Journal of Business Research 69 (2016) 992–999

Contents lists available at ScienceDirect

Journal of Business Research

Increasing sample size compensates for data problems in segmentation studies

Sara Dolnicar [a,*], Bettina Grün [b,1], Friedrich Leisch [c,2]

[a] University of Queensland, St Lucia, Brisbane 4072, Australia
[b] Johannes Kepler University, Altenbergerstraße 69, 4040 Linz, Austria
[c] University of Natural Resources and Life Sciences, Gregor-Mendel-Straße 33, 1180 Wien, Vienna, Austria

ARTICLE INFO

Article history:
Received 29 January 2015
Received in revised form 1 September 2015
Accepted 8 September 2015
Available online 12 October 2015

Keywords:
Market segmentation
Sample size
Survey data
Response bias

ABSTRACT

Survey data frequently serve as the basis for market segmentation studies. Survey data, however, are prone to a range of biases. Little is known about the effects of such biases on the quality of data-driven market segmentation solutions. This study uses artificial data sets of known structure to study the effects of data problems on segment recovery. Some of the data problems under study are partially under the control of market research companies, some are outside their control. Results indicate that (1) insufficient sample sizes lead to suboptimal segmentation solutions; (2) biases in survey data have a strong negative effect on segment recovery; (3) increasing the sample size can compensate for some biases; (4) the effect of sample size increase on segment recovery demonstrates decreasing marginal returns; and—for highly detrimental biases—(5) improvement in segment recovery at high sample size levels occurs only if additional data is free of bias.

1. Introduction

Market segmentation "is essential for marketing success: the most successful firms drive their businesses based on segmentation" (Lilien & Rangaswamy, 2002, p. 61) and "tools such as segmentation […] have the largest impact on marketing decisions" (Roberts, Kayande, & Stremersch, 2014, p. 127).

Despite the importance of market segmentation and its widespread use in industry, segmentation experts have repeatedly raised concerns about discrepancies between academic progress in the field and practical application challenges (Dibb & Simkin, 1997, 2001; Greenberg & McDonald, 1989; Wind, 1978; Young, Ott, & Feigin, 1978) pointing to an overemphasis on the data analytic aspect at the expense of developing solutions for conceptual and implementation challenges. This is particularly true for data-driven segmentation studies which construct segments by applying a statistical algorithm to several variables in the segmentation base as opposed to commonsense segmentation studies where segments result from dividing the population according to prior knowledge (Dolnicar, 2004).

One key implementation challenge companies face every time they conduct a segmentation study is that of data quality. Recent work by Coussement, Van den Bossche, and De Bock (2014) studies the extent to which data accuracy problems in databases affect the performance of direct marketing actions and segmentation solutions and investigate the robustness of different segmentation algorithms against inaccurate data. Despite the extensive body of work on survey data quality, targeted investigations of the effect of data quality on segmentation solutions have only recently started to emerge: Dolnicar and Leisch (2010) speak to the issue of segmentability of the market, first raised by Wind (1978) and Young et al. (1978), and offer a framework for data structure analysis before constructing segments.

The present study contributes a novel solution to the data quality challenge in data-driven market segmentation by investigating whether increasing the sample size can compensate for typical survey data quality problems. Specifically, the study investigates (1) the extent of the detrimental effect of data characteristics typical for survey data on the correctness of market segmentation solutions, (2) the general potential of increasing sample sizes to improve the correctness of market segmentation solutions, and (3) the potential of increased sample sizes to improve the correctness of market segmentation solutions when encountering typical survey data challenges. While it is to be assumed that larger sample sizes improve data analysis, the present study aims at deriving recommendations about the extent of required sample size increase to counteract specific kinds of survey data problems. Increasing the sample size to the required level represents—in times where survey data is increasingly collected online—a simple and affordable measure. The results of this study, therefore, will generate managerial recommendations which can easily be implemented.

[*] Corresponding author. Tel.: +61 7 3365 6702.
E-mail addresses: s.dolnicar@uq.edu.au (S. Dolnicar), bettina.gruen@jku.at (B. Grün), friedrich.leisch@boku.ac.at (F. Leisch).
[1] Tel.: +43 732 2468 6829.
[2] Tel.: +43 1 47 654 5061.

http://dx.doi.org/10.1016/j.jbusres.2015.09.004

2. Literature review

The potentially detrimental effect of bad data on market segmentation solutions has been discussed in the earliest studies on market segmentation: Claycamp and Massy (1968) point to the challenge of measuring response elasticities for segments; Young et al. (1978) argue that each segmentation problem is unique, and consequently, it is critical to select carefully who is interviewed, which questions are asked, and in which way. Wind (1978) discusses shortcomings related to segmentation bases typically used, and calls for increased efforts in determining the unit of analysis, the operational definition of dependent and independent variables, sample design, and checking of data reliability.

Several data characteristics that can reduce the validity of segmentation solutions have been known for a long time with no generally accepted solutions to reduce their impact available to date. For example, masking variables in the segmentation base which “hide or obfuscate the true structure in the data” (Brusco, 2004, p. 511) and consequently lead to inferior segmentation solutions (Carmone, Kara, & Maxwell, 1999; DeSarbo, Carroll, Clark, & Green, 1984; DeSarbo & Mahajan, 1984; Milligan, 1980) led to the development of a range of different variable selection and weighting approaches (Maugis, Celeux, & Martin-Magniette, 2009a, 2009b; Raftery & Dean, 2006; Steinley & Brusco, 2008a, 2008b).

Survey data are also known to be susceptible to response styles. Response styles result from response tendencies regardless of the content (Paulhus, 1991) and can manifest in extreme or acquiescence response styles (Baumgartner & Steenkamp, 2001). Again, different approaches have been proposed to address this problem, such as standardization of the data prior to the analysis (Schaninger & Buss, 1986) or a joint segmentation approach (Grün & Dolnicar, in press) which allows for response style and content-driven segments simultaneously.

Many other survey data characteristics—the effect of which on market segmentation analysis has not been studied to date—can also reduce the ability of a segmentation algorithm to identify naturally existing market segments or to construct managerially useful segments: sampling errors due to the decreasing willingness of people to participate in survey studies (Bednell & Shaw, 2003); respondents not answering survey questions carefully (Krosnick, 1999) or in a socially desirable way (Goldsmith, 1988; Tellis & Chandrasekaran, 2010); respondents interpreting survey questions differently, respondent fatigue leading to some low-quality responses (Johnson, Lehmann, & Horne, 1990); questionnaire items not being selected carefully (Rossiter, 2002, 2011); and the provision of binary or ordinal answer options to respondents where continuous measures could be used, which leads to less information available for data analysis (Kampen & Swyngedouw, 2000). An overview of these challenges affecting the quality of survey data is provided in Table 1.

