Influence of Error Factors in Marketing Research: A Monte Carlo Method Based Analysis in the Retail Context

Michael Brusch, Ines Brusch and Eva Stüber

Abstract Marketing decisions are often based on empirically collected data and the goodness of data plays an important role here. Data quality can be affected by several types of errors, which are distinguished in particular into systematic and random errors, and is also influenced by the sample size. Accordingly, market researchers have to consider the goodness of data and should know which factors will have which kind of influence. In our paper, the influence of different factors of the goodness of data in the vibrant retail context will be investigated within a Monte Carlo experiment. For this purpose, a real empirical data set (n=1,500) of a survey regarding buying behavior in stationary and online shopping is used as "true" data and will be compared with "generated" data. The "generated" data are randomly disturbed and systematically varied alternatives of the "true" data. The data sets will be compared with respect to their conformity values and allow influence estimations.

Michael Brusch ✉ michael.brusch@hs-anhalt.de
Anhalt University of Applied Sciences, P.O. Box 1458, D-06354 Köthen, Germany

Ines Brusch ✉ ines.brusch@b-tu.de
Brandenburg University of Technology Cottbus-Senftenberg, P.O. Box 101344, D-03013 Cottbus, Germany

Eva Stüber ✉ e.stueber@ifhkoeln.de,
IFH Köln GmbH Dürener Str. 401 b, D-50858 Cologne, Germany

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 6, No. 1, 2020

DOI: 10.5445/KSP/1000098011/05
ISSN 2363-9881

1 Introduction

Decisions in marketing are mainly supported by methods of marketing research due to their far-reaching consequences. Such methods are often based on empirically collected data and evaluated using complex analysis methods. The goodness of data of derived results is important and can be affected by several factors as systematic and random errors as well as the sample size (e.g. Cochran, 1968). Accordingly, market researchers should consider the quality of data and know which factors have which influence. This problem is particularly relevant in the case of new or difficult to describe offers such as (stand-alone or partial) services with their intangibility (e.g. Parasuraman et al., 1985; Baumert et al., 2011). Especially the increasing importance of services in several markets makes it necessary to consider this in market research projects. However, the interesting issue in all kind of empirical measurements is the goodness of data collected within a survey and in the following of the derived results.

In our paper, the influence of different factors of the goodness of empirical data in the vibrant retail context (e.g. Kushwaha and Shankar, 2013; Li and Kannan, 2014) is investigated via a Monte Carlo experiment (e.g. Fishman, 1996; Brusch and Baier, 2010). Therefore, a real-world empirical data set (n=1,500) of a survey regarding buying behavior in brick-and-mortar shopping and online shopping is used as "true" data. This data will be compared with "generated" data which are randomly disturbed and systematically varied alternatives of the "true" data. Both types of data will be compared with regard to common conformity values to allow conclusions about the focused influence estimations considering the frequently occurring problem of missing values and the frequent use of descriptive measures in practice (in contrast to multivariate method based measures in research; e.g. Gatty, 1966; Sheth, 1971).

Accordingly, this paper is structured as follows. In Sect. 2 a short overview about issues of the goodness of data will be given. The following Sect. 3 presents the empirical investigation, which serves as basis for later analyses. Sect. 4 gives an overview about the Monte Carlo comparison and its main results. A part with a conclusion and an outlook in Sect. 5 closes this contribution.

2 Goodness of Data

When evaluating scales and measured values researchers have to consider the goodness of data. Here the measurement accuracy has to be mentioned. The measurement of a characteristic of an object can be influenced by several potential sources of error (e.g. Malhotra and Birks, 2007):

- Individual characteristics (e.g. intelligence and social desirability),
- Short-term or transient personal determinants (e.g. health and emotions),
- Situational determinants (e.g. the presence of other people and noise),
- Item sampling (e.g. deletion or changes in the scale items),
- Scale clarity (e.g. item formulation or instructions),
- Mechanical factors (e.g. poor printing and overcrowding items),
- Scale administration (e.g. differences among interviewers),
- Analysis factors (e.g. differences in scoring and evaluation).

