2007

Accepted standards undermining the validity of tourism research

Sara Dolnicar
University of Wollongong, s.dolnicar@uq.edu.au

Publication Details

Dolnicar, S, Accepted standards undermining the validity of tourism research, in Woodside, A (ed.) Advances in Culture, Tourism and Hospitality Research, volume 1, 2007, Emerald Group, 131-182.

Accepted standards undermining the validity of tourism research

Abstract
This paper draws attention to accepted measurement and research method standards in empirical research on tourism. Some standards stand out because they are superior to alternative approaches. However, many have emerged because the measurements and methods used in prior work were assumed to be optimal (or at least valid) for solving particular problems. Unfortunately this assumption is inaccurate. Yet the reviewing process favors the use of such standards (often without demanding evidence) over the introduction of novel approaches, even if these are justified. This paper focuses on three accepted standards in empirical tourism research which have the potential to undermine the validity of findings: the uncritical use of ordinal multi-category answer formats, the derivation of cross-cultural comparisons that do not consider cultural response biases resulting from response styles, and the standard step-wise procedure used in data-driven market segmentation. This paper describes the potential dangers of these standard approaches and makes recommendations for researchers to consider before choosing to adopt any of the above approaches.

Keywords
accepted standards; ordinal answer formats; cross-cultural studies; data-driven market segmentation.

Disciplines
Business | Social and Behavioral Sciences

Publication Details
Dolnicar, S, Accepted standards undermining the validity of tourism research, in Woodside, A (ed.) Advances in Culture, Tourism and Hospitality Research, volume 1, 2007, Emerald Group, 131-182.

Accepted Standards Undermining the Validity of Tourism Research

Sara Dolnicar, University of Wollongong

This research was supported by the Australian Research Council (through grants DP0557257 and LX0559628). The author thanks Katrina Matus for her help as a research assistant for this study and Friedrich Leisch, Bettina Grün and John Rossiter for their feedback on various components of the study.

Please send correspondence to:

Sara Dolnicar

School of Management and Marketing

marketing research innovation centre (mric)

University of Wollongong

Wollongong, NSW 2522, Australia

Phone +61 2 4221 3862

Fax +61 2 4221 4120

Email sara_dolnicar@uow.edu.au

Abstract

This paper draws attention to accepted measurement and research method standards in empirical research on tourism. Some standards stand out because they are superior to alternative approaches. However, many have emerged because the measurements and methods used in prior work were assumed to be optimal (or at least valid) for solving particular problems. Unfortunately this assumption is inaccurate. Yet the reviewing process favors the use of such standards (often without demanding evidence) over the introduction of novel approaches, even if these are justified.

This paper focuses on three accepted standards in empirical tourism research which have the potential to undermine the validity of findings: the uncritical use of ordinal multi-category answer formats, the derivation of cross-cultural comparisons that do not consider cultural response biases resulting from response styles, and the standard step-wise procedure used in data-driven market segmentation. This paper describes the potential dangers of these standard approaches and makes recommendations for researchers to consider before choosing to adopt any of the above approaches.

Key words: accepted standards; ordinal answer formats; cross-cultural studies; data-driven market segmentation.

1 INTRODUCTION

Much research accepts the approaches and techniques used and published in the past as established, valid procedures. The practise of citing several authors of prior studies (the more the better) who use a certain approach or technique, instead of explaining the reason for choosing this procedure and justifying why it is the best solution for the

problem is a dominating logic in the research community. This insight is not new. Thomas Kuhn (1970, p. 6) refers to this phenomenon as the 'tradition-bound activity of normal science' and defines 'scientific revolutions' as 'tradition-shattering complements' that move science forward.

The practise of uncritically following other authors' approaches (or in a seemingly uncritical manner) is prevalent in tourism research. While not uncommon in other area of research, it is nevertheless undesirable, and has several effects that contradict the fundamental principles of scientific research, by: (1) tolerating the uncritical use of approaches and techniques; (2) not providing incentives to introduce new approaches and techniques; and (3) discouraging the use of new approaches. Authors who introduce new approaches must justify their deviation from the norm in the most rigorous fashion; whereas uncritical acceptance of the current standard does not require extensive explanation in the reviewing process.

