Psicológica (2010), 31, 491-508.

A comparison of Web-based and face-to-face Functional Measurement experiments

Frederik Van Acker*[1] & Peter Theuns[3]

Open Universiteit, The Netherlands[1]; Vrije Universiteit Brussel, België[2];

Information Integration Theory (IIT) is concerned with how people combine information into an overall judgment. A method is hereby presented to perform Functional Measurement (FM) experiments, the methodological counterpart of IIT, on the Web. In a comparison of Web-based FM experiments, face-to-face experiments, and computer-based experiments in the lab it is found that the computer-based method is less sensitive to experimental manipulations. However, different integration rules can be distinguished. The inability to monitor motivation in the unproctored setting can partly explain this effect. Consequently it is argued that Web-based experiments enable the researcher to test larger groups which enable more in-depth analysis of individual differences using single subjects analyses and clustering methods.

Everyday decisions are rarely based on a single piece of information: when deciding if some book is worth buying, several features are likely to be taken into account simultaneously: its price, the reputation of the author, its size, print, and maybe even the pictures that it contains. The decision on whether or not to buy the book will then be based on the integrality of the available information: somehow we integrate several single facts into an overall decision or judgment. This forms the basis for Information Integration Theory (IIT) (Anderson, 1981).

In marketing research as well as in more fundamental studies IIT has been a useful framework to explain how several pieces of information are internally combined. These studies usually require a rather cumbersome experimental procedure. In this paper we will describe a method to perform information integration experiments on the Internet and we discuss how

* Acknowledgments: The authors would like to thank the editor and two anonymous reviewers for their very helpful comments. Correspondence address: Frederik Van Acker. Open Universiteit Nederland - School of Psychology, PO Box 2960 - 6401 DL Heerlen, The Netherlands. Email: frederik.vanacker@ou.nl.

such method can be useful to marketers. The current method is compared empirically with the traditional way of testing participants individually in a controlled environment.

Ever since Web-based studies were introduced into the scientist's toolbox, researchers have been searching for mode differences, for example between Internet questionnaires and their paper-and-pencil counterpart. Findings suggest that Internet samples are demographically diverse, do not differ from nonusers on most variables (e.g. depression) and are sufficiently motivated (Gosling, Vazire, Srivastava, & John, 2004). The online population can thus be considered as a valid alternative to most offline samples. Moreover, presentation format (online versus paper) does not seem to affect validity and reliability (Pettit, 2002; Meyerson, & Tryon, 2003). Equivalence of Web-based experiments versus controlled lab experiments has received less attention. An extensive literature search revealed one published study comparing a Web-based versus class-based experiment, reporting higher variability and lower risk-aversion in the Web-based experiment (Shavit, Sonsino, & Benzion, 2001). The current study will compare IIT experiments in several data collection modes hereby focusing on essential aspects, such as sensitivity, participant motivation and linearity of the response scale.

Applications of IIT in market research. Studies on how people integrate information can be very useful from a theoretical as well as a practical point of view. For example, based on the finding that children do not merely add the utilities of two objects, but rather average them (Schlottman, 2000), it may be inadvisable for marketers to offer product combinations to children. While the cited study was intended as a means to gain theoretical insights into the information processing capabilities of children, several IIT studies reported in marketing journals have direct implications for the field and will be cited briefly. There is some evidence that consumers rather average than add the information of different product attributes whereas models on consumer decision making are predominantly additive (Troutman, & Shanteau, 1976). Similarly, multi-product bundles show subadditivity (that is the utility of their combination is perceived to be less than the sum of the consecutive utilities), especially when some poor quality object is combined with one of high quality (Gaeth, Levin, Chakraborty, & Levin, 1990).

The cited studies show that empirical testing of the integration rules for attributes or product information may be advantageous. With Functional Measurement, the active integration rule can be recognized both on individual and on aggregate level. Moreover, the subjective evaluation of each attribute and its levels can be determined and in the case of an averaging integration pattern weights for individual attributes can be estimated. Functional Measurement thus offers a set of tools deemed to be most useful in marketing research. The method may appear to be very similar to conjoint analysis, which is widely used in this field. In conjoint analysis the importance of product attributes can be estimated using logistic regression analysis on data from incomplete factorial designs. Interactions between attributes can be modeled as well with this technique, but there is some evidence that FM may have additional advantages in finding a decision rule compared to conjoint analysis: the most important advantage of Functional Measurement is that it has the necessary power to detect even small interactions (and thus presumed non-additivity) where conjoint analysis fails to do so (Shanteau, Pringle & Andrews, 2007). Some advocates of conjoint analysis consider the interactions found with Functional Measurement to be an artifact of the assumption that the rating scale used in FM experiments is interval level (Coxon, 2006). Therefore, the linearity of the response scale is an important aspect of FM that has been established through a rigorous experimental procedure (Anderson, 1982).