These factors are, to some degree, in the control of the firm, because good item and scale development, questionnaire design, and fieldwork administration can reduce the incidence of survey data contamination. General recommendations for developing good survey questions are offered by Converse and Presser (1986), who recommend the use of short, simple, intelligible, and clear questions which employ straightforward language. An overview on survey sampling is given in Kalton (1983) who emphasizes the importance of reducing nonresponse because of the limitations of statistical procedures based on weighting to account for or remove nonresponse bias. Respondent fatigue, for example, can be reduced by employing procedures requiring fewer judgments from the respondents (Johnson et al., 1990). Yet, all these quality issues can never be totally excluded because, for example, some respondents always fail to take the task of completing the survey seriously. In some cases, statistical methods can be employed to account for data contaminations in the analysis, as is the case for response styles (see Grün & Dolnicar, in press; Schaninger & Buss, 1986). As a preprocessing tool to remove delinquent respondents Barge and Gehlbach (2012), for example, suggest determining the amount of satisficing of each respondent and to then assess the influence of exluding these respondents from the subsequent analysis.

Furthermore, all of the data issues discussed above can occur in situations where market characteristics already complicate the task for segmentation algorithms. For example, segment recovery is more complicated for segments of unequal size (De Craen, Commandeur, Frank, & Heiser, 2006) and for segments which overlap (Steinley, 2003) and depends on the number of segments. Such factors are entirely out of the control of the firm.

One aspect of segmentation analysis is usually *in* the control of the firm: the sample size. If shown to be effective in counteracting the detrimental effects of survey data problems, adjusting the sample size represents a simple solution. Increased sample sizes should improve solutions because market segmentation studies typically use data sets containing large numbers of variables and are thus subject to the so-called “curse of dimensionality” (Bellman, 1961).

Little research has been conducted to date to understand the effect of sample size on the correctness of segment recovery, although researchers as early as in the late 1970s noted that increasing sample size “can increase the confidence in a particular structure” and that reducing “the dimensionality can have the effect of increasing sample size” (Dubes & Jain, 1979, p. 240). Sample size recommendations for segmentation studies have, until recently, not been available at all, and the issue of sample size requirements has not been discussed as being critical—not even by authors who emphasize the importance of data quality. Only three discussions of sample size in the context of market segmentation analysis exist, none of which represent generalizable recommendations: (1) Formann (1984), in a monograph on latent class analysis, provides a sample size recommendation in the context of goodness-of-fit testing using the Chi-squared test for binary data: a minimum of two to the power of the number of variables in the segmentation base and preferably five times this number; (2) Qiu and Joe (2009) recommend that, for the purpose of generating artificial data for clustering simulations, the sample size should be at least ten times the number of variables in the segmentation base, times the number of clusters in the simplest case where clusters are of equal size; and (3) Dolnicar, Grün, Leisch, and Schmidt (2014) simulate

Table 1
Sources of quality issue problems in survey data

Problem	Description	Reference
Sampling error	Nonresponse bias occurring due to nonresponse or noncontacts, i.e., a subset of the population is not covered by the survey	e.g., Bednell and Shaw (2003)
Delinquent respondents	Satisficing respondents minimizing effort involved or respondents giving socially desirable answers	e.g., Goldsmith (1988), Krosnick (1999), Tellis and Chandrasekaran (2010)
Respondent fatigue	Respondents becoming tired of the survey task leading to a deterioration of data quality	e.g., Johnson et al. (1990)
Construct measurement and scale development	Surveys can use either single or multiple questions to measure constructs, where multi-item scales often lead to answers being highly correlated	e.g., Rossiter (2002, 2011)
Response alternatives	Choices provided to the respondents determining the measurement scale, i.e., metric, ordinal, or binary	e.g., Kampen and Swyngedouw (2000)
Response style	A systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content (Paulhus, 1991, p. 17)	e.g., Baumgartner and Steenkamp (2001)

sample size requirements for data sets of limited variability typically encountered in tourism surveys, and recommend, conservatively, that 70 times the number of variables in the segmentation base is required, arguing that results using typical tourism data do not improve beyond this point.

The two leading market segmentation textbooks (McDonald & Dunbar, 2004; Wedel & Kamakura, 2000) provide no general guidance. Wedel and Kamakura mention sample size in two very specific contexts: for the estimation of the correct segment *size* when using stratified samples and in the context of mixtures of structural equation models where a minimum sample is required to ensure that the variance–covariance matrix is positive definite. McDonald and Dunbar discuss sample size in the context of ensuring that the full heterogeneity of the market is captured and do not mention a minimum sample size to ensure valid segment recovery. Similarly, Lilien and Rangaswamy (1998)—in their very practical chapter on how to undertake market segmentation analysis—redirect the reader to market research textbooks. Market research textbooks (e.g., Iacobucci, 2013), however, fail to provide any guidance. A data analyst striving to conduct a high-quality segmentation analysis is left guessing which sample size is required. Not surprisingly, therefore, authors of applied segmentation studies demonstrate little concern about the sample size. For example, authors of segmentation studies using raw data as input which are published in the *Journal of Business Research* (Alvarez, Dickson, & Hunter, 2014; Athanassopoulos, 2000; Bhatnagar & Ghose, 2004; Floh, Zauner, Koller, & Rusch, 2014; Mariorty & Reibstein, 1986; Peterson & Sharpe, 1973; Steenkamp & Wedel, 1993; Sullivan & Miller, 1996; Theysohn, Klein, Völckner, & Spann, 2013) tend not to explain or justify the sample size used, providing support to the assumption that there is a lack of guidance and, as a consequence, a lack of concern that sample sizes may be too low.