In the increasingly competitive research market a new researcher who needs to build their CV and rationally analyzes how the acceptance rate of their publications can be maximized is likely to conclude that: (1) they can spend significantly less effort to conduct a study if they follow the established standards in a field of research because no justification will be required for the choice of method, measurement technique, or data analytic approach; and (2) using the established standard is much safer because the risk of rejection will be significantly lower. In sum, uncritically following the approaches taken by authors in the past, and reviewers' willingness to accept citations (rather than justifications) as reasons for adopting a particular

approach lead to stagnation, rather than development and innovation of a field of research — in this case empirical tourism research.

This paper describes three common aspects of empirical tourism research that lead to the development of emerging standards: the uncritical use of ordinal multi-category answer formats in data collection, cross-cultural comparisons ignoring response style bias, and the use of a particular step-wise procedure in data-driven market segmentation. For each of these three aspects: (1) a review of the discussion in the broader scientific community is presented; (2) hypotheses are formulated about the precise nature of the respective accepted standard within tourism research; (3) empirical evidence is provided to support or reject these hypotheses; (4) potential dangers resulting from the uncritical use of outlined standard approaches are discussed; and (5) a series of questions or aspects is provided, which may be helpful to empirical tourism researchers in deciding whether or not to adopt these standard approaches in their future studies.

2 ANSWER FORMATS

In empirical social sciences where the responses of subjects to the researcher's questions form the basis of theoretical (or practical) insight, the *question* is the scientist's measurement instrument. Where an atmospheric chemist uses a thermometer or a barometer, the social scientist uses a question. An atmospheric chemist would never consider using an uncalibrated or untested thermometer or barometer to measure temperature or air pressure. The same should be true for an empirical social scientist. The question asked is the main measurement instrument,

and it has to be carefully chosen or developed to ensure that it reliably measures what it is supposed to measure.

How to best ask questions to get valid results is therefore an issue of interdisciplinary interest, and is of fundamental importance to any research field in which the collection of primary data is required to investigate a research question. The importance of question formulation has been acknowledged by social scientists in many areas since the early 19th century. Consequently, a vast body of literature exists in psychology, sociology, psychometrics, and marketing, which investigates the effects of various aspects of questionnaire design on the validity of results. Some researchers go as far as to question whether survey responses can at all be viewed as valid measurements. Feldman and Lynch (1988, p.431), for instance, "show how observed correlations among beliefs, attitudes, intentions and behaviours may be affected by the process of measurements". The underlying argument is that respondents frequently do not have cognitions which are easily accessible to answer a question. Instead, they compute or create them in response to questions. Such a computation process is strongly influenced by the order, context and wording of the question in the survey. Earlier responses will have an influence on how later questions will be answered. Feldman and Lynch's study demonstrates clearly that the validity of measurement in the social sciences is by no means a given. Instead a significant amount of effort in pre-analysis of questions for specific populations is needed to minimize the self-validation and other detrimental effects on the validity of survey findings.

The present study, however, focuses on only one area of questionnaire design: alternative response formats. The typical aim of prior studies into the effects of alternative response formats has been to determine which response format is optimal, where optimality is defined differently, depending on the study. The terms ‘answer format’ and ‘response format’ are used interchangeably here, and are understood to mean the format in which respondents are requested to answer questions. A large number of different answer formats have been proposed in the past; most of them can, however, be classified as nominal, binary ordinal, or metric in nature. An example of each of those response formats is provided in Figure 1.

Figure 1 here.