Possible issues in online FM studies. In Functional Measurement, it is critical that a given rating is a linear function of the internally represented response. Failing to elicit linear responses will result in the inability to find the expected integration model, even though participants would have integrated the provided information exactly according to one of the FM integration rules. Imprecision induced by respondents who do not employ the rating scale as a linear one, cannot be corrected for (Anderson, 1982). This is a very important issue in Functional Measurement that can be addressed with the procedures described below (Anderson, 1982).

In FM experiments the format of the rating scale that is used to express the result of the stimulus integration is crucial to obtain linear ratings. A graphical rating scale should be preferred for FM experiments (Anderson, 1982), however, 20-point category rating scales have been used frequently with success (e.g. Lampel & Anderson, 1968; Zhu & Anderson, 1991). In Web surveys, the use of graphical rating scales such as sliders is discouraged since they are suspected to cause drop-out and missing data as the time needed to complete the questions increases (Couper, Tourangeau, Conrad & Singer, 2006). Newer methods have been developed that impose fewer requirements on respondents' computers (Reips & Funke, 2008), however, the impact of this technology on response behavior has to our knowledge not been studied yet. In the current study a 20-point rating scale will therefore be used. The applicability for this kind of rating scale has been validated in previous research (Hofmans, Theuns & Mairesse, 2008). The software accompanying this paper can however be used to create FM experiments with graphical rating scales as well as category rating scales.

Category rating scales in Web surveys usually contain about five to eleven response categories. In FM experiments rating scales count up to 20 possible responses. This large number of response options compared to usual surveys can be justified since Anderson (1982) notes that distribution effects need to be avoided. One type of distribution effects can be influenced by using a large number of response options, as it was shown empirically that the tendency to use all response categories equally often decreases as the number of response categories increases (Parducci, 1982), which justifies the use of up to 20 scale points in FM experiments. Other distribution effects should be taken care of as well: when using rating scales, end effects, such as floor or ceiling effects, may occur (Anderson, 1982). One way to avoid such effects is to train respondents in using the scale with a series of practice trials in which the response scale is calibrated. The main purpose of this calibration phase is to instruct respondents to use the entire scale and avoid overly using the end categories, as this would result in nonlinearity near the ends of a rating scale. Although this is hard to control for in an online setting, clearly written instructions could possibly suffice to motivate people to employ the scale properly. In face-to-face experiments it is not uncommon to give feedback on scale use during the calibration phase. Some experimenters instruct participants to change their answer. If, for example, a respondent rates a certain stimulus near the upper end of the rating scale, the experimenter can tell the participant to lower the rating when a stimulus will follow that is probably going to be rated higher (e.g. Schlottman, 2000). Although more difficult to program, this kind of interactions with respondents can be performed in a standardized way in computer administered experiments. This will however not be done in the current study, although research upon this may yield even more insight into the possibilities of online FM experiments.

Finally, another procedure to avoid floor and ceiling effects is to present stimuli during practice trials that are more extreme than the actual stimuli used in the experiment (e.g. as in Munos Sastre, Mullet & Sorum, 2000). A possible drawback of the method is that due to this manipulation several categories near the extremes are likely to remain unused.

To test whether the described response effects actually occur more in self-administered than in face-to-face experiments, the current study was thus designed to compare response modes. One way to test for the linear use of a response scale is actually finding the hypothesized integration rule (Hofmans & Theuns, 2008; Hofmans & Theuns, in press; Hofmans, Theuns

& Mairesse, 2007). Finding a specific integration rule simultaneously confirms the linear use of the response scale (Anderson, 1982).

Motivation and response behavior. In many experiments participants are first year university students who get some course credit for participation. Not all students are equally motivated to take part in lab experiments and so they may be careless in the way they perform in certain studies. Monitoring people individually can help to prevent such carelessness. When running batch experiments, the greatest concern is with increased response error resulting from lesser attention and dedication in the respondents. One way to overcome the problem of increased response error would be to test larger groups to reduce the standard error of estimation in the subsequent statistical analysis.

Low motivation may thus yield less valid data, for example when participants do not conscientiously follow instructions. Low motivation may also lead to incomplete data when participants are self-selected (e.g. by clicking on a banner on some Website) or in other situations where they are in no way obliged to take part in the experiment. Basing on results from nonresponse in survey research (Dillman, 2007), decisions on (non) participation in experiments may well be seen as a kind of social exchange behavior. In most situations participants are free to leave a survey or experiment at will. Social exchange theory comprises three important determinants of behavior: cost, reward and trust (Homans, 1958). Cost of taking part in FM studies should thus be minimized, especially in a Web-based context. In general, research on nonresponse has yielded a number of strategies to reduce the cost and to maximize the social reward and trust in online research, which can be translated to the context of online FM experiments.

