2501/S1470785309201430

The American Association for Public Opinion Research. (2011). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 7th edition*.

Toepoel, V., & Couper, M. P. (2010). Can Verbal Instructions Counteract Visual

Context Effects in Web Surveys? *Public Opinion Quarterly*, *75*(1), 1-18. doi:10.1093/poq/nfq044

Winters, L. (1989). SRI Announces VALS 2. *Marketing Research*, *1*(2), 67.