On First or On Second Thought – How Response Instructions May Impact on the Quality of Measurement in Marketing Research

Thomas Salzberger, WU Wien (Austria), Thomas.Salzberger@wu.ac.at
Monika Koller, WU Wien (Austria), Monika.Koller@wu.ac.at

Abstract

Questionnaires, particularly when self-administered, often come with instructions for the respondent as to the speed with which the responses should be given. While instructions sometimes lack such a rule, spontaneous responses are asked for much more often than well-considered responses. Currently, decisions concerning this matter seem to be made largely in an atheoretical manner. In principle, spontaneous responses are assumed to be less susceptible to social desirability. However, well-thought-out responses may potentially lead to more accurate measurement and thus increase validity. Therefore, the effects of different instructions are empirically investigated using the example of an environmental attitude scale. Data analyses are primarily based on the Rasch measurement model, since the authors argue that this model is superior to the traditional approach of traditional measurement theory. The results based on the Rasch model suggest that no difference in terms of social desirability exists between spontaneous and well-considered responses. However, well-considered responses lead to enhanced construct validity. Consequently, the authors suggest providing the instruction to consider the responses carefully, or at least refraining from the request to answer spontaneously. Interestingly, comparative analyses based on factor analysis point in the opposite direction, but these findings are deemed spurious by the authors.

On First or On Second Thought – How Response Instructions May Impact on the Quality of Measurement in Marketing Research

Introduction

Instructions included in self-administered questionnaires should, first, help the respondent fill in the questionnaire without any difficulty, and second, they should ensure the highest possible validity of the measures of latent variables we infer from responses to manifest items. In this regard, instructions in terms of response speed suggest themselves as a means to elicit accurate responses and at the same time to avoid socially desirable answers. Currently, decisions concerning this matter are largely based on plausibility with asking for spontaneous responses being much more popular than requesting well-thought-out responses. Well-considered responses are assumed to be more strongly influenced by irrelevant attitudes and, above all, by social desirability. In marketing research, a thorough empirical investigation based on different measurement theories is overdue.

Time to Respond

Besides the adequacy of the items, a plethora of other factors (for example, the mode of administration, the response scale provided, the involvement of the respondent, whether or not the respondent has all necessary information to answer the question, etc.) impact on the quality of measurement. Before providing a response, the respondent has to decode the item and retrieve feelings and beliefs which are relevant to the question (Tourangeau and Rasinski, 1988). The interpretation of the item and the retrieval of information require time.

The administration of a questionnaire, particularly when self-administered, typically allows for very little control, if any, with respect to response speed. However, the researcher has the opportunity to instruct respondents and, by this means, try to influence the speed with which respondents provide their judgments. While we do not have full control over actual compliance of respondents to instructions provided, response instructions do frame the response process (Tourangeau and Ye, 2009; Wenke and Frensch, 2005). On the one hand, the respondent might be urged to respond spontaneously. On the other hand, the researcher might ask the respondent to thoroughly think about his/her opinion and provide a carefully considered response. Finally, no guideline at all might be given in terms of response speed. There is no clear general rule as to which strategy is advisable. This is due to the fact that either approach (spontaneous or well-considered response) entails potential advantages as well as disadvantages. In principle, a well-considered response should be more accurate as a consequence of the depth of information processing. By the same token, however, the likelihood increases that the respondent accounts for irrelevant beliefs and feelings and, particularly problematic, takes into consideration what is socially desirable. Another drawback of a well-considered response process lies in its being more time-consuming, which might result in higher cost, lower acceptance by respondents and, consequently, a higher dropout rate. By contrast, a spontaneous response should lead to increased measurement error and inaccurate responses as respondents are more likely to react superficially.