Table 2
Artificial data set characteristics

Market characteristics (uncontrollable)	
Number of market segments	2 5 10
Equal or unequal size of market segments	All segments equally sized One small segment (5% of data) present
Separation/overlap of market segments	Misclassification rate .01 Misclassification rate .05 Misclassification rate .3
Data characteristics (controllable to some degree)	
Proportion of noisy variables (possible consequence of including variables into the segmentation base which do not contribute to the segmentation solution)	0 .25
Proportion of answers randomly replaced (possible consequence of careless responding due to factors such as respondent fatigue)	0 .25
Proportion of noisy respondents (possible consequence of sampling errors or response bias)	0 .25
Number of variables in the segmentation base (possible consequence of including too many survey questions in the survey)	4 16 32
Correlation of variables in the segmentation base (possible consequence of including redundant items in the survey)	Correlation 0 Correlation .6 Correlation of .6 among groups of 4 variables
Measurement scale (possible consequence of using a binary answer format instead of directly measuring a continuous construct)	Metric Binary
Sample size (controllable)	10 times the number of variables 30 times the number of variables 50 times the number of variables 100 times the number of variables 200 times the number of variables 500 times the number of variables

3. Method

The research question calls for an approach that allows determining under which circumstances a segmentation algorithm can identify the correct segmentation solution. A simulation study based on artificial data with known structure achieves this aim. This approach is recommended for validating clustering algorithms (see, for example, Cormack, 1971) and commonly used to compare segmentation methods (see, for example, Andrews & Currim, 2003a, 2003b; Brusco, 2004; De Soete, DeSarbo, & Carroll, 1985; Dimitriadou, Dolnicar, & Weingessel, 2002; Milligan, 1980). The market segments constructed were distinguishable through the mean distribution of the variables in the segmentation base, that is, direct differences in the observed variables are of interest and not differences in implicit response behavior. Response behavior is included as one of the variables in the segmentation base (representing heterogeneity).

A total of 2592 different artificial data scenarios emerged as a combination of market characteristics relevant to market segmentation and specific data characteristics, as shown in Table 2. For each scenario, the sample size increased as a multiple of the number of variables with six different values. To enable isolation of the impacts of these characteristics, the values for the market and data characteristics were selected to range from simple cases (with a clear segmentation structure and no data contamination) to difficult cases.

General characteristics modeled included the number of segments present in the data, the separation or extent of overlap of segments, and whether or not the segments were of equal size. In market segmentation studies, these factors represent market characteristics out of the control of the market research department. The number of market segments was selected to range from two over five to ten because two segments are typically too few to capture distinct segments whereas studies with three segments are common (e.g., Sullivan & Miller, 1996), and solutions with five segments are frequently reported (Athanassopoulos, 2000; Theysohn et al., 2013). Solutions with ten segments are hard to interpret and thus represent the upper bound of market segments typically considered. Segment overlap was selected to represent data situations with a very clear data structure (misclassification rate = 0.01) and data situations where segments overlap substantially (misclassification rate = 0.3). Market segments are generally not of equal size, but the difficulty of the segmentation problem depends on the smallest segment, thus leading to the comparison of scenarios with equal segment sizes and with one small segment (containing only 5% of the consumers, see also Athanassopoulos, 2000). In total, this led to 18 scenarios when combining the different market characteristics: number of market segments (3 different levels) × equality of market segment size (2) × separation/overlap of market segments (3).

Regarding differences in data characteristics, the amount of distortion included regarding each of the answer biases is 25%, representing substantial—but not unreasonably high—error given that Goldsmith (1988) and Tellis and Chandrasekaran (2010) find that about 40% of respondents are susceptible to socially desirable response bias and agree to at least one out of two questions where agreement is impossible.

Survey data characteristics modeled in the artificial data sets included:

(1) The proportion of noisy variables (as pointed out by Brusco, 2004; Carmone et al., 1999; DeSarbo et al., 1984; DeSarbo & Mahajan, 1984; Milligan, 1980). Noisy variables are variables which contribute nothing to the segmentation solution. Such

variables can result from the inclusion of irrelevant survey questions. In the artificial data, they manifest as an entire column that is unrelated to the segmentation solution. Data scenarios contain either no or 25% noisy variables.

(2) Partially noisy variables contain meaningful information for the segmentation solution, but include unrelated random values, which can result from respondent fatigue (Johnson et al., 1990). Partially noisy variables manifest in the artificial data sets as meaningful variables with a pre-specified proportion of entries randomly replaced. Data scenarios contain either no or 25% randomly replaced entries.

(3) Noisy respondents are not informative for the segment structure of the data. They can result from suboptimal sampling or delinquent respondents (Krosnick, 1999) and manifest in the artificial data sets as data rows unrelated to the segmentation solution. The artificial data sets either contain no or 25% noisy respondents.

(4) The number of variables in the segmentation base (Dolnicar et al., 2014; Formann, 1984), over which the market research department has some influence. Four, sixteen, and 32 variables are used which extends the range used in previous studies (ten in Athanassopoulos, 2000; 14 in Mariorty & Reibstein, 1986).

(5) The extent to which the variables in the segmentation base correlate and whether the correlation occurs among all variables or among groups of variables, which is typically the case with scales of items measuring the same construct (Rossiter, 2011). The case of no correlation reflects the situation where correlated variables are excluded from the segmentation base (as in Athanassopoulos, 2000), while groups of correlated items occur if factors measured using multi-items scales are included with their items (as in Sullivan & Miller, 1996). The artificial data sets model three scenarios: no correlation, a Pearson correlation of .6 among all variables, and a correlation of .6 among groups of four variables that represent items measuring the same construct.

(6) The metric or binary measurement scale of the data (Kampen & Swyngedouw, 2000). The artificial data was generated on a metric scale. Binary versions were derived from these by assigning zeros and ones, depending on whether the metric variable was negative or positive.

These settings led to 144 data characteristics: proportion of noisy variables (2) × proportion of random answers (2) × proportion of noisy respondents (2) × number of variables (3) × correlation of variables (3) × measurement scale (2). The total of 2592 scenarios emerges from 18 market characteristic scenarios × 144 different data characteristic scenarios. Finally, data sets that represent combinations of all the above factors were created for sample sizes between 10 and 500 times the number of variables. The choice of sample sizes was informed by Formann (1984); Qiu and Joe (2009), and Dolnicar et al. (2014).