Within each of the four broad categories several different response formats exist, and these differ in subtle but very important ways. For example, a binary scale can force people either to commit to answering either ‘yes’ or ‘no’ (full binary scale). Alternatively, respondents might be asked only to tick the ‘yes’ box if they agree with a statement (affirmative binary scale). An example of an affirmative binary scale case is the question, ‘Which of the following European cities do you perceive as expensive?’, followed by a list of European cities, where respondents are asked to tick all cities they intend to visit. This is the typical format (referred to also as pick-any data) that is widely used in brand image measurement, where respondents are asked to tick how they perceive several brands (or destinations) with regard to several

attributes. The full binary scale allows the researcher to interpret a ‘yes’ answer as meaning that a respondent perceives Paris as expensive, and a ‘no’ answer as meaning the respondent does not perceive Paris as expensive. Another option, ‘I have never heard of Paris,’ could be added to ensure that no irrelevant judgments are recorded. In the case of the binary affirmative scale, the ‘yes’ option can be interpreted as meaning that the respondent perceives Paris as expensive. However, if the respondent does not tick ‘yes,’ it is unclear what they are expressing. No answer could mean ‘I have never heard of Paris,’ a state which could again be included as an answer option in order to exclude irrelevant answers. But it could also capture people who do not want to make a choice, are tired at the end of a long questionnaire or cannot be bothered thinking about another question.

Both alternatives are useful in different contexts. In the brand image measurement context the binary affirmative scale is typically used, presumably because it is not essential to know precisely what the respondent’s perception is, if it is not ‘Paris is expensive.’ For other research questions, however, it is essential that the respondent commits to one response. A recent study into alternative water sources (Dolnicar and Schaefer, 2006) is a good example. One aim of this study was to assess the level of knowledge the Australian population had about recycled water and desalinated water. Given the relatively low level of knowledge, pre-tests using the affirmative binary answer format showed that respondents who were unsure about whether, for example, recycled water was purified sewage, simply did not respond. The result was a data set that contained very few responses that answered the research questions regarding the population’s knowledge level about these water sources, or which aspects the

population was well- or ill-informed about and therefore required information campaigns.

The selection of the kind of response formats, such as affirmative binary versus full binary, can lead to major variations in results, among other factors. Even within one response format option, the wording of the question can lead to dramatic variations in results. For example, different responses can be elicited from a question worded, ‘Do you think Paris is expensive?’ compared to ‘Do you think Paris is very expensive?’.

This example illustrates two important effects. First, small modifications in response format can have major impacts on the data obtained, and consequently on the results. These effects may make it impossible to answer the very research question they were developed to investigate, as illustrated by the alternative water sources study. Second, there is no single best option for all problems. Each research question requires the social scientist to investigate alternative answer formats, evaluate their advantages and disadvantages and choose the most valid measurement instrument for the problem and the sample under study.

The above example was based on questions requiring binary answers, arguably the simplest possible response format. The complexity of potential response format side-effects increases further when a multi-category ordinal response is required from respondents.

The methodological dangers of both ordinal and binary scales have been extensively discussed by Scharf (1991), Peterson (1997) and Kampen and Swyngedouw (2000). Kampen and Swyngedouw (2000) classify most of the ordinal variables used in

tourism research, such as items capturing agreement levels with statements or satisfaction levels with service components, as 'unstandardized discrete variables with ordered categories,' and state that this is the most undesirable of all ordinal options. As opposed to categories of income or age, there is no underlying objective measure that is simply divided using known threshold values.

On the contrary, different scale points (for example, 'slightly agree' or 'quite satisfied') are likely to mean different things to different respondents, which makes interpretation extremely difficult. Also, equidistance is not assured. The distances between scale points are likely to be unequal, and could be perceived as different by different respondents. The results will not necessarily be invalid, but they could be. If ordinal data is used, it is safer to use data analytic methods that have been developed for this scale level. For example, computing a mean value on a five-point ordinal response format does not produce valuable insights, because the result cannot be interpreted unambiguously — what exactly does an average of 3.7 mean, if 3 is 'moderately satisfied' and 4 is 'very satisfied'? The widely spread treatment of data resulting from multi-category ordinal response formats as being interval-level data is very common. The typical justification is that Likert scale data can be assumed to be interval scaled. This was certainly not intended by Likert (1932), who claimed metric properties only for the summated scale, not the single items.