While many published marketing scales do not offer instructions referring to the speed of the response process, some explicitly invoke a well-thought-out response. A much higher number, though, suggest asking for a spontaneous response. The *Imagery Vividness* scale (Childers and Heckler, 1985), a 16-item scale measuring the clarity of mental images a person evokes, instructs respondents to "consider carefully the picture that comes before your mind's

eye" (Bruner and Hensel, 1996, p.321). By contrast, the *Novelty Experience Seeking* scale (Venkatraman, 1991; Venkatraman and Price, 1990) asks respondents to "[w]ork rapidly and give your first impression" (Bruner and Hensel, 1996, p.454). Similarly, the *Motivation to Work* scale (Bruner and Hensel, 1996, p.967) prompts respondents to "answer with the first response that comes to mind." The World Health Organization (WHO, 2009), in their *Perception Survey for Health-Care Workers* scale, also asks to respond spontaneously. The *Occupational Personality Questionnaire* (OPRA Consulting group) requests respondents to "try to be spontaneous in your answers and give your first-response answers rather than thought out ones." (p.1); the Online Alexithymia Questionnaire (Thompson, 2005) asks to "[a]nswer the (...) questions as spontaneously as possible." Practical guidelines of questionnaire design often mention the lack of spontaneous responses as a drawback of questionnaires. Nimkar (2010) advises questionnaire developers to "[a]sk [the] respondent to give [a] spontaneous response to know his top of mind awareness". In summary, whenever response time instructions are given, typically spontaneous responses are requested.

Empirical studies on the effects of response instructions are scarce in marketing research. To the best of our knowledge, there are only a few studies related to experimental research in psychology (e.g., Lievens, Sackett, and Buyse, 2009; McDaniel *et al.*, 2007; Wenke and Frensch, 2005). Instructions regarding speed of responding usually aim at eliciting impulsive versus elaborate answering processes. Impulsive answers are thought to mirror more intuitive, emotional and subconscious attitudes whereas elaborate answers result from a greater extent of cognitive mental processing. It has been shown repeatedly that when decisions have to be made under time-constraints, individuals use different strategies of decision-making. The strategies applied vary regarding the amount of information considered and the mental effort invested (Glöckner and Betsch, 2008). Automated strategies of decision-making are likely to be used under two different conditions: first, if a quick inspection of information is possible; and second, if information has to be retrieved from memory. Concerning decisions under time constraints, individuals predominantly use non-compensatory strategies rather than complex compensatory strategies with contradictory empirical evidence as to why this is the case. It could be due to limitations of cognitive capacity but also due to limitations of the search for information (Glöckner and Betsch, 2008). However, Barnard, Wright and Wilcox (1979) point out that respondents have presuppositions in terms of, for example, filling in a form. Such schemas may contradict instructions given. In an experimental setting the authors noted that "[i]t cannot be assumed that they will stop, read and respond appropriately to instructions which run counter to these presuppositions." (p. 224).

Empirical Study

The empirical study aims at testing the effect of instructions in terms of response speed on the mean value as well as on construct validity. Three experimental groups are distinguished. The first group ("well-considered") was instructed to carefully consider their responses and think about pros and cons before choosing a response category, while the second group ("spontaneous") was prompted to reply spontaneously without pondering on their reaction. The third group was the control group, which did not get any additional instruction.

While environmental issues have always been on the marketing research agenda, recently, interest in environmental issues has increased even more (Griskevicius, Tybur, and Van den Bergh, 2010; Kilbourne, Beckmann, and Thelen, 2002). In this context, the *New Environmental Paradigm Scale* (NEP) was proposed by Dunlap and Van Liere (1978), consisting of 12 items. Recently, Dunlap *et al.* (2000) suggested a revised NEP scale (new ecological paradigm) comprising 15 items. Four items have been discarded from and seven

new ones added to the original scale. The NEP scale has been widely used in the measurement of pro-environmental orientation over the past 20 years (Dunlap *et al.*, 2000).

We decided to administer the original version of the NEP scale as an exemplary construct in the empirical study, since the revised version does not seem to be superior from a Rasch measurement perspective (Rasch, 1960; see methodology section). However, future empirical research has to examine how the new items blend in. The response scale consists of a seven-point rating scale with endpoints being labelled fully agree and fully disagree, respectively. It is plausible that at least some respondents deem it socially desirable to respond in a pro-environment fashion.

The data collection was administered online among students at a major European Business School. The sample size amounts to a total of 1644 respondents. However, the results reported refer to one half of the sample, since the other half of respondents were presented a different response scale (the results do not differ substantially, though). The sample is relatively homogeneous in terms of age (mean 25 years, 97% are under 41) and certainly not representative of the total population. However, the invariance property of the Rasch model allows for a valid assessment of item properties regardless of the sample composition (Rasch, 1977). Respondents were allocated randomly to one of the three groups.