Computations were done in the statistical software environment R (R Core Team, 2014) using the flexmix package (Grün & Leisch, 2008; Leisch, 2004) for fitting mixture models. Metric data were generated by drawing from finite mixtures of multivariate normal distributions. The underlying mixture model, as well as the full data set (containing the maximum number of observations required), was generated to ensure that they corresponded to the required characteristics for the metric data, based on ideas from Maitra and Melnykov (2010) and Melnykov, Chen, and Maitra (2012). The average misclassification rate over all classes for that data set if the true mixture model were known needed to be close to the pre-specified value; that is, within a tolerance of .01. The noisy variables, the randomly replaced answers, and noisy respondents were generated by drawing independently from normal distributions with mean values and standard deviations similar to the data informative of the segmentation solution. Metric data was binarized to obtain the binary data using zero as the threshold value. Note that the binary data emerges from the same metric data, but only contains the information if the observed value is positive or negative.

For each of the 2592 data scenarios and six sample size settings, finite mixture models derived estimates for 20 replication data sets. The best solution across ten random starting points was used to avoid local optima. For finite mixture models for metric data, the components come from a multivariate normal distribution with a full variance–covariance matrix, and for binary data, the component distribution is the product of Bernoulli distributions assuming conditional independence given segment membership. To avoid degenerate solutions, a regularization was undertaken, corresponding to adding information with weight equivalent to one observation to each component. In the metric case, a diagonal variance–covariance matrix containing the empirical variances was added as information to the variance–covariance matrices (compare, for example, Fraley & Raftery, 2007). In the binary case, the information added was with respect to the success probabilities and equal to the same probabilities of success than the empirically observed ones (after having added one success and one failure to the observed data; compare, Galindo-Garre & Vermunt, 2006).

The correctness of the resulting segmentation solution served as the dependent variable. The result for a simulated respondent is correct if the segment membership resulting from the analysis is the same as the known, true membership. The adjusted Rand index (Rand, 1971; Hubert & Arabie, 1985) serves as the measure of performance because of its invariance to the cluster labeling; it corresponds to comparing the partition of the respondents based on the segmentation solution to the partition induced by the true segment memberships. The adjusted Rand index takes into account the proportion of pairs of respondents that are assigned to the same or different clusters across the two partitions, corrected for agreement by chance. A value of 1 indicates the exact same solution across two partitions; a value of 0 indicates that the agreement across the two partitions could also occur by chance. The benchmark adjusted Rand index emerged from an additional testing data set containing 5000 respondents generated from the true model without random answers or noisy respondents, but with noisy variables. The benchmark data set was the same for all different sample sizes for a given scenario. Again, the binary data resulted from binarizing the metric one.

For each data set, a number of finite mixture models was calculated containing the correct number of clusters and up to two numbers of clusters lower or higher. BIC values served as criterion for selecting the number of clusters. Rand indices were then only calculated for the selected number of clusters. This approach mimics that in reality the number of clusters is not known in advance.

Linear models with the adjusted Rand index as dependent variable (after averaging over replications) and categorical variables as regressors assisted in identifying significant factors. For a given sample size, the main effect (that is, the marginal effect) of each specific factor was determined. Results show—for a fixed sample size—how the performance varies depending on the difficulty of the task and the quality of the data. Holding task difficulty and data quality fixed, the average adjusted Rand index values across sample sizes indicate the change in performance if the sample size is increased. Ultimately, each fitted model includes main effects for sample size and each factor, as well as interaction effects between sample size and each factor. The interaction effect was probed by comparing the full model with the model without this interaction effect using *F*-tests in order to check if the effect of increasing sample size differs for the different levels of the factors. The marginal effects of each factor resulted for different sample sizes with the effects of other factors averaged out. Results from the linear models formed the basis of estimations of required sample size increases using linear interpolation to compensate for the full range of survey data issues considered.

Table 3
Regression coefficients ($n = 10^{*}d$)

	Estimate	Std. Est.	Std. Error	t value	p-value
(Intercept)	0.175		0.004	49.653	<0.001
5 segments compared to 2	0.110	0.358	0.005	22.095	<0.001
10 segments compared to 5	−0.028	−0.093	0.005	−5.726	<0.001
Small segment present	−0.010	−0.035	0.004	−2.494	0.013
Misclassification rate: .05 compared to .01	−0.082	−0.267	0.005	−16.479	<0.001
Misclassification rate: .3 compared to .05	−0.079	−0.258	0.005	−15.937	<0.001
Noisy variables .25	−0.004	−0.015	0.004	−1.103	0.270
Random .25–0	−0.039	−0.136	0.004	−9.723	<0.001
Noisy respondents .25	−0.023	−0.080	0.004	−5.681	<0.001
16 variables compared to 4	−0.078	−0.254	0.005	−15.672	<0.001
32 variables compared to 16	0.001	0.002	0.005	0.133	0.894
Correlation of .6	−0.103	−0.334	0.005	−20.646	<0.001
Groups of variables with correlation of .6	−0.081	−0.262	0.005	−16.202	<0.001
Measurement binary	0.000	0.000	0.004	0.004	0.997

Std. Est.: standardized coefficient estimate indicating the change in standard deviations of the dependent variable per standard deviation increase of the independent variable; Std. Error: standard error; $R^2 = 0.49$.

4. Results

4.1. The effect of data characteristics on segment recovery

To test the effect of data characteristics on segment recovery, all simulated market and data characteristics serve as independent variables and segment recovery for the smallest of the sample sizes serves as dependent. The combination of scenarios covers the full range from easy to difficult segmentation tasks with the majority including at least one challenging factor leading only to a modest overall adjusted Rand index for the smallest sample size of .113. This average is observed if the number of components is estimated by BIC; but BIC points to the correct number of clusters in only 7% of cases. Had the correct number of clusters been chosen each time, the average adjusted Rand index would have been .143. Note, however, that—across all scenarios—95% of the adjusted Rand index values lie between .000 and .496.

A linear regression model assists in determining the effect of data characteristics on segment recovery. The adjusted Rand indices serve as the dependent variable. This model leads to baseline values for the effects of data problems in cases where only a limited amount of data is available. Results are provided in Table 3.

The intercept captures the average adjusted Rand index for the baseline setting with two segments, four variables, a misclassification rate of .01, no correlation between variables, all segments being of equal size, no noisy variables, no noisy respondents, no randomly replaced answers, and using metric data. The regression coefficients reflect the change in the average adjusted Rand index values if the explanatory variables change (value indicated in the label). For explanatory variables with three different levels the regression coefficients give the effect of changing between successive levels taking the ordinal nature of the explanatory variables into account. Thus, the coefficient for “10 segments compared to 5” indicates how much the average adjusted Rand index changes if 10 instead of 5 segments are fitted and to determine the effect on the average adjusted Rand index when changing from 2 to 10 segments, the coefficients for “5 segments compared to 2” and “10 segments compared to 5” need to be combined.