The total measurement error consists of both a systematic error and a random error. A systematic error has a constant influence on the measurement (e.g. as a mechanical factor) and a random error does not have a constant influence on every measurement but rather in different ways (e.g. as short-term transient personal or situational factors; Malhotra and Birks, 2007).

The evaluation of a measured scaled can involve an assessment of objectivity, generalizability, reliability and validity. In addition to objectivity (as independence of the market research results from the persons involved) and generalizability (which refers to the fact that one can infer from the present observation generally admitted relationships), reliability and validity are particularly important and must be taken into account (e.g. Malhotra and Birks, 2007; Herrmann et al., 2008).

Reliability as freedom from random errors encompasses the formal accuracy of the characteristic recording. Accordingly, a measuring instrument is reliable (if the measuring conditions are constant) if the measurement results are both precise and stable and thus reproducible in repeated measurements. The validity as freedom from systematical errors describes the content-related accuracy of the measurement results. The validity of a measurement method exists when the actual facts of interest are actually recorded, that is only then given, when exactly that is measured what should be measured (e.g. Berekoven et al., 1999; Herrmann et al., 2008). This has already been taken into account in the literature through information on the correct handling of selected errors, e.g. as guidelines for investigating construct validation (e.g. Bagozzi et al., 1998), for avoiding Type IV errors (resulting from the inappropriate treatment of interactions in an analysis of variance, e.g. Umesh et al., 1996), for considering nonsampling and sampling errors (e.g. Assael and Keon, 1982; Anderson and Gerbing, 1984).

Another problem are missing values. These can be differentiated according to whether individual values of an observation (“item non-response”) or entire observations (“total non-response”) are missing. Accordingly, these must be given special consideration in the analysis (e.g. Brand et al., 1994; Acock, 2005; Decker and Wagner, 2008).

3 Empirical Retail Data

The basis of the later Monte Carlo comparison is an investigation in the retail context by the E-Commerce Center Köln (ECC Köln, 2017). The base of operations was the changing buyer structure: The number of traditional retail buyers are dying out and the proportion of selective buyers is growing. To this end, retailers are upgrading and expanding their cross-channel offerings. However, from the customer’s point of view, some basic prerequisites for successful cross-channels are lacking.

In this context the goal of the ECC Köln investigation was the identification of starting points for a better understanding of the customer, taking into account branch-specific conditions. The focus was on obtaining reliable results for prioritizing investments in the company’s own cross-channel strategy, taking into account different types of buyers and purchasing needs.

In this study 1,500 Internet users were surveyed with a branch-specific view and taking into account the real turnover shares in 2017. Their allocation according to sectors and purchasing channels (in a brick-and-mortar or an online shop) is shown in Fig. 1.

In Fig. 2 some results regarding their multi-channel behavior is shown. For example, it can be seen that 45.1% of purchases in stores are preceded by an information search on the Internet. The figure for smart consumers is in this case as high as 54.3%. In 81.3% of online purchases, consumers inform themselves exclusively online. Overall, it becomes clear from Fig. 2 that online information

	CE and home appliances	Leisure and hobby	Living and furnishing	Do-it-yourself and garden	Fashion and accessories
Brick-and-mortar shop	**21 %**	**13 %**	**17 %**	**21 %**	**26 %**
Online shop	**32 %**	**14 %**	**11 %**	**6 %**	**36 %**

Figure 1: Branch-specific proportions of purchases of the investigated online buyers (n=1,500) (ECC Köln, 2017).