Hypotheses

Well-considered responses are expected to be most consistent inasmuch as manifest responses are most closely related to the latent variable. We therefore propose that the measurement model fits best when responses are well-considered (hypothesis A).

Spontaneous responses are assumed to be less susceptible to socially desirable responses, while the instruction to provide a well-thought-out response should lead to a biased mean. We therefore expect the mean of the latter group to be shifted towards a pro-environmental attitude compared to the control group and the spontaneous group (hypothesis B).

Methodology

The scale analyses were primarily based on item response theory (IRT), specifically the Rasch measurement model for categorical data (Andrich, 1978, 1988) using RUMM 2030 (Andrich, Sheridan, and Luo, 2010). Since classical test theory is still the predominant measurement model in marketing research, comparative confirmatory factor analyses (CFA) were conducted employing Mplus (Muthen and Muthen, 2008). As regards the differences between IRT and CFA, we focus on the most important issue in the context of the present study (for details the reader is referred to Ewing, Salzberger, and Sinkovics, 2005). In IRT, for each item a location parameter is estimated (as the mean of a set of threshold parameters characterising the response categories). For valid measurement, sufficient variation in terms of item locations is required in order to cover a wide range of the latent variable (Singh, 2004). The model states a probability for each response category based on the person location parameter and the item parameter. In the test of fit, actual responses are then compared with expected responses. Consistency implies that respondents choose categories in line with their location on the latent variable and the item characteristics. By contrast, CFA relies on correlations between item scores. Since floor and ceiling effects may bear on the size of correlations when only a part of the items are affected, responses which consistently replicate across all items contribute to stronger item inter-correlations and, as a consequence, to model fit.

Findings Based on the Rasch Model

The first step of the data analysis was the assessment of the fit of the data to the Rasch model. During this stage, the sample was confined to the control group (n=273). Following an iterative process, three items had to be discarded before acceptable fit could be achieved (χ^2 (81)=90.286, p=0.225). The same set of items also fits best in all other experimental groups. Further tests specifically sensitive to violations of unidimensionality (principal component analysis of item residuals) confirmed the construct validity of the instrument. This in itself is an important finding, as it shows that, in general, the Rasch model is applicable to marketing scales and, specifically, that the NEP scale possesses satisfactory psychometric properties.

Subsequently, hypothesis A was tested by comparing the fit across all three experimental groups. In the group where respondents were asked to provide a well-considered response, the fit of the data to the model was even better (χ^2 (81)=81.657, p=0.459), while in the spontaneous response group the data fitted the model unacceptably badly (χ^2 (81)=121.217, p=0.003). Consequently, there is clear evidence that the instruction to respond spontaneously compromises the accuracy of the responses. The superiority of the "well-considered" instruction is tentative, though, as hypothesis B suggests well-considered responses being more likely prone to social desirability. However, the group means of all three experimental groups prove otherwise. In the spontaneous group, the mean value (0.889) is highest, while the mean in the well-considered group amounts to 0.805. The mean in the control group is 0.800. Higher values indicate a more pronounced pro-environmental attitude. However, the three mean values do not differ significantly. Consequently, there are no signs whatsoever that a well-considered response implies a stronger effect of social desirability compared to a spontaneous response. The conclusion that it is advisable to instruct respondents to consider their response carefully rather than asking them for spontaneous reactions is therefore supported. However, refraining from any instructions at all in terms of response speed seems to be a viable option as well.

As mentioned above, the instruction to respond spontaneously or in a well-considered way does frame the response process (also demonstrated by the significantly faster actual response speed in the spontaneous group), but it cannot force the respondents to comply. It is therefore interesting to see, how the data fit the measurement model depending on the actual response speed. For this purpose, we divided the sample into three groups (fast, medium, slow) based on the 33rd and 67th percentile of the distribution of response speed. The best fit could be achieved in the medium-speed group (p=0.366), followed by the slow respondents (p=0.131) and the fast respondents (p=0.013). Again, the fast respondents respond less accurately (that is, deviating from the Rasch model expectations) suggesting that respondents should not be prompted to react quickly. Of course, the time needed to provide a well-considered response varies between subjects. Therefore, we cannot conclude that a fast response time necessarily implies that a respondent did not consider the response carefully.