Most of the factors lead to a significant reduction in segment recovery. In terms of market characteristics, the most influential factors are overlap between segments and the presence of a small segment, as opposed to all segments being of equal size. In addition, the presence of more than two segments in the market increases the likelihood of correctly identifying the true segment membership. Unsurprisingly so, given that solutions with larger numbers of segments can better capture the data space.

Among the data characteristics, correlation between variables has the strongest detrimental effect on segment recovery, followed by large numbers of variables being present in the segmentation base, the substitution of 25% of the data with random values, and the presence of noisy respondents. No significant difference in segment recovery occurred for metric data as opposed to binary data and if noisy variables were present or not.

Clearly, data characteristics common to survey data can have detrimental effects on segmentation solutions computed using such data.

4.2. Improving segment recovery by increasing sample size—market characteristics

Generally, increasing sample size leads to an increase in the average adjusted Rand index. For example, when averaging across all data scenarios, an increase in sample size from 10 times to 500 times the number of variables increases the average adjusted Rand index from .113 to

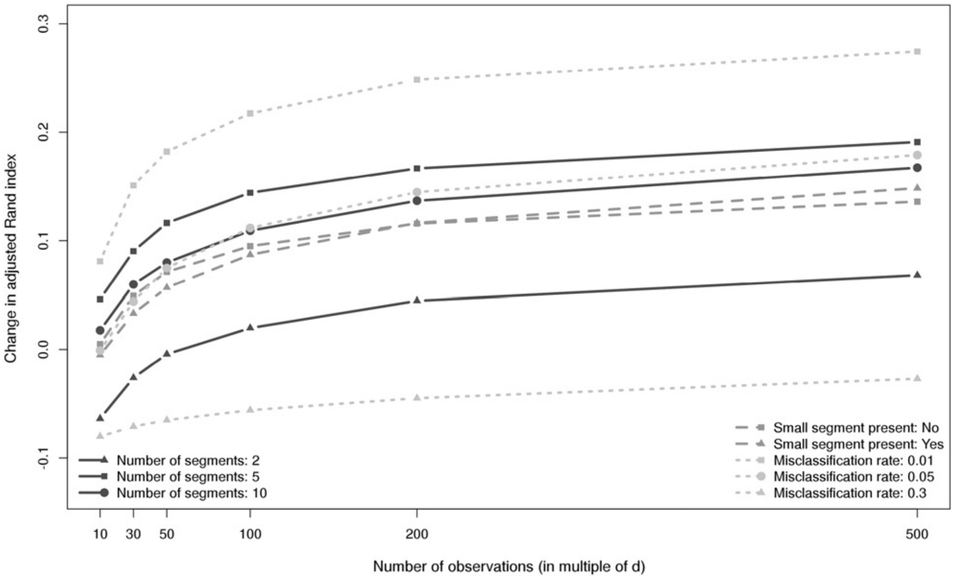

Fig. 1. Segment recovery effect of sample size increase (market characteristics).

Table 4
Model comparison results of the full model with the model where the interaction effect between the data or market characteristics and sample size is omitted

	res. DF	RSS	ΔDF	ΔRSS	*F* value	*p*-value
Full model	15,468	234.36				
Omitted interaction effect with sample size and						
Number of segments	15,478	234.49	10	0.13	0.87	0.564
Small segment	15,473	234.74	5	0.38	4.95	<0.001
Misclassification rate	15,478	241.74	10	7.38	48.71	<0.001
Noisy variables	15,473	234.37	5	0.00	0.06	0.998
Random	15,473	234.42	5	0.06	0.82	0.537
Noisy respondents	15,473	234.50	5	0.13	1.76	0.118
Number of variables	15,478	234.76	10	0.40	2.64	0.003
Correlation structure	15,478	241.96	10	7.59	50.12	<0.001
Measurement binary	15,473	237.44	5	3.08	40.65	<0.001

res. DF: residual degrees of freedom; RSS: residual sum of squares; ΔDF: difference of degrees of freedom; ΔRSS: difference of residual sum of squares.

.256. The difference in average adjusted Rand index between the solutions where the true number of components is known or estimated by BIC decreases when sample size is increased from .030 to .005 for smallest to largest sample size. Thus, not knowing the correct number of clusters is less critical with large sample sizes. Also, when sample size increases, more high values of the adjusted Rand index are achieved. Across all data scenarios the adjusted Rand index range for 95% of observations lies between .000 and .496 when the sample size is only 10 times the number of variables. When a sample of 500 times the number of variables is available, this range lies between .000 and .754.

Fig. 1 illustrates the effect of sample size (*x* axis) on segment recovery (*y* axis) in market situations of increased complexity, including the presence of more segments (solid, dark gray lines), the presence of a niche segment (long dashed, gray lines) and different separation values between segments (short dashed, light gray lines). Each line joining the dots represents a certain value of the variable characterizing the segmentation task. The lines join the fitted change in adjusted Rand indices averaged over all other variables varied in the simulation study for different sample sizes.

The primary effect of increasing sample size—that of improving segment recovery—occurs across all factors known to affect the complexity of the segmentation problem. Also, the diminishing marginal returns of sample size increase can be observed in all cases. Sample size increase has the strongest effect from 10 to 30 times the number of variables with a leveling off after 100 times the number of variables. The results in Fig. 1 show the model where the incremental effect of sample size is allowed to differ for the different levels of overlap. This model is compared to a restricted model where the effect of sample size would correspond to parallel lines in this figure, i.e., the incremental effect of sample size is the same regardless of the overlap, but overlap leads to different offsets. Results are given in Table 4, where the goodness-of-fit test between the full and restricted models indicates that the full model fits significantly better (*p*-value < 0.001). Thus, the effect of increasing sample size differs depending on the overlap of segments with the effect of an increase in sample size being larger for data sets which have a clearer structure. This difference in sample size effect also occurs if a niche segment is present or not, but is not significant for the number of segments (see Table 4).

4.3. Improving segment recovery by increasing sample size—data characteristics

Figs. 2 and 3 illustrate the effect of increasing sample size (*x* axis) on segment recovery (*y* axis) for typical survey data characteristics. All data problems negatively affect segment recovery: the lines with triangles—representing the highest level of each data challenge—are consistently located lower than the lines with squares and circles—representing either the absence or a lower level of a data-related challenge.