As proposed by Anderson, Functional Measurement based experiments are usually performed in a face-to-face situation, because running batch experiments could possibly yield low-grade data (Anderson, 1982). However, we do not know of any studies investigating whether batch experiments would actually produce unreliable or invalid data. The current paper describes a method to run Internet-based Functional Measurement experiments using a software: OSuCre. Several exploratory analyses are performed to compare the data that were obtained in different settings. The main focus will be on the sensitivity of the data, the linearity of the response scale and proxies of motivation, such as response time.

METHOD

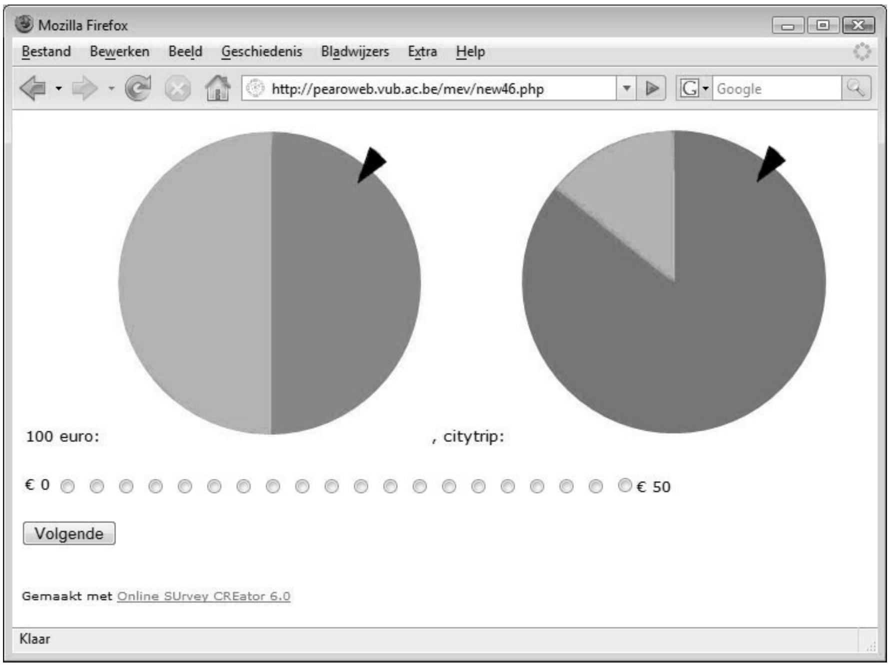
A randomized experiment was designed to test for equivalence of results from 3 kinds of experiments, which represent the three conditions of this study: face-to-face, computerized lab and Web experiment.

Participants and procedure. Seventy seven participants were randomly assigned to one of these three conditions. All of them were students who received some partial course credit for performing in one or more experiments. The average age of participants was 19.39 (*SD* = 1.98) years and 71% were women. All 77 students were invited to come to the lab at the psychology department. Once they arrived at the lab they were either assigned to the computerized lab, the face-to-face or the online condition. Students in the online condition were sent home with a note containing a hyperlink to the experiment, the other students took the experiment on the spot. After this they were asked to take part in a series of experiments, among which the Functional Measurement experiment described below. After screening the data, the ratings of four (two in the Web-based condition and two in the computerized lab condition) participants were found to have very low variances, indicating that they had been showing no or little variation in their responses. These participants were omitted from the analysis and thus 73 valid cases were used.

Design. During the experiment, which was analogous to the one by Schlottman (2000), participants were required to indicate the amount of money (between €0 and €50) they would spend on a certain gamble. Participants were shown a roulette-type spinner with two colors: red and blue as shown in Figure 3.3. The proportion red/blue was either 3/4, 1/2 or 1/8. If the arrow of the spinner would stop on red, the participant was told she/he was to win a prize. Next, participants were introduced to the prizes that could be won: either €100 in banknotes or a citytrip to a destination of choice. Gambles consisted of either a single spinner, where only the amount of money could be won, or two spinners, where both prizes could be won. The two spinner games were arranged in a 3×3 factorial design combining the three different probabilities (3/4, 1/2 or 1/8) of winning the €100 and the citytrip. Three ("uncombined") conditions were included in this design wherein the participant could only win the money, thus resulting in a total of twelve different stimuli. The single spinner games were necessary to test for the averaging integration rule. According to economic theory, additivity is a core feature of the standard approach to judgment (von Neumann, & Morgenstern, 1967). Violations of this independence axiom have however been shown with children (Schlottman, 2000) as well as with adults (Gaeth et al, 1990) and therefore a test for averaging is added to the experimental design.