Findings Based on Confirmatory Factor Analysis (CFA)

For comparative purposes, we also analysed the data based on CFA. Generally speaking, the CFA model showed rather poor fit. However, we are not focusing on the absolute assessment but look primarily at the relative fit in the three experimental groups. Strikingly the data from the group with the spontaneous instruction fit the CFA model best (χ^2(27)=97.07, CFI=0.86, RMSEA=0.098). The model in the control group performs second best (χ^2(27)=106.93, CFI=0.82, RMSEA=0.104), while the well-considered responses lead to the worst fitting model (χ^2(27)=172.46, CFI=0.70, RMSEA=0.139). In terms of variance explained the three groups rank in the same order as according to fit. Consequently, the CFA approach would

strongly advise against instructions to respond in a well-considered way. In fact, the opposite, which is, asking for spontaneous responses, seems to be advisable in order to ensure that respondents do not agonise about the decision which response category to choose,

The diametrically opposed conclusions from Rasch modelling and traditional CFA raise serious concerns. The lack of consensus seems to be due to spontaneous responses being more consistent inasmuch as similar manifest response categories are chosen for all items (the respondent basically replicates a particular choice of a category across all items resulting in a response set). In addition, relatively more respondents consistently resort to more extreme responses and do not take intermediate categories into consideration leading to higher variance of manifest scores. Robinson, Rush, and Head (1968) already alerted scale builders that speeding through the instrument leads to response set contamination. Both, the higher consistency as well as the increased variance (which is picked up by CFA as true variance), lead to a better fit of the spontaneous group compared to the well-considered group.

In the Rasch model, these response patterns have an opposite effect on the fit, since respondents are not expected to provide the same response to all items irrespective of the item characteristics. Therefore, what appears to be response consistency in CFA is actually response inconsistency in the Rasch model. Furthermore, more extreme responses do not inflate the model fit in Rasch modelling but rather lead to more unexpected responses. Consequently, the better fit of the data based on the spontaneous response instructions in CFA and, vice versa, the worse fit of the data from the well-considered instruction, are a methodological artefact pointing out that a measurement model building upon correlations is inappropriate.

Conclusions and Further Research

The assumption that spontaneous responses are more likely to reflect true attitudes and are less affected by social desirability turned out to be questionable. The empirical study clearly showed that well-considered responses (or rather the instruction to behave in this manner) lead to more accurate measurement and higher construct validity without necessarily evoking a social desirability bias. This is in sharp contrast to the prevalence of instructions urging subjects to respond spontaneously, which may also be due to the fact that CFA seems to reward such an approach. However, the consistency of spontaneous responses is spurious and misleading. There is good reason to assume that the results would replicate for a construct which is less susceptible to socially desired responses. However, whether spontaneous responses would be superior (or at least not worse) in case of more intuitive, emotional and subconscious attitudes, which are said to be better measured by spontaneous responses, remains an issue to be investigated empirically. In any case, as a default recommendation we would advise refraining from specific directions in terms of response speed.

References

Andrich, D., 1978. A Rating Formulation for Ordered Response Categories. Psychometrika, 43 (4), 561-573.

Andrich, D., 1988. A General Form of Rasch´s Extended Logistic Model for Partial Credit Scoring. Applied Measurement in Education 1 (4), 363-378.

Andrich, D., Sheridan, B.S., Luo, G., 2010. Rumm2030: Rasch Unidimensional Measurement Models [computer software]. RUMM Laboratory Perth, Western Australia.

Barnard, P.J., Wright, P., Wilcox, P., 1979. Effects of response instructions and question style on the ease of completing forms. Journal of Occupational Psychology 52, 209-226.

Bruner, G.C., Hensel, P.J., 1996. Marketing scales handbook: a compilation of multi-item measures, American Marketing Association, Chicago, IL.

Childers, T.L., Heckler, S.E., 1985. Measurement of individual differences in visual versus verbal information processing. Journal of Consumer Research 12 (September), 125-134.