For all data-related challenges investigated, increasing the sample size can, to a limited extent, counteract negative impacts, although increasing sample size alone can never lead to absolutely correct segment recovery. For example, for the artificial data sets studied, increasing the sample from 10·d to 12·d (where d is the number of variables) can compensate for 25% of noisy variables, increasing the sample from 10·d to 22·d can compensate for 25% of noisy respondents, increasing the sample from 10·d to 32·d can compensate for 25% of random data, and increasing the sample from 10·d to 29·d can compensate for all survey questions being correlated compared to only groups of correlated questions. For a sample size of 10·d having data collected on a binary scale has approximately the same performance as if the data were measured on a continuous scale.

No increase in the sample size can compensate for all variables having a correlation of 60%, or having groups of variables with 60% correlation in the segmentation base, with the performance always being

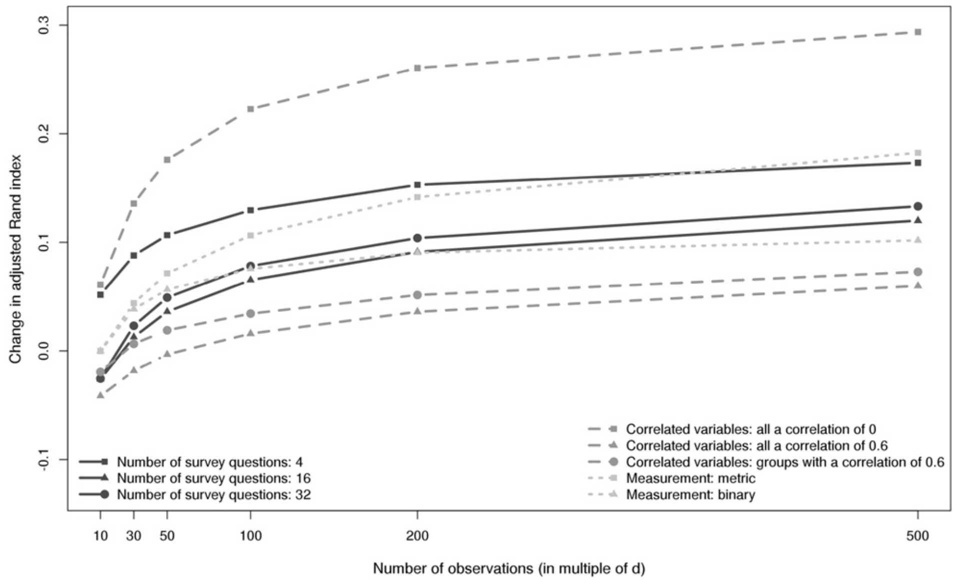

Fig. 2. Segment recovery effect of sample size increase (data characteristics: number of survey question, correlation between variables, measurement scale).

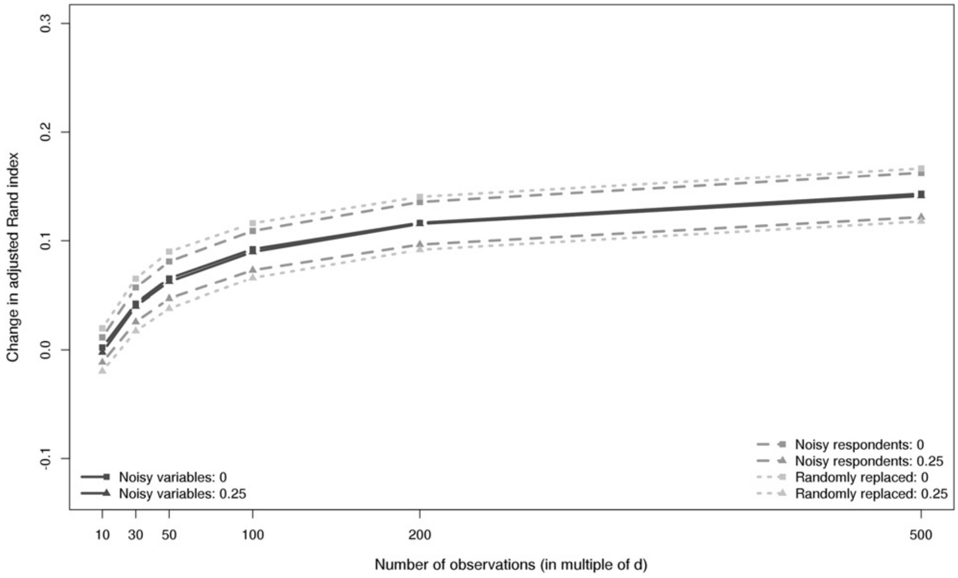

Fig. 3. Segment recovery effect of sample size increase (data characteristics: noisy variables, noisy respondents, random replacement).

much worse than for uncorrelated variables. Similarly, the performance of a small set of good questions (that is, four questions) cannot be achieved by increasing the sample size if the number of questions is higher (that is, 16 or 32).

Figs. 2 and 3 also indicate the diminishing marginal returns of sample size increase. However, the improvement attained by adding observations also differs across the different factors and their levels (see Table 4). Additional observations are of higher value if collected at metric level and if uncorrelated. Thus, improvement through sample size increase is much more pronounced if the additional observations are free of bias.

5. Conclusion

The results of an extensive simulation study with artificial data sets of known structure lead to the following key conclusions:

(a) Insufficient sample size of the segmentation base can have serious negative consequences on segment recovery.
(b) Segment recovery can be substantially improved by increasing sample size from 10 to 30 times the number of variables. This improvement levels off subsequently, but is still noticeable up to a sample size of approximately 100 times the number of variables; whereas the effect on correctness of the segmentation solutions is much smaller thereafter.
(c) Several common problems that affect the quality of survey research (such as respondents' fatigue, the inclusion of irrelevant survey questions, and/or imperfect sampling) can negatively affect segment recovery. If at all possible, and as recommended many decades ago by Wind (1978), such quality issues should be prevented through high-quality research design and data collection at the early stages of a segmentation study.
(d) When data quality issues are unavoidable, increasing the sample size represents a simple measure to compensate—not fully, but to some degree—for the detrimental effects caused by poor data quality.
(e) The effect of sample size increase on segment recovery demonstrates decreasing marginal returns.
(f) Improvement in segment recovery at high sample size levels occurs only if additional data is free of bias.
(g) And, supporting the finding of Coussement et al. (2014), good data is critical to good segmentation studies.