Each stimulus was presented twice in order to assess stability of the responses on the one hand and to be able to perform statistical analysis at the individual level on the other.

Spinners and prizes were presented randomly on a paper sheet or in a Web browser window on a computer. The online and computer experiment were identical. In the computerized lab condition a Web server was installed on lab PCs to enable presentation of the Web-based instrument. The instrument was created using OSuCre which has the possibility to randomize pages. Moreover, the software Website (http://www.osucre.be) offers an instrument to construct the needed script for complete factorial designs, hereby facilitating the work needed to create a Functional Measurement experiment. As the researcher has to enter all details about the design, full factorial designs as well as nested designs can be created with the software. An example of how to create a FM experiment can be found on the Website by clicking the example link.



100 euro: , citytrip:

€ 0 ○ € 50

Gemaakt met Online SUrvey CREator 6.0

Figure 1. Example of a two spinner game as presented in both the online and computerized lab condition. For the face-to-face condition, spinner games were printed on a paper sheet.

In the online and computerized lab conditions, instructions and stimuli were presented onscreen. No additional instructions were given by the lab assistant. In the face-to-face condition instructions were read aloud by the experimenter and participants were asked if the instructions were clear. Stimuli were presented on a piece of paper together with a printed image of the same rating scale as seen in Figure 1. The three data collection modes were kept as similar as possible and therefore no feedback on the task was provided by the experimenter. In all three conditions, the first phase of the experiment consisted of six practice trials. These practice trials included the best and worst gambles, namely 1/8 chance of winning only €100 and a two times a 3/4 chance of winning both €100 and a citytrip. Participants were instructed to take these extremes into account during the actual experiment and to utilize the entire response scale. To reduce the chance that participants would overly bet the maximum amount (namely €50), they were instructed that only a limited amount of money could be spent on the entire set of gambles (€750 on 24 gambles).

RESULTS

The analysis will focus on several aspects of the data. First we assess whether visual as well as statistical analysis of the data on an aggregate level yield similar results across the three data collection modes. Next, the same analysis will be performed on individual basis. Several other aspects of data quality will be assessed, especially focusing on the use of the response scale. Finally, overall response times will be compared between the online and the computerized lab conditions.

Group analysis. The group analysis will be performed for each mode separately. To test for an additive integration rule, an ANOVA on the 3×3 design will be performed. To confirm the additive rule, both main effects need to be significant in absence of an interaction. As aforementioned, our design included a test for averaging. A separate ANOVA on the 3×4 (which includes the uncombined level of the second variable) design should not yield a significant interaction for the additive integration rule to hold true. If an interaction does appear from these data, this would be in support of an averaging integration rule. This same two-stage procedure will be pursuited to test integration rules on the group as well as individual level.

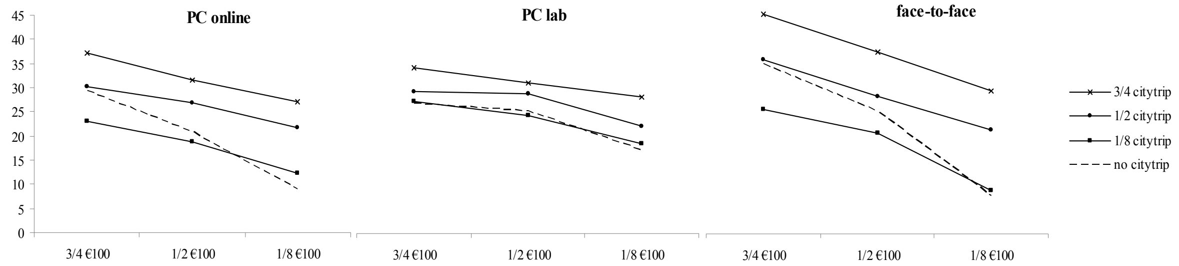


Figure 2. Results of the FM experiment in the online condition (left panel), the computerized lab condition (center panel) and the face-to-face condition (right panel), basing on group data. Mean amount of money spent plotted against spinner game 1 (chances 1/8, 1/2 and 3/4 to win €100) for each associated second spinner game (chances 0, 1/8 , 1/2 and 3/4 to win a citytrip).