Dunlap, R.E., Van Liere, K.D., 1978. The "new environmental paradigm": a proposed measuring instrument and preliminary results. Journal of Environmental Education 9, 10-19.

Dunlap, R.E., Van Liere, K.D., Mertig, A.G., Jones, R.E., 2000. Measuring endorsement of the New Ecological Paradigm: A revised NEP scale. Journal of Social Issues 56 (3), 425-442.

Ewing, M., Salzberger, T., Sinkovics, R., 2005. An alternate approach to assessing cross-cultural measurement equivalence in advertising research. Journal of Advertising 34 (1), 17–36.

Glöckner, A, Betsch, T., 2008. Multiple-reason decision making based on automatic processing. Journal of Experimental Psychology: Learning, Memory, and Cognition 34 (5), 1055-1075.

Griskevicius, V., Tybur, J.M., Van den Bergh, B., 2010. Going green to be seen: status, reputation, and conspicious conservation. Journal of Personality and Social Psychology 98 (3), 392-404.

Kilbourne, W.E., Beckmann, S.C., Thelen, E., 2002. The role of the dominant social paradigm in environmental attitudes. A multinational examination. Journal of Business Research 55 (3), 193-204.

Lievens, F., Sackett, P.R., Buyse, T., 2009: The effects of response instructions on situational judgment test performance and validity in a high-stakes context. Journal of Applied Psychology 94 (4), 1095-1101.

McDaniel, M.A., Hartman, N.S., Whetzel, D.L., Grubb, W.L. III 2007. Situational judgment tests, response instructions, and validity: a meta analysis. Personnel Psychology 60, 63-91.

Muthén, B., Muthén L., 2008. Mplus (http://www.statmodel.com/) [computer software]. Muthén & Muthén, Los Angeles, CA.

Nimkar, C.Y., (2010). Questionnaire design. Available online: http://www.scribd.com/doc/25419095/Questionnaire-Design, last accessed: June 10, 2010.

OPRA Group Limited (undated). The Occupational Personality Questionnaire. http://www.opragroup.com/, last accessed. June 10, 2010.

Rasch, G., 1960. Probabilistic Models for Some Intelligence and Attainment Tests. MESA, Chicago, IL, Reprint 1980, Danish Institute for Educational Research.

Rasch, G., 1977. On Specific Objectivity. An Attempt at Formalizing the Request for Generality and Validity of Scientific Statements. Danish Yearbook of Philosophy 14, 58-94.

Robinson, J.P., Rush, J.G., Head, K.B., 1968. Criteria for an attitude scale. In: Robinson, J.P. (Ed.), Measures of political attitudes, Survey Research Center, Ann Arbor, MI, pp. 9-21; Reprinted in: Maranell, G.M. (Ed.), Scaling: A sourcebook for behavioral scientists, AldineTransaction, New Brunswick (U.S.A.) and London (U.K.), pp. 244-257.

Singh, J., 2004. Tackling Measurement Problems with Item Response Theory: Principles, Characteristics, and Assessment, with an Illustrative Example. Journal of Business Research 57 (2), 184–208.

Thompson, J., (2005). Alexithymia Questionnaire – G2. Available online: http://oaq.blogspot.com/, last accessed: June 10, 2010.

Tourangeau, R., Rasinski, K.A., 1988. Cognitive processes underlying context effects in attitude measurement. Psychological Bulletin 103 (3), 299–314.

Tourangeau, R, Ye, C., 2009. The Framing of the Survey Request and Panel Attrition. Public Opinion Quarterly 73 (2), 338-348.

Venkatraman, M.P., 1991. The impact of innovativeness and innovation type on adpotion. Journal of Retailing 67 (Spring), 51-67.

Venkatraman, M.P., Price, L.L., 1990. Differentiating between cognitive and sensory innovativeness: concepts, measurement, and implications. Journal of Business Research 20 (June), 293-315.

Wenke, D., Frensch, P. A., 2005. The influence of task instruction on action coding: Constraint setting or direct coding? Journal of Experimental Psychology: Human Perception and Performance 31, 803-819.

World Health Organization, 2009. Perception Survey for Health-Care Workers. Available online: http://www.who.int/entity/gpsc/5may/Perception_Survey_for_Senior_Managers.doc, last accessed: June 10, 2010.