The implications are of benefit to industries that rely on market segmentation studies, thus contributing to the reduction of the “discrepancy between academic developments and real-world practice” first raised by Wind (1978) and subsequently reinforced by others (Dibb & Simkin, 1997, 2001; Greenberg & McDonald, 1989; Young et al., 1978). Specifically, results from the present study lead to a number of recommendations for managers involved in market segmentation: the quality of data is extremely important for segmentation studies and good care needs to be taken in the development of survey studies for market segmentation that the least possible data contamination by biases occurs. Some biases, of course, are inevitable. It is therefore recommendable to err on the side of caution and collect slightly larger samples. Most critically, however, in terms of sample size: sample affordability should not determine sample size; rather, the sample size should be determined in view of the number of variables used as the segmentation base and in view of the extent of survey data contamination that is expected to be present in the data. While collecting data from additional respondents costs additional money, it is still more resource-efficient than risking the validity of the entire segmentation study by trying to save on a few additional responses, especially given that working with a larger than necessary sample size does not have negative effects on the segmentation solution. The fact that market research is increasingly conducted online helps reduce the cost of additional respondents.

In terms of limitation, this study is based on artificial data, which was necessary in order to have available a criterion for segment recovery. Note, however, that the kind of study conducted here can only be conducted using artificial data given that the true segment structure is not known for empirical data sets. So, in this case, the artificial data sets do not represent the next best thing after real empirical survey data. Rather, the artificial data sets represent the only option given the aims of this study. It is expected that the generalizability of the findings presented in this study to real empirical data sets is high given that the artificial data sets were designed to specifically model challenging characteristics known to occur in empirical survey data. Future work could focus on developing ways to determine—before segmentation analysis—the extent to which an empirical data set is affected by any data characteristic problems.

Acknowledgements

We are grateful to the Australian Research Council (ARC) and the Austrian Science Fund (FWF) for supporting this research through projects DP110101347, LX0881890, and V170-N18. Computations were partially run on the Vienna Scientific Cluster (VSC) under approval number 70419. We thank Martin Natter and Dominik Ernst for feedback on previous versions of the manuscript.

References

Alvarez, C.M.O., Dickson, P.R., & Hunter, G.K. (2014). The four faces of the Hispanic consumer: An acculturation-based segmentation. *Journal of Business Research*, *67*(2), 108–115.

Andrews, R.L., & Currim, I.S. (2003a). Recovering and profiling the true segmentation structure in markets: and empirical investigation. *International Journal of Research in Marketing*, *20*(2), 177–192.

Andrews, R.L., & Currim, I.S. (2003b). Retention of latent segments in regression-based marketing models. *International Journal of Research in Marketing*, *20*(4), 315–321.

Athanassopoulos, A.D. (2000). Customer satisfaction cues to support market segmentation and explain switching behavior. *Journal of Business Research*, *47*, 191–207.

Barge, S., & Gehlbach, H. (2012). Using the Theory of Satisficing to Evaluate the Quality of Survey Data. *Research in Higher Education*, *53*(2), 182–200.

Baumgartner, H., & Steenkamp, J. -B.E.M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156.

Bednell, D.H.B., & Shaw, M. (2003). Changing response rates in Australian market research. *Australasian Journal of Market Research*, *11*(1), 31–41.

Bellman, R.E. (1961). *Adaptive control processes.* Princeton: Princeton University Press.

Bhatnagar, A., & Ghose, S. (2004). Segmenting consumers based on the benefits and risks of Internet shopping. *Journal of Business Research*, *57*(12), 1352–1360.

Brusco, M. (2004). Clustering binary data in the presence of masking variables. *Psychological Methods*, *9*(4), 510–523.

Carmone, F.J., Jr., Kara, A., & Maxwell, S. (1999). HINoV: A new model to improve market segment definition by identifying noisy variables. *Journal of Marketing Research*, *36*(4), 501–509.

Claycamp, H.J., & Massy, W.F. (1968). A theory of market segmentation. *Journal of Marketing Research*, *5*(4), 388–394.

Converse, J.M., & Presser, S. (1986). Survey questions—Handicrafting the standardized questionnaire. *Sage series on quantitative applications in the social sciences, number 63*. Newbury Park: Sage.

Cormack, R.M. (1971). A review of classification. *Journal of the Royal Statistical Society. Series A (General)*, *134*(3), 321–367.

Coussement, K., Van den Bossche, F.A., & De Bock, K.W. (2014). Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees. *Journal of Business Research*, *67*(1), 2751–2758.

De Craen, S., Commandeur, J.J.F., Frank, L.E., & Heiser, W.J. (2006). Effects of group size and lack of sphericity on the recovery of clusters in K-means cluster analysis. *Multivariate Behavioral Research*, *41*(2), 127–145.

De Soete, G., DeSarbo, W.S., & Carroll, J.D. (1985). Optimal variable weighting for hierarchical clustering: An alternative least-squares algorithm. *Journal of Classification*, *2*(1), 173–192.

DeSarbo, W.S., & Mahajan, V. (1984). Constrained classification: The use of a priori information in cluster analysis. *Psychometrika*, *49*(2), 187–215.

DeSarbo, W.S., Carroll, J.D., Clark, L.A., & Green, P.E. (1984). Synthesized clustering: A method for amalgamating alternative clustering bases with differential weighting of variables. *Psychometrika*, *49*(1), 57–78.

Dibb, S., & Simkin, L. (1997). A programme for implementing market segmentation. *Journal of Business and Industrial Marketing*, *12*(1), 51–65.

Dibb, S., & Simkin, L. (2001). Market segmentation: diagnosing and treating the barriers. *Industrial Marketing Management*, *30*(8), 609–625.

Dimitriadou, E., Dolnicar, S., & Weingessel, A. (2002). An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, *67*(1), 137–160.

Dolnicar, S. (2004). Beyond "commonsense segmentation": A systematics of segmentation approaches in tourism. *Journal of Travel Research*, *42*(3), 244–250.

Dolnicar, S., & Leisch, F. (2010). Evaluation of structure and reproducibility of cluster solutions using the bootstrap. *Marketing Letters*, *21*(1), 83–101.

Dolnicar, S., Grün, B., Leisch, F., & Schmidt, K. (2014). Required sample sizes for data-driven market segmentation analyses in tourism. *Journal of Travel Research*, *53*(3), 296–306.