Visual inspection of Figure 2 reveals a parallel pattern of lines. In both the face-to-face and the online version the uncombined level which shows an interaction. This specific pattern, which is indicative for an averaging integration rule, was supported statistically by the ANOVA results. All main effects were significant in all three data collection modes. None of the interactions were significant when the uncombined level was excluded. The test for averaging yielded a significant money × citytrip interaction for both the online ($F(3.94, 98.51) = 7.99, p < .001, \eta^2_p = 0.24$) and the face to face condition ($F(4.63, 101.88) = 11.50, p < .001, \eta^2_p = 0.34$) however not for the computerized lab condition ($F(3.42, 78.74) = 0.91, p = .49, \eta^2_p = 0.04$). Contrary to the two previous conditions, the graphical inspection of the factorial plot (see center panel of Figure 1) gives no clear support for an averaging integration rule. One could argue that, basing on the ANOVA results, the data are well represented by an additive integration rule, however, for this explanation to hold marginal means (or scale values) of the blank level should be lower than is currently the case.

Individual analysis. Group analyses cannot reflect individual differences in integration patterns. Therefore the data of each individual subject were analyzed consecutively both visually and statistically to investigate the integration rule on an individual basis. First all individual data underwent a visual inspection. For conciseness, an overview of the results is given as a percentage of integration rules that were found in each condition either visually or statistically.

Table 1. Percentage of different integration rules found across data collection modes, basing on visual analysis of individual data (left side) and ANOVA (right side).

	VISUAL INSPECTION			ANOVA		
RULE	**Face-to-face**	**PC online**	**PC lab**	**Face-to-face**	**PC online**	**PC lab**
Averaging	21.7%	25.0%	23.1%	21.7%	19.2%	26.9%
Adding	26.1%	25.0%	30.7%	69.6%	59.9%	38.5%
Multiplication	None	None	None	None	None	None
Other	52.2%	50.0%	46.2%	8.7%	20.9%	34.6%
N	26	24	23	26	24	23

As can be seen in Table 1, the distribution of integration rules tested visually across the three data collection modes was similar. None of the participants' plots showed any evidence of having used multiplication as integration rule.

The right part of Table 1 presents an overview of the statistical analysis of the individual data. As for the group data, individual ANOVA's are performed in two stages. First, the analysis of the 3×3 design should show two main effects and no interaction; secondly, the ANOVA including the blank level will discern between an adding and an averaging rule. The categorization in Table 1 is based on this two stage analysis. Since statistical power is lower for individual analysis than for group analysis, alpha was set to .1.

In contrast with the visual inspection, Table 1 shows a noteworthy difference in the distribution of integration rules across the data collection modes. About 35 to 43 percent of the cases could not be categorized in the online or computerized lab condition compared to only 9 percent in the face-to-face condition. In 82% of the cases in the online and computerized lab conditions the results were designated as "other" because one or two main effects were not significant.

Analysis of scale use and sensitivity. Several indicators of the quality of ratings were calculated and compared across the data collection modes. Here we will compare the range of ratings used and the variance of the ratings and their relationship with sensitivity. Finally we will analyze whether distribution effects occur.

For each participant the minimum and maximum rating were subtracted and the average range was compared across the three response modes. There were no significant differences in mean range across the three groups (*F*(2, 70) = 0.31, *p* = .74). As can be seen in Table 2, the variability in the range is smallest in the face-to-face condition, indicated by the smallest standard deviation.

The range of the obtained ratings is not necessarily indicative for the quality of the ratings. Range in combination with the variability of the ratings may provide a better insight into the use of the response scale. Table 2 also reports the mean variability in the participants ratings and these do not seem to differ significantly (*F*(2, 70) = 0.32, *p* = .73). Variability in the use of the scale is more homogeneous in the face-to-face condition.

Table 2. Mean range of scale points (on a 20-point scale) used and mean *SD* in use of scale points across participants. The *SD* was calculated for each participant as the standard deviation in the used scale points.

Mode	Range of individual scores overall		*SD* of individual scores overall	
	M_{range}	SD_{range}	M_{SD}	SD_{SD}
Face-to-face	15.65	2.38	4.75	0.94
PC online	15.23	2.86	4.78	1.25
PC lab	15.08	2.47	4.54	1.17

Participants in different contexts may be less sensitive to the experimental manipulations of the FM experiment itself. To test these differences in sensitivity we assessed the interaction of the mode factor with each of our expected effects, namely citytrip × mode (*F*(2.92, 102,36) = 3.44, *p* = .02, η^2_p = .09), € 100 × mode (*F*(2.61, 91.32) = 4.41, *p* = .01, η^2_p= .11) and finally citytrip × € 100 × mode (*F*(8.86, 310.18) = 1.93, *p* = .05, η^2_p = .05). As can be observed in the factorial plots of the group analysis (Figure 1) the size of the effects is the largest in the face-to-face condition. Additional pairwise comparisons revealed no significant differences between the online and the face-to-face mode. Differences between the online mode and the computerized lab were also non-significant. Mode differences did appear between the computerized lab and the face-to-face condition were all effects were significantly different (money × mode: F(1.29, 57.83) = 8.43, p < .001, η^2_p = .16; money × mode: F(1.52, 68.21) = 7.54, p < .001, η^2_p = .14 and money × city × mode: F(4.22, 190.08) = 3.00, p = .02, η^2_p = .06).