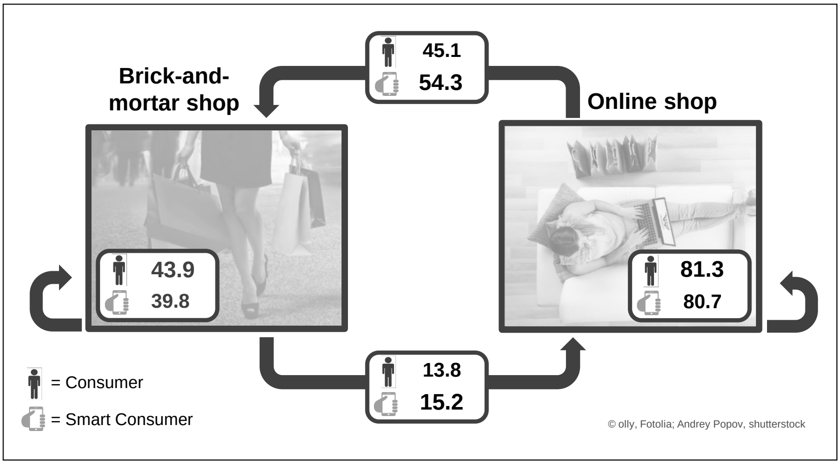

Figure 2: Multi-channel behavior of the investigated online buyers (71 ≤ n ≤ 1,221; values in %) (ECC Köln, 2017).

is highly relevant before the stationary purchase and that the behavior of smart consumers shows a further increasing relevance of channel linking.

Fig. 3 presents summarized results regarding the satisfaction with a purchase. It becomes clear that a cross-channel offer leads to the most satisfied customers for brick-and-mortar purchases, whereas online, on the other hand, the mono-channel purchase is more convincing, which also takes place much more frequently. At this place we would also like to point out the problem of missing values, which can be seen in Fig. 3 on the basis of the partly low sample sizes (ranging between 50 and 1,015) compared to the total sample size (1,500) and is often found in practice.

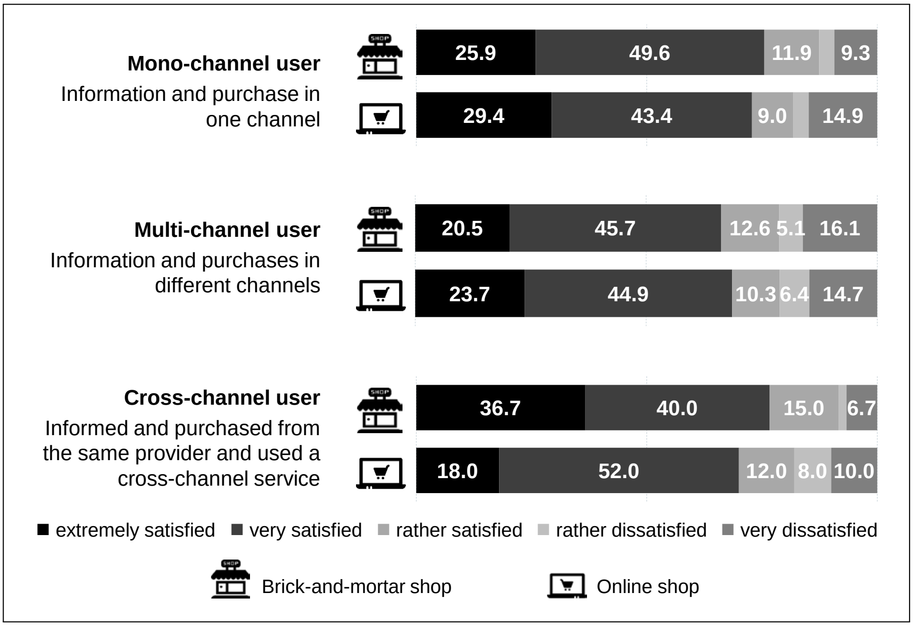

Figure 3: Satisfaction with purchase ($50 \leq n \leq 1{,}015$; values in %) (ECC Köln, 2017).

The data set presented in this section is thus the basis for further analyses within a Monte Carlo comparison, which is presented in the following.

4 Monte Carlo Comparison

In order to answer the question which factors have which influence on the goodness of data, a Monte Carlo comparison is carried out. In the following, an overview of the considered factors and their levels as well as the determination of comprehensible conformity values is given (Sect. 4.1). Last but not least, the main results are presented and discussed (Sect. 4.2).