Dubes, R., & Jain, A.K. (1979). Validity studies in clustering methodologies. *Pattern Recognition*, *11*, 235–254.

Floh, A., Zauner, A., Koller, M., & Rusch, T. (2014). Customer segmentation using unobserved heterogeneity in the perceived-value–loyalty–intentions link. *Journal of Business Research*, *67*(5), 974–982.

Formann, A.K. (1984). *Latent Class Analyse Einführung in die Theorie und Anwendung [Latent class analysis—Introduction to theory and application].* Weinheim: Beltz.

Fraley, C., & Raftery, A.E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, *24*(2), 155–181.

Galindo-Garre, F., & Vermunt, J.K. (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. *Behaviormetrika*, *33*(1), 43–59.

Goldsmith, R.E. (1988). Spurious response error in a new product survey. *Journal of Business Research*, *17*(3), 271–281.

Greenberg, M., & McDonald, S.S. (1989). Successful needs/benefits segmentation: A user's guide. *The Journal of Consumer Marketing*, *6*(3), 29–36.

Grün, B., & Dolnicar, S. (2015). Response style corrected market segmentation for ordinal data. *Marketing Letters*http://dx.doi.org/10.1007/s11002-015-9375-9 (in press).

Grün, B., & Leisch, F. (2008). FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, *28*(4), 1–35.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

Iacobucci, D. (2013). *Marketing models: Multivariate statistics and marketing analytics.* Mason: South-Western.

Johnson, M.D., Lehmann, D.R., & Horne, D.R. (1990). The effects of fatigue on judgments of interproduct similarity. *International Journal of Research in Marketing*, *7*(1), 35–43.

Kalton, G. (1983). Introduction to survey sampling. *Sage series on quantitative applications in the social sciences, number 35*. Newbury Park: Sage.

Kampen, J., & Swyngedouw, M. (2000). The ordinal controversy revisited. *Quality and Quantity*, *34*(1), 87–102.

Krosnick, J.A. (1999). Survey research. *Annual Review of Psychology*, *50*(1), 537–567.

Leisch, F. (2004). FlexMix: a general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, *11*(8), 1–18.

Lilien, G.L., & Rangaswamy, A. (1998). *Marketing engineering: Computer-assisted marketing analysis and planning.* Reading: Addison-Wesley.

Lilien, Gary L., & Rangaswamy, A. (2002). (2nd edition). New Jersey: Marketing Engineering Pearson Education Upper Saddle River.

Maitra, R., & Melnykov, V. (2010). Simulating data to study performance of finite mixture modeling and clustering algorithms. *Journal of Computational and Graphical Statistics*, *19*(2), 354–376.

Mariorty, R.T., & Reibstein, D.J. (1986). Benefit segmentation in industrial markets. *Journal of Business Research*, *14*(6), 463–486.

Maugis, C., Celeux, G., & Martin-Magniette, M.L. (2009a). Variable selection for clustering with Gaussian mixture models. *Biometrics*, *65*(3), 701–709.

Maugis, C., Celeux, G., & Martin-Magniette, M.L. (2009b). Variable selection in model-based clustering: a general variable role modeling. *Computational Statistics & Data Analysis*, *53*(11), 3872–3882.

McDonald, M., & Dunbar, I. (2004). *Market segmentation: How to do it, how to profit from it.* Oxford: Elsevier Butterworth-Heinemann.

Melnykov, V., Chen, W. -C., & Maitra, R. (2012). MixSim: an R package for simulating data to study performance of clustering algorithms. *Journal of Statistical Software*, *51*(12), 1–25.

Milligan, G.W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, *45*(3), 325–342.

Paulhus, D.L. (1991). Measurement and control of response bias. In J.P. Robinson, P.R. Shaver, & L.S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Peterson, R.A., & Sharpe, L.K. (1973). Market segmentation: Product usage patterns and psychographic configurations. *Journal of Business Research*, *1*(1), 11–20.

Qiu, W., & Joe, H. (2009). clusterGeneration: Random cluster generation (with specified degree of separation). R package version 1.2.7.

R Core Team (2014). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Raftery, A.E., & Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, *101*(473), 168–178.

Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850.

Roberts, J.H., Kayande, U., & Stremersch, S. (2014). From academic research to marketing practice: Exploring the marketing science value chain. *International Journal of Research in Marketing*, *31*, 127–140.

Rossiter, J.R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing*, *19*, 305–335.

Rossiter, J.R. (2011). *Measurement for the social sciences: The C-OAR-SE method and why it must replace psychometrics.* New York: Springer.

Schaninger, C.M., & Buss, W.C. (1986). Removing response-style effects in attribute-determinance ratings to identify market segments. *Journal of Business Research*, *14*(3), 237–252.

Steenkamp, J.B.E., & Wedel, M. (1993). Fuzzy clusterwise regression in benefit segmentation: Application and investigation into its validity. *Journal of Business Research*, *26*(3), 237–249.

Steinley, D. (2003). Local optima in K-means clustering: What you don't know may hurt you. *Psychological Methods*, *8*(3), 294–304.

Steinley, D., & Brusco, M.J. (2008a). A new variable weighting and selection procedure for K-means cluster analysis. *Multivariate Behavioral Research*, *43*(1), 77–108.

Steinley, D., & Brusco, M.J. (2008b). Selection of variables in cluster analysis. *Psychometrika*, *73*(1), 125–144.

Sullivan, M.K., & Miller, A. (1996). Segmenting the informal venture capital market: Economic, hedonistic, and altruistic investors. *Journal of Business Research*, *36*(1), 25–35.

Tellis, G.J., & Chandrasekaran, D. (2010). Extent and impact of response biases in cross-national survey research. *International Journal of Research in Marketing*, *27*(4), 329–341.

Theysohn, S., Klein, K., Völckner, F., & Spann, M. (2013). Dual effect-based market segmentation and price optimization. *Journal of Business Research*, *66*(4), 480–488.

Wedel, M., & Kamakura, W. (2000). *Market segmentation—Conceptual and methodological foundations* (2nd ed.). Boston: Kluwer.

Wind, Y. (1978). Issues and advances in segmentation research. *Journal of Marketing Research*, *15*(3), 317–337.

Young, S., Ott, L., & Feigin, B. (1978). Some practical considerations in market segmentation. *Journal of Marketing Research*, *15*(3), 405–412.