Additional analysis. A participant with a large range of scores and a high value for the standard deviation in the use of the scale may possibly have used a specific response strategy, resulting in invalid data. For example, limiting responses to the first and the last category, would yield such data pattern. Moreover, specific floor or ceiling effects may show in the data. To test for distribution effects, differences in response distributions are assessed. It can be seen in Figure 3 that the distributions of the given responses on the rating scale are quite similar across the three data collection modes, which was confirmed statistically (χ^2(38, N = 848) = 53.03, p = .05). No specific floor or ceiling effects are observed in the response patterns.

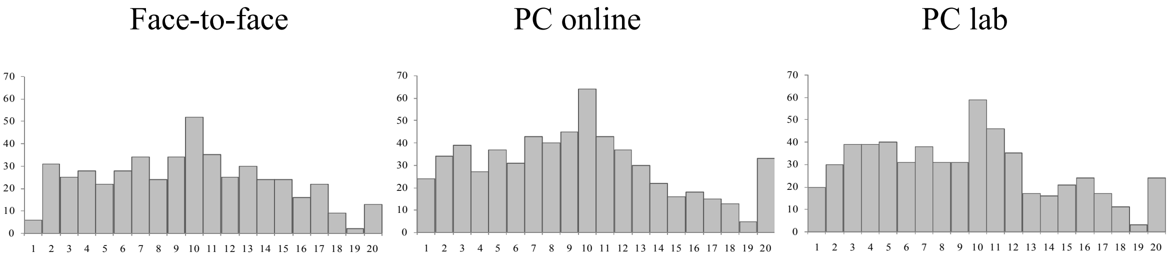


Figure 3. Distribution of the endorsed scale points for each data collection mode.

As each stimulus was presented twice in each condition, it is possible to assess the stability of the given responses and to see whether differences in stability occur between the conditions. This resulted in a 3×4×2×3 (€100 × citytrip × repetition × mode) analysis.

To assess the stability of the responses, the effect of repetitions must be considered. The largest effect of repetitions was the repetitions × citytrip × money × mode interaction (F(9.98, 349.45) = 0.45, p = .92) which is non-significant. An additional test to assess stability was performed using the data from the repetitions. For each participant the squared deviation between the first presentation of each stimulus and the repetition was calculated. An average squared deviation was then calculated across stimuli for each participant. These average deviations did not seem to differ between the online condition (*M* = 39.89, *SD* = 27.68), the computerized lab condition (*M* = 37.74, *SD* = 26.02) and the face-to-face condition (*M* = 31.08, *SD* = 16.31) (*F*(2, 70) = .95, *p* = .39).

Paradata. When collecting data by means of self-report instruments presented in some electronic format, it is possible to collect additional data on the response process. For this study, the time the experiment was started and the time the experiment was completed were registered for each person in the online and computerized lab condition. These paradata can thus provide an estimate of the time needed for each participant to complete the entire experiment. A final analysis was performed to assess differences in response time (in minutes) between the online condition (*M* = 8.37, *SD* = 1.70) and the computerized lab condition (*M* = 7.57, *SD* = 3.00). These differences were non-significant (*t*(41.56) = 1.22, *p* = .23).

Analysis of the response patterns and paradata (i.e. response times) indicated that the 2 participants in the computerized lab condition and another 2 participants in the online condition whose data were omitted from the analysis, performed very poorly. Not only did their responses show very little variation, also, response time for these participants was very fast, possibly indicative of a suboptimal processing of the stimuli.

DISCUSSION

The following paragraphs will first address methodological implications focusing on differences in data quality across data collection modes. Next we will briefly discuss the substantial implications of the current study.

Methodological implications. Contrary to the prediction of the independent integration of utility, the current experiment suggests that people employ an averaging rule when judging the worth of two gambles. It must be noted that this integration function is based on aggregated data and that in one out of three data collection modes no averaging rule was retrieved, neither from a visual inspection of the factorial plots, nor statistically using ANOVA. The factorial plot of the computerized lab condition (center panel of Figure 1) does not confirm an averaging integration rule, but gives no clear evidence for an additive rule either.