4.1 Overview

Within the Monte Carlo comparison two data sets have been used. One (named as "true" in the following) is based on the responses of the retail investigation previously described (see Sect. 3). The other one (named as "generated") consists of randomly disturbed and systematically modified variations of the "true" data set. Therefore, a factorial design with five factors – each with three levels – was used. The first three factors consider possible influences with respect to (w.r.t.) the data gathering step and vary

- the measurement error w.r.t. the evaluation of the respective question (with small, medium or large standard deviations of an additive error) – factor 1,
- the systematical error w.r.t. the evaluation of the respective question (with a positive, negative or no displacement of the basic evaluation score) – factor 2 and
- the underlying distribution for disturbances (with normal, uniform or lognormal distribution) – factor 3.

The disturbance of factor 1 considers a limitation of the scale borders (i.e., $min = 1$ and $max = 5$), allows discrete values only and used a specified distribution (see factor 3). This factor is used to simulate random errors.

Similarly, the disturbance of factor 2 considers also a limitation of the scale borders (i.e., $min = 1$ and $max = 5$) and allows discrete values only. This disturbance leads then to a displacement of plus or minus one scale point (in every case when applied). This factor is used to simulate uncertainty regarding the overall evaluation range and/or a basically positive or negative attitude.

Factor 3 takes into account the kind of the error influence, using the normal distribution to simulate variability depending on the form of the day, uniform distribution for simulating general indifference and lognormal distribution for simulating decisions close to reality.

The disturbance is realized in a way that additions or deductions are applied to the discretely collected original values ("true" data) depending on the first three factors. The new continuous values are then transformed back into discrete values by rounding while limiting to the permissible scale borders. The distribution parameters are set to relevant values, i.e. to a minimum of -1 and a maximum of

Table 1: Interpretation of the levels of F4: Weighting of question types.

Characteristic	Type1	Type2	Type3
Content	Purchase category	Satisfaction	Involvement
Scale type	Nominal	Ordinal	Interval
Response scale	1 = Fashion and accessories 2 = Living and furnishing 3 = Consumer electronics and home appliances 4 = Leisure and hobby 5 = Do-it-yourself and garden	1 = very dissatisfied 2 = rather dissatisfied 3 = rather satisfied 4 = very satisfied 5 = extremely satisfied	1 = does not apply at all 2 . . . 3 . . . 4 . . . 5 = applies completely
Level “1:1:1”	Weight 1	Weight 1	Weight 1
	Interpretation: Equal relevance of all types		
Level “1:1:2”	Weight 1		Weight 1
	Interpretation: Equal relevance of non-metric and metric scaled questions		
Level “1:1:4”	Weight 1		Weight 2
	Interpretation: High relevance of metric scaled questions		

+1 for uniform, to a mean of 0 and a standard deviation of 1 for the normal and lognormal distribution, whereby a randomly drawn sign change was integrated in case of the lognormal distribution (to simulate positive and negative disturbance). The other factors are related to the calculation of conformity measures and vary

- the weighting of the question types (with an increasing relevance of metric scales) – factor 4 and,
- the share of the sample size (with an increasing proportion of respondent data used) – factor 5.

Factor 4 considers the influences of question types which are applied in real-world research practice (in contrast to scientific investigations). The three levels used (which are described in Table 1) allow insights regarding the different use of metric and non-metric scales in questionnaires.

For factor 5 and the consideration of sample size effects there are in principle two alternatives. One is the selection of respondents after arbitrary sorting with an exclusion of uninterested (i.e., not valid) answers at the beginning

Table 2: Overview about considered values and applied disturbing factors of the Monte Carlo comparison (with "X" when applied during Monte Carlo simulation and "TSD" as total sum of differences).

Characteristic	Type 1	Type 2	Type 3
Content	Purchase category	Satisfaction	Involvement
Basis value: mean	Brick-and-mortar / online shop	Brick-and-mortar / online shop	Brick-and-mortar / online shop
F1: Measurement error			X
F2: Systematical error			X
F3: Underlying distribution			X
F4: Weighting of question types	X	X	X
F5: Sample size	X	X	X
Measure of relationship	"Correlation": Spearman rank order correlation between "true" and "generated" data (possible range: -1 to +1)		
Measure of difference	"TSD": rescaled absolute value of all differences of both percentages scales ("true" and "generated") (possible range: 0 to 100)		

and/or the end. An other one is a random selection with a reduction of negative systematical influences. The latter one is here chosen and used for a simulation of the necessary number of respondents.