Billiet and McClendon (2000), McClendon (1991), and Watson (1992) draw attention to another problem inherent in Likert scales: the susceptibility to acquiescence or yea-saying bias, or, more generally, response styles. This aspect will be discussed in detail in the section on cross-cultural response styles.

A further difficulty is that multi-category ordinal scales lead to different responses independent of the number of answer categories offered to the respondents. The decision whether to use three, five, seven or nine points in the answer format affects the results. The effect of the choice of answer format is clearly a methodological artifact, and should not be interpreted as content.

The quality of ordinal scales strongly depends on the rigor of operationalization of the construct under study, the extent to which the validity of the chosen response format for the research question has been studied, and the calibration to mean the same to all respondents. If the construct is not well defined, and leaves a lot of space for interpretational differences, ordinal scales are not a very precise measurement instrument, and results remain ambiguous. The main dangers are that ordinal answer formats typically used in tourism research: (1) are not operationalized well (what does it mean to a hotel manager that a tourist slightly agrees that having a swimming pool is important? Should the hotel manager build the pool?); (2) do not offer answer options that mean the same thing to all respondents ('moderately satisfied' does not mean the same thing to all hotel guests); (3) capture both individual and cross-cultural response styles to a higher extent than alternative answer formats, as will be discussed in detail later; (4) do not have equal intervals between answer options (the difference between 'very satisfied' and 'moderately satisfied' is not necessarily the same as the difference between 'very dissatisfied' and 'moderately dissatisfied'); and (5) typically have not been tested for validity for the research problem at hand.

Despite the above insecurities involved in using multi-category ordinal scales, such response formats (particularly specific answer formats within that group, such as the

Likert scale (Likert, 1932)) have become the 'industry standard' in empirical tourism research as well as other fields, such as marketing.

While (to the author's knowledge) theoretical comparisons of the difficulties and insecurities related to each of the alternative answer formats are rare (some have been discussed above), a large number of empirical studies have been conducted in the past comparing binary with ordinal response formats. The aim of these studies was to determine which of the two generally is the better response format in the social sciences. Although, as previously mentioned, the author does not hold to the notion that a generally better scale can be found, the findings of these studies are briefly reviewed below (based on Dolnicar, 2006), with different streams of prior work using different criteria for 'better' or 'optimal.'

Several authors define optimality as 'reliability' and compare results from different scales according to how reliable they are. Interestingly, the majority of this type of work concludes that the number of response options given to respondents does not influence reliability (Bendig, 1954; Peabody, 1962; Komorita, 1963; Komorita and Graham, 1965; Matell and Jacoby, 1971; Jacoby and Matell, 1971; Remington, Tyrer, Newson-Smith, and Cicchetti, 1979; Preston and Colman, 2000). More recently, Rungie et al. (2005) demonstrate reliability issues in the context of brand image measurement using affirmative binary scales, and hypothesize that ordinal multi-category measurement would lead to similar levels of unreliability. However, several studies conclude that an association between reliability and response options exists (Symonds, 1924; Nunnally, 1967; Oaster, 1989; Finn, 1972; Ramsay, 1973).

A similar variety of conclusions emerges when validity is used as a criterion of optimality of an answer format. Jacoby (1971), Jacoby and Matell (1971), Chang (1994), and Preston and Colman (2000) conclude that response options and validity of findings are not related. Contrarily, the results obtained by Loken, Pirie, Virnig, Hinkle, and Salmon (1987), and Hancock and Klockars (1991) indicate that a larger number of options (for example, using a seven-point scale instead of a five-point scale) increases validity.

A third stream of research into the effects of response options uses factor analysis results to compare whether different scale formats result in different interpretations, thus using structural equivalence as the criterion for the quality of a response format. Martin, Fruchter, and Mathis (1974), Percy (1976), Green and Rao (1970), and Dolnicar, Grun, and Leisch (2004) chose this research approach. Green and Rao conclude that at least six answer options should be included, whereas Martin, et al., Percy, and Dolnicar, et al. found no significant differences in the factor results.