The sensitivity analysis indicated that the experimental manipulations produced the largest effects in the face-to-face condition and the smallest in the computerized lab condition. The different results may thus be due to a lack of sensitivity within this data collection mode. Post hoc analyses revealed significant differences in sensitivity between the computerized lab and the face-to-face mode, while both other pairwise comparisons remained non-significant. This finding suggests that participants in the computerized lab condition were less sensitive to the experimental manipulations. A possible explanation is that the unproctored setting reduced motivation of respondents. In the absence of a lab assistant, participants may have chosen to process the stimuli less thoroughly. A result in accordance with this finding is that both in the online and the computerized lab condition, participants were omitted from the dataset due to a lack of variability in their responses, which is indicative of a straight lining strategy (i.e. giving the same response irrespective of the intensity of the stimulus). However, after removing these data, the significantly different sensitivity remains. There must be some other factor explaining these differences as both the online and the computerized lab condition are unproctored. Participants in the computerized lab, however, were asked to complete the experiment at a specific time in a fixed location, while the online participants could complete the experiment at their own convenience. The latter may thus have chosen a moment to participate when their motivation was highest. Although untestable in the current study, this hypothesis would be in line with social exchange theory as the cost for these participants would be lower resulting in a better cost-reward ratio.

Carelessness can be assessed by looking at the stability of ratings. Each stimulus was presented twice and thus differences between the ratings for each presentation should be minimal. The stability analysis did not reveal any differences indicating that participants in each data collection mode show a similar conscientiousness when completing the experiment. Carelessness can thus not explain the differences in sensitivity. Whenever one has doubts concerning the quality of responses, paradata, such as response times can be used to objectively detect invalid data.

The main focus of this study was on testing whether the online method could provide us with data comparable to classical face-to-face FM experiments. An averaging integration rule was found with ANOVA as well as visually. Given that the same result was found with a face-to-face collection mode, this is a first indication of sufficient data quality in online FM experiments. Moreover, finding a specific integration rule simultaneously validates the linear use of the response scale (Anderson, 1982) which is a major concern in FM studies.

FM offers researchers the possibility to analyze data at the individual level. Aggregate data may conceal individual differences concerning the integration rule or subjective utility assigned to a set of stimuli. The results of our analyses show that, for those participants for which an integration rule could be identified, about fifty percent followed an averaging rule, while the other half showed the characteristics of an additive integration. An important difference appears when the individual data are analyzed statistically. The lack of sensitivity in the computerized lab condition results in insufficient power to detect significant (main) effects in about half of the cases compared to only about ten percent in the face-to-face condition. One could address this issue through additional presentations of stimuli to improve the statistical power of the design. However, as has been argued, this could lead to an increased proportion of dropout when participation is voluntary. Also, as dropout will probably occur proportionally more often in participants who lack motivation, the validity of the results may become insufficient. This may especially be the case when more complex designs are employed consisting of more than two attributes. An argument in favor of adding repetitions is that responses become more stable with an increasing number of repetitions. About half of the participants' integration patterns could not be categorized as either adding or averaging. In none of these cases any apparent pattern was discernable in the data, possibly caused by a lack of stable estimates. A counterargument for increasing the number of repetitions is that the factorial design of the experiments is more likely to become apparent to participants and that they may start behaving accordingly. The effect of the number of repetitions on data quality in FM experiments has to our knowledge not been tested yet. Therefore, in practice, one should consider the size of the design when setting up an FM study.

Substantial findings. A general conclusion concerning the integration rule on the group level would be that an averaging strategy is being used, rather than an additive one. These results have quite important practical implications as in the case of gambles, combining a high with a low value gamble could possibly result in a lower subjective utility than when only the high value gamble is presented. This effect was already observed in children (Schlottmann, 2000) as well as for the integration of product bundles (Gaeth et al, 1990) where adding a low to a high quality product decreased the net worth of the primary high value product. Knowing the exact integration rule may provide marketers with better insight into consumer decision processes than when solely relying on normative rules.

The results of the individual analysis also have clear implications for practitioners. If different individuals use different integration rules to evaluate subjective utility, then differential market strategies may be appropriate. In the study of Schlottmann (2000) it was clear that there was an age difference in the employed integration rule. Integration patterns may well vary with other social, demographic or psychological variables. The cumbersome experimental procedure described by Anderson (1982) requires participants to perform in experiments in a psychological lab. The current study however suggests that the online method can yield useful data, without the experimental control of a laboratory. This new method enables researchers to test a more heterogeneous sample as compared to face-to-face experiments where mostly students are participating. Finally, one can assess possible relationships with other variables and marketing strategies can then be segmented according to possible differences in integration functions. Hence, more insights into methods to cluster integration patterns might prove a useful addition to the field.