The conformity is determined by two measures – the Spearman rank order correlation coefficient (named as "correlation" in the following) and the total sum of differences (named as "TSD"). The TSD quantifies thereby the difference in the relative frequencies of the rated response categories (as an absolute value) between the two data sets ("true" and "generated") on a percentage scale. As shown in Table 2, all five types of disturbances apply only to metric scale-based questions (i.e. to "involvement").

All in all a full factorial 3^5-design results and leads with hundredfold replication to a total of 24,300 datasets which will be analyzed in the following.

Table 3: Monte Carlo comparison concerning the influence of error factors using mean Spearman rank order correlation ("mean correlation") and mean total sum of differences ("mean TSD").

Factor	Level	Mean correlation	Mean TSD
F1: Measurement error	$\sigma = .1$	0.809^{***}	12.95^{***}
	$\sigma = .2$	0.784	13.54
	$\sigma = .3$	0.710	14.66
F2: Systematical error	-1	0.573	17.28
	0	0.893^{***}	6.43^{ns}
	+1	0.838	17.43
F3: Underlying distribution	Normal	0.751	13.20
	Uniform	0.807^{*}	12.81^{***}
	Lognormal	0.746	15.13
F4: Weighting of question types	1:1:1	0.838^{***}	9.67^{***}
	1:1:2	0.768	13.71
	1:1:4	0.698	17.80
F5: Sample size	1/3	0.756	14.60
	2/3	0.768	13.69
	3/3	0.779^{***}	12.85^{***}
Overall		0.768	13.72

***. . . significant differences between the influences of the levels of a factor, i.e. within columns of each factor (F-test), where significance is given at the best value (i.e. highest correlation or lowest TSD) at the $p < .001$ level;
** . . . at the $p < .01$ level;
* . . . at the $p < .1$ level; ns . . . not significant.

4.2 Main Results

For each data set comparing "true" and "generated" data the correlation and TSD values are calculated and shown as mean values in Table 3. Additionally, the significance of the differences with regard to levels (i.e., columns) is presented (based on F-tests).

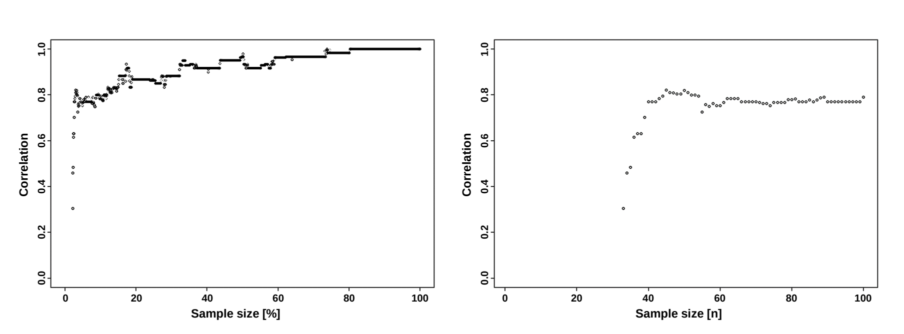

Figure 4: Detailed analysis of the correlation and the share of sample size (F5) of the whole sample (n=1,500) in percentages (left) and in absolute values for the first 100 respondents (right).

When looking at the correlation values, it is noticeable that the overall mean with a value of 0.768 is quite acceptable. This also becomes clear when the small variations across all correlation values (from Table 3) is taken into account. The extreme values (not shown in the Table) also indicate (with an overall minimum correlation of -0.200 and an overall maximum correlation of $+1.000$) mainly positive relations. The best results are found, unsurprisingly, with the least measurement error (level "σ=.1"), without a systematic error (level "0") and with the largest sample (level "3/3"). It is rather surprising that the uniform distribution (level "uniform") and the equal relevant question types (level "1:1:1") lead to the best results regarding the correlation.