Finally, a few authors have investigated the perspective of consumer friendliness of surveys. Jones (1968) and Preston and Colman (2000) conclude that respondents prefer to have more options, and also found that this reduced perceived speed. Dolnicar (2003) and Dolnicar, Grun, and Leisch (2004) conclude that ordinal scales are perceived as significantly more difficult to answer and take significantly more time to complete.

As illustrated, prior studies comparing response formats lead to quite different conclusions: a frequency count of recommendations across response option studies would lead the scientific community to believe that seven-point scales are the optimal

choice (Cox, 1980). The popularity of such multi-category ordinal response formats in the social sciences has been noted by Peterson (1997), Van der Eijk (2001), and Dolnicar (2002). This supports the notion that response format decisions are not able to be generalized: depending on the research questions, the construct under study, and the nature of the sample, different response formats will be appropriate or inappropriate, will produce very valid, moderately valid or invalid results. However, the acceptance of an emerged standard is potentially a very dangerous decision that can — in the worst case — lead to invalid results or the inability to even answer the research question.

2.1 Are we 'following the recommendation of…' in our choice of answer formats?

Often we cannot assess whether an author has uncritically chosen a particular scale, or whether they have invested considerable time and effort in their decision. Typically, information allowing us to make this judgment is not available in manuscripts. We cannot therefore empirically evaluate the extent to which emerged standards are accepted uncritically, or compare the proportion of studies that are based on a thorough analysis of the response format before fieldwork is conducted. Consequently, testing the level of validity of conclusions drawn in published empirical tourism research. If in doubt, we should assume that the authors have thoroughly evaluated their response format. The empirical illustration here cannot aim to state the proportion of uncritical use of response formats or the proportion of findings with questionable validity. It can, however, analyze the proportion of studies that are prone to the abovementioned problems due to the use of accepted standards

without explanation and discuss their potential dangers, or look at how these potential dangers have been addressed. The inability to test for critical use and validity of findings thus limits the empirical investigation to the following three hypotheses derived from the insights from prior research as reviewed above:

H1.1 Empirical tourism researchers predominantly (in more than 80 percent of studies) use multi-category ordinal scales.

H1.2 The majority of empirical tourism researchers (more than 50 percent) do not provide reasons for their choice in the manuscript.

H1.3 The majority of empirical tourism researchers (more than 50 percent) do not point out the dangers or insecurities associated with their choice in the manuscript.

H1.4 The majority of empirical tourism researchers (more than 50 percent) use data analytic techniques, which are not suitable for the response format used.

The method selected to investigate the research aims of this study is a literature review of academic tourism research published in 2005 in three of the main journals that publish empirical social sciences[1] research: the *Journal of Travel Research*, *Annals of Tourism Research*, and *Tourism Management*. All articles published in 2005 were screened and classified as being either empirical or non-empirical in

[1] According to the Oxford English Dictionary the social sciences encompass 'The scientific study of the structure and functions of society; any discipline that attempts to study human society, either as a whole or in part, in a systematic way.'

nature. In order to be classified as empirical for the purpose of this review, disaggregate data had to form the basis of the investigation, and the subjects under study had to be tourists.

Sixty-five studies published in the three outlets in 2005 were classified as empirical and included in the review. Each study was reviewed in detail and coded with seven variables relevant to hypotheses H1.1 to H1.4: (1) the answer format used, following the classification illustrated in Figure 1; (2) the specific answer format, such as 'Likert scale,' if mentioned; (3) whether someone else's work was cited to justify the use of the used response format; (4) whether an explanation was provided as to why the chosen response format was deemed the best choice for the research question at hand; (5) whether the dangers associated with this response format were discussed; (6) which method was used to analyze the data; and (7) whether raw data or a summated scale value was used for data analysis.

Coding was undertaken separately for each of the constructs investigated in each of the published studies, and descriptive statistics were computed to test hypotheses H1 to H4.

The results are provided in Table 1. Citations of prior work were used in 17 percent of the studies, which at first appears to indicate that researchers may not be following emerged standards at all with regard to answer formats. However, only five percent of the authors provided an explanation of why they chose the answer format. An even lower proportion (three percent) discussed potential dangers of the answer formats used.