General conclusion: limitations and possibilities. The current study shows that online FM experiments can reveal several interesting facts about the integration of utility to marketers. The method may well be used for other purposes, depending on the substantive field of interest. Experiments on decision making, impression formation or consumer behavior lend themselves to the online method. Most such studies are performed with adults who nowadays have sufficient computer proficiency to perform in similar experiments. Children may lack these skills or may be too young to provide ratings on a standard rating scale and therefore researchers in the field of developmental psychology may not profit from the proposed procedure. Personal computers show large hardware variability and therefore psychophysical experiments, for example, may neither be suitable. To test the suitability of the online method one can easily pre test an experiment online using the proposed software, OSuCre. It must also be noted that the proposed method is less expensive and that data can be collected more rapidly.

A limitation of this study is that its design was less complex than many FM studies performed in the past (e.g. see Anderson, 1996). The complexity of the study will probably not impact on the scale usage. Complexity could however have an important impact on the experimental dropout when participants volunteer to take part in a study. Strategies from research on (online) survey dropout (e.g. incentives) can be taken into account when designing an online FM study.

REFERENCES

Anderson, N. H. (1981). *Foundations of Information Integration Theory*. London: Academic Press.

Anderson, N. H. (1982). *Methods of Information Integration Theory*. London: Academic Press.

Anderson, N. H. (1996). *A Functional Theory of Cognition*. New Jersey: Lawrence Erlbaum Associates.

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.

Dillman, D. A. (2007). Mail and Internet Surveys. The Tailored Design Method. New Jersey: John Wiley & Sons.

Gaeth, G. J., Levin, I. P., Chakraborty, G., & Levin, A. M. (1990). Consumer Evaluation of Multi-Product Bundles: An Information Integration Analysis. *Marketing Letters, 2,* 47-57.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires. *American Psychologist, 59,* 93-104.

Hofmans, J., & Theuns, P. (2008). On the Linearity of Predefined and Self-Anchoring Visual Analogue Scales. *British Journal of Mathematical & Statistical Psychology*, *61*, 401-413.

Hofmans, J., & Theuns, P. (in press). Testing the Impact of Predefined and Self-Defined end Anchors on the Linearity of the Category Rating Scale. *Quality & Quantity*.

Hofmans, J., Theuns, P., & Mairesse, O. (2007). On the Impact of the Number of Response Categories on Linearity and Sensitivity of 'Self Anchoring Scales'. A Functional Measurement Approach. *Methodology*, *3*, 160-169.

Homans, G. C. (1958). Social Behavior as Exchange. *American Journal of Sociology, 63*, 597-606.

Lampel, A. K., & Anderson, N. H. (1968). Combining Visual and Verbal Information in an Impression-Formation Task. *Journal of Personality and Social Psychology, 9*, 1-6.

Munos Sastre, M. T., Mullet, E., & Sorum, P. C. (2000). Self-Assessment of Inebriation From External Indices. *Addictive Behaviors, 5*, 663-681.

Meyerson, P., & Tryon, W. W. (2003). Validating Internet Research: A Test of the Psychometric Equivalence of Internet and in-Person Samples. *Behavior Research Methods, Instruments, & Computers, 35,* 614-620.

Parducci, A. (1982). Category Ratings: Still More Contextual Effects! In B. Wegener (Ed.), *Social attitudes and psychophysical measurement.* New Jersey: Erlbaum.

Pettit, F. A. (2002). A Comparison of World-Wide Web and Paper-and-Pencil Personality Questionnaires. *Behavior Research Methods, Instruments & Computers, 34,* 50-54.

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.

Schlottmann, A. (2000). Children's Judgements of Gambles: A Disordinal Violation of Utility. *Journal of Behavioral Decision Making, 13*, 77-89.

Shanteau, J., Pringle, L. R., & Andrews, J. A. (2007). Why Functional Measurement is (still) Better Than Conjoint Measurement: Judgement of Numerosity by Children and Adolescents. *Teorie & Modelli, 12*, 199-210.

Shavit, T., Sonsino, D., & Benzion, U. (2001). A Comparative Study of Lotteries-Evaluation in Class and on the Web. *Journal of Economic Psychology, 22,* 483-491.

Troutman, C. M., & Shanteau, J. (1976). Do Consumers Evaluate Products by Adding or Averaging Attribute Information? *Journal of Consumer Research, 3,* 101-106.

Von Neumann, J., & Morgenstern, O. (1967). *Theory of Games and Economic Behavior*. New York: Wiley.

Weiss, D. J. (2005). *Analysis of Variance and Functional Measurement: a Practical Guide*. Oxford: University Press.

Zhu, S., & Anderson, N.H. (1991). Self-Estimation of Weight Parameters in Multiattribute Analysis. *Organizational Behavior and Human Decision Processes, 48*, 36-54.

(Manuscript received: 12 June 2009; accepted: 6 October 2009)