Similar findings are observable when looking at the TSD. The overall mean TSD is with 13.72% also quite acceptable (having in mind the range of this percentage scale is between 0 and 100%). The extreme values (again not shown in the Table) have also (with an overall minimum total sum of differences of 0% and an overall maximum total sum of differences of 32.72%) no real outliers.

To enhance the analyses a special focus will be in the following on the sample size (see Fig. 4–7). For this reason, the sample size is no longer just differentiated between the three categories (i.e., the levels "1/3", "2/3" and "3/3"), but directly analyzed by the number of respondents.

Accordingly, Fig. 4 shows the course of the correlation for the total sample of 1,500 respondents as an overall view on the left and for the first 100 respondents as a detailed view on the right. It is identifiable that after approx. 60% of the total

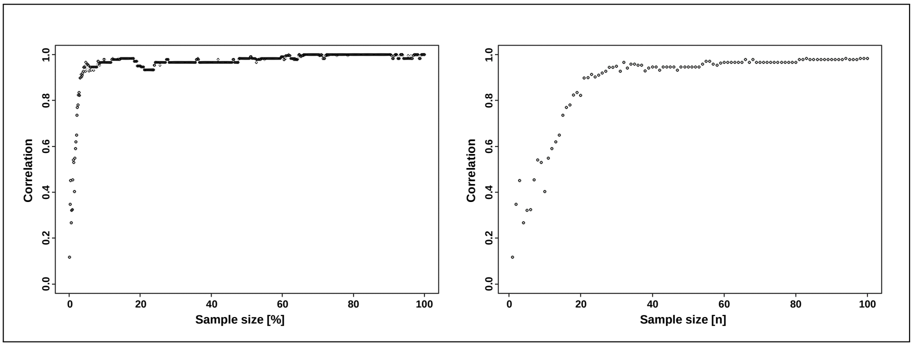

Figure 5: Detailed analysis of the correlation and the share of sample size (F5) of the sample without missing values (n=677) in percentages (left) and in absolute values for the first 100 respondents (right).

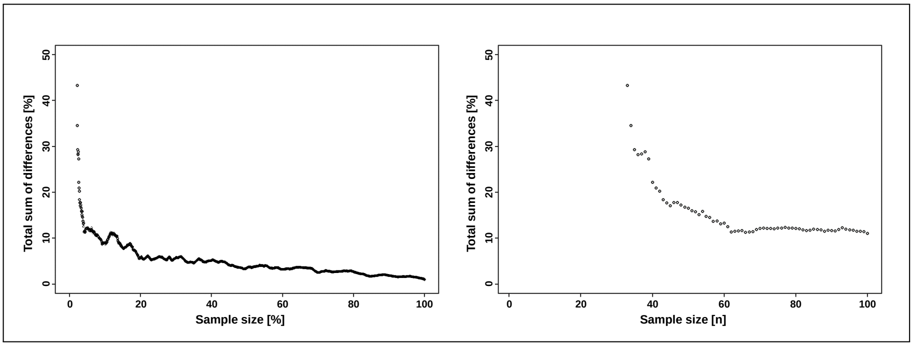

Figure 6: Detailed analysis of the TSD and the share of sample size (F5) of the whole sample (n=1,500) in percentages (left) and in absolute values for the first 100 respondents (right).

sample a correlation of around 0.9 is reached and for the first 100 respondents of around 0.8.