With respect to which answer formats were actually used, the high proportion of studies using multi-category ordinal scales assumed in hypothesis H1.1 was actually surpassed, with 88 percent of all studies using ordinal scales (either ordinal only or in combination with other scales). Only two studies explicitly stated the nature of the scale as being nominal, binary, ordinal, or metric. Mostly the authors showed scale but did not discuss its mathematical properties and implications.

The main methods of data analysis were factor analysis (either alone or in combination with other analytic techniques more than half of all empirical studies used this technique), descriptive statistics, analysis of variance, and logistic regression. Factor analysis and analysis of variance require metric data levels. More than half of the empirical studies undertaken in tourism research do not use methods appropriate for the answer format in the instrument.

Table 1 here

This leads to the following findings with respect to H1.1 to H1.4. Hypothesis H1.1 cannot be rejected because the vast majority of empirical tourism studies uses a multi-category ordinal answer format, either as the sole measurement instrument or in combination with other answer formats. Hypothesis H1.2 cannot be rejected either, because no explanation for the use of the response format was provided in 95 percent of the studies reviewed. Dangers associated with response styles are only discussed in three percent of all studies. Consequently, H1.3 cannot be rejected. Finally, in more

than half of the studies analytic techniques are used, which require higher than ordinal data level, indicating that a significant proportion of research applies analytic techniques unsuitable for the data properties.

The fact that hypotheses H1.1 to H1.4 could not be rejected indicates that empirical tourism research as a field is very prone to confounding the result component of a finding with the measurement artefact component, resulting from the use of response formats that often appear not to have been validated and calibrated for the particular research problem at hand.

2.2 A few things to consider when choosing a response format

This section cannot provide the magic solution to all research problems: the single best response format. Instead it aims to list several aspects that empirical tourism researchers may want to consider when choosing their response format. Each aspect is discussed independently. The selection of the optimal response format, however, requires us to account for all the following points in an integrated manner:

1. Is the speed of completing the questionnaire critical? Decreasing the time required to complete a survey can be necessary for at least one of two reasons: (a) longer questionnaires are more expensive, because respondents are paid more to compensate for their time in self-administering the survey, or the expenses for interviewer time increase; and (b) longer questionnaires are known to lead to a reduction in data quality (Johnson, Lehmann, and Horne, 1990). Typically, high quality data is the primary aim of an empirical social scientist. If questionnaire length is a concern, the use of binary response formats is recommended, because

the time saved is about 30 percent, with aggregated results showing very little deviation from ordinal scales (Dolnicar, 2003; Dolnicar, Grün, and Leisch, 2004).

2. Is the increased detail available by using ordinal multi-category scales required to answer the research question? Multi-category ordinal response formats enable frequency counts of all options, which the binary answer format cannot provide. The crucial question is: is this additional detail required? The best way to answer this question is to determine what the research question is, and in which way the data will be analyzed. If it is essential to know what proportion of respondents are 'moderately satisfied,' and if frequently counts will be computed to assess this proportion, a multi-category ordinal or metric response format is needed. If analyses will be based on means or analytic techniques that use the mean value as a basis, a binary answer format may be sufficient because it has been demonstrated in the past that at the aggregate level the mean derived from a binary scale essentially leads to the same interpretation as the typically (incorrectly) computed mean across a limited number of ordinal multi-category scale points (Dolnicar, 2003; Dolnicar, Grün, and Leisch, 2004).

3. Can it be reasonably assumed that all respondents will perceive the ordinal response options in the same way? For example, will 'very satisfied' mean the same thing to all respondents? If this can be reasonably assumed, the multi-category ordinal response format is a suitable choice. If not, then the seemingly higher level of precision is contaminated to an extent that it is questionable whether the responses can be interpreted beyond positive versus negative (and thus binary) statements.