As shown at the description of the underlying real-world data set (see Sect. 3) a problem is sometimes the high number of missing values. Therefore, an analysis has been added where the respondents with missing values have been excluded. As shown in Figure 5 the correlation reaches much earlier high values, i.e., after approx. 5% or 30 respondents a correlation of about 0.9 is observable. The same analyses regarding the influence of the sample sizes are made for the TSD. They are first made for the total sample (n=1,500, cf. Figure 6). After approx. 30% of the sample size a TSD lower than 5% and after the first 100

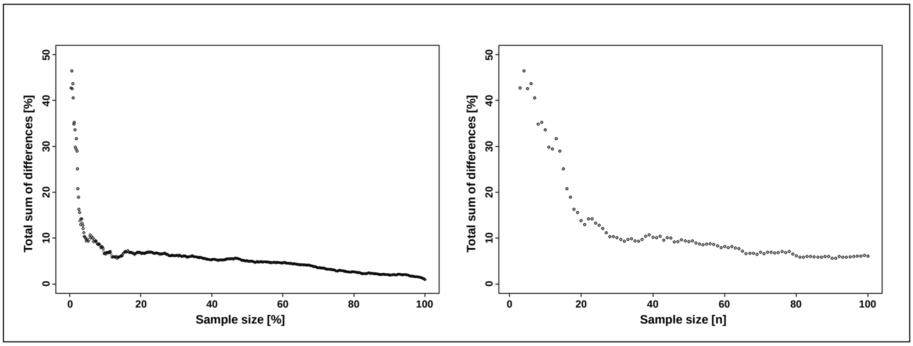

Figure 7: Detailed analysis of the TSD and the share of sample size (F5) of the sample without missing values (n=677) in percentages (left) and in absolute values for the first 100 respondents (right).

respondents of around 10% is identifiable. They are then complemented for the part of the sample without missing values (n=677, cf. Figure 7). The curves show again an extremely fast improvement resulting in a TSD lower than 10% after approx. 5% of the sample or after approx. 30 respondents).

Overall, our results show that the type of question to be addressed on the one hand and the avoidance of missing values on the other are important. Since the questions of market research practice are very often non-metric in nature, the results show that stable results are achievable relatively simple (here identifiable as high correlation and low TSD values for the level "1:1:1" of F4 in Table 3). Hence, this has positive effects on the goodness of data obtained and on the determination of the necessary sample size. Independently of this, avoiding missing values also means that a smaller sample size can produce stable results (here a number of approx. 30 respondents yields an acceptable correlation of 0.9, see on right plot in Figure 7).

5 Conclusion and Outlook

In general this contribution shows two main results. First, that Monte Carlo experiments are able to identify relevant influence factors (even without enormous efforts for empirical data collections). Second, that uniform distributed errors have the least effects (which could indicate a compensatory influence across all respondents).

In case of retail surveys the researcher should test whether the scale range is appropriate (here a fixed displace of -1 has a different effect than a displace of $+1$ for the correlations). Furthermore, the researcher should consider the right number of respondents, especially for metric values and the subsequent application of multivariate methods (here, e.g., the worst mean TSD in case of level "1:1:4" of factor 4 has to be noticed). Finally, the researcher should definitely check the possibility to avoid missing values to reduce the number of necessary respondents (here a number of about 30 respondents seems to lead to an acceptable correlation).

Further research with more real-world data sets is needed to identify generalizable results. This research can additionally focus on three issues. First, the integration of additional data analysis methods can be tested (e.g., using cluster analysis to identify different respondent groups – identifiable, e.g., by means of the standard deviation of factor 5). Second, the refinement of the data collection step (e.g., via enlargement of the response scale with seven instead of five points – identifiable, e.g., by means of the standard deviation of factor 2). Third, the determination of the optimal sample size (ideally compared with cost issues) can be placed on a broader basis using further real-world data sets.

References

Acock AC (2005) Working with Missing Values. Journal of Marriage and Family 67(4):1012–1028. DOI: 10.1111/j.1741-3737.2005.00191.x, https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1741-3737.2005.00191.x

Anderson JC, Gerbing DW (1984) The Effect of Sampling Error on Convergence, Improper Solutions, and Goodness-of-Fit Indices for Maximum Likelihood Confirmatory Factor Analysis. Psychometrika 49(2):155–173. DOI: 10.1007/BF02294170.

Assael H, Keon J (1982) Nonsampling vs. Sampling Errors in Survey Research. Journal of Marketing 46(2):114–123. DOI: 10.2307/3203346.