4. Can it reasonably be assumed that the distances between the ordinal multi-category answer alternatives are perceived as the same? For example, will the difference between ‘satisfied’ and ‘very satisfied’ be perceived as identical to the difference between ‘dissatisfied’ and ‘very dissatisfied’? One could argue that this is not the case; that instead, the jump to ‘very dissatisfied’ is significantly larger from ‘dissatisfied’ than the distance between the two positive scale points. If equidistance can be shown in presets or can be reasonably assumed, the multi-category ordinal scale is suitable. Otherwise, choosing an ordinal multi-category format has the consequence that data-analytic techniques assuming metric data have to be eliminated from the portfolio of applicable methods for such data, because computation of distance is meaningless — even misleading. In such cases a metric response format could be considered if a high level of detail is required in the response, or a binary format if this is not the case.

5. Further decisions that need to be made based on the construct under study if multi-category ordinal response formats are chosen include (a) whether the response format should be unipolar (for example, from ‘not expensive’ to ‘very expensive’) or bipolar (from ‘very cheap’ to ‘very expensive’); (b) how many scale points should be used; (c) whether all response alternatives should be verbalized, or only the endpoints; and (d) how will the response alternatives or endpoints should be verbalized (as ‘very expensive’ or ‘strongly agree’).

6. Will the sample include respondents from different cultural backgrounds? If so, binary response formats may be the preferable solution if there is not sufficient time to undertake rigorous testing of various levels of equivalence before the

fieldwork. Binary formats were recommended in this context by Cronbach (1950), in order to reduce cross-cultural response bias. This aspect will be discussed in more detail in the next section.

3 CROSS-CULTURAL RESPONSE STYLES

Much empirical tourism research will be confronted with individuals from different cultural backgrounds. The more global that tourism becomes, the smaller the proportion of demand-oriented studies that can use samples of respondents from only one country or cultural background. Empirical research in tourism often aims to reveal differences between cultural groups or tourists from different countries of origin.

The need to compare respondents from different countries or cultural backgrounds exposes the discipline to several potential result contaminants: culturally biased response norms can cause different scale usages independent of the information passed on by completing a questionnaire; questions can be interpreted differently; and the underlying constructs measured might not be identical. Therefore, the most concerning potential mistake resulting from cross-cultural response styles is that differences in group means can become uninterpretable (Chun, Campbell, and Yoo, 1974), although typically the comparison of means across countries or cultures constitutes the central analysis in cross-cultural comparisons.

The tourism research literature has not broadly discussed the potential dangers of interpreting empirical data derived from surveys conducted in different languages in different places (with the exception of Kozak, Bigne, and Andreu, 2003, in the context of satisfaction research, and Dolnicar and Grün, in press). However, psychologists,

sociologists and market researchers have investigated cross-cultural issues in empirical research extensively. The following overview is based on the review by Dolnicar and Grün (in press).

The number of potential pitfalls is huge, as Sekaran (1983) discusses in detail. The central problem is equivalence. However, equivalence has to be ensured at several different levels. Sekaran categorizes them into the following areas: functional equivalence, equivalence of instruments (vocabulary equivalence, idiomatic equivalence, grammatical, and syntactical equivalence), conceptual equivalence, transferability of concepts, data collection, sampling, scaling, data analysis, and measurement equivalence. Functional equivalence means that the behavior to be measured should be naturally occurring. Conceptual equivalence refers to the requirement that the object of study should have the same meaning in all cultures included in the study. The criterion of transferability looks at whether concepts can be transferred to different cultures. Vocabulary, idiomatic, grammatical, and syntactical equivalence are part of the equivalence of instruments, and relate to the translation process of survey instruments, the use of idioms that may not be directly translatable to another language, and the grammatical form of the questions (which is particularly important when long or complex text components need to be translated). Data collection could cause bias if there are different methods of data collection in different countries. Sampling can cause bias if the samples of different countries are not all representative of the local population or directly matched. Scaling equivalence requires that the response format used should elicit responses in the same way from all groups of respondents. Measurement bias could result from different cultural

sensitivities to topics studied. Data analysis could distort findings if data from different cultural groups are analyzed in different ways.

Kozak, Bigne, and Andreu (2003) provide a similar review specifically for the context of cross-cultural satisfaction research in tourism, and they distinguish between