Bagozzi RP, Yi Y, Nassen KD (1998) Representation of Measurement Error in Marketing Variables: Review of Approaches and Extension to Three-Facet Designs. Journal of Econometrics 89(1-2):393–421. DOI: 10.1016/S0304-4076(98)00068-2.

Baumert K, Baier D, Brusch M (2011) Identifying and Evaluating Characteristics that are Difficult to Quantify Using the Repertory Grid Technique. In: Jens J. Dahlgaard. SMDP (ed.), Proceedings of the 14th QMOD Conference on Quality and Service Sciences, San Sebastian, Spain, San Sebastian (Spain), pp. 225–231. URL: `https://eref.uni-bayreuth.de/6968/`.

Berekoven L, Eckert W, Ellenrieder P (1999) Marktforschung: Methodische Grundlagen und praktische Anwendung. Gabler. DOI: 10.1007/978-3-8349-8267-4.

Brand J, van Buuren S, van Mulligen EM, Timmers T, Gelsema E (1994) Multiple Imputation as a Missing Data Machine. In: Proceedings. Symposium on Computer Applications in Medical Care, pp. 303–306. ISSN: 0195-4210.

Brusch M, Baier D (2010) Analyzing the Stability of Price Response Functions: Measuring the Influence of Different Parameters in a Monte Carlo Comparison. Studies in Classification, Data Analysis, and Knowledge Organization 38:527–535. DOI: 10.1007/978-3-642-01044-6_48.

Cochran WG (1968) Errors of Measurement in Statistics. Technometrics 10(4):637–666, Taylor & Francis. DOI: 10.1080/00401706.1968.10490621.

Decker R, Wagner R (2008) Fehlende Werte: Ursachen, Konsequenzen und Behandlung. In: Herrmann A, Homburg C, Klarmann M (eds.), Handbuch Marktforschung: Methoden - Anwendungen - Praxisbeispiele. Gabler, pp. 53–79.

ECC Köln (2017) Cross-Channel – Quo Vadis?, Vol. 8, Cologne, Germany. URL: `https://www.ifhshop.de/studien/cross-channel/210/cross-channel-quo-vadis`.

Fishman GS (1996) Monte Carlo: Concepts, Algorithms, and Applications. Springer. DOI: 10.1007/978-1-4757-2553-7.

Gatty R (1966) Multivariate Analysis for Marketing Research: An Evaluation. Journal of the Royal Statistical Society. Series C (Applied Statistics) 15(3):157–172. DOI: 10.2307/2985298.

Herrmann A, Homburg C, Klarmann M (2008) Marktforschung: Ziele, Vorgehensweise und Nutzung. In: Herrmann A, Homburg C, Klarmann M (eds.), Handbuch Marktforschung: Methoden - Anwendungen - Praxisbeispiele. Gabler, pp. 3–19.

Kushwaha T, Shankar V (2013) Are Multichannel Customers Really More Valuable? The Moderating Role of Product Category Characteristics. Journal of Marketing 77:67–85. DOI: 10.2307/23487404.

Li A, Kannan PK (2014) Attributing Conversions in a Multichannel Online Marketing Environment: An Empirical Model and a Field Experiment. Journal of Marketing Research 51:40–56. DOI: 10.1509/jmr.13.0050.

Malhotra N, Birks D (2007) Marketing Research: An Applied Approach: 3rd European Edition. Pearson Education. ISBN: 978-0-134735-04-7.

Parasuraman A, Zeithaml VA, Berry LL (1985) A Conceptual Model of Service Quality and Its Implications for Future Research. Journal of Marketing 49(4):41–50. DOI: 10.2307/1251430.

Sheth JN (1971) The Multivariate Revolution in Marketing Research. Journal of Marketing 35(1):13–19. ISSN: 0022-2429, DOI: 10.2307/1250558.

Umesh UN, Peterson RA, McCann-Nelson M, Vaidyanathan R (1996) Type IV Error in Marketing Research: The Investigation of ANOVA Interactions. Journal of the Academy of Marketing Science 24(1):17–26. DOI: 10.1007/BF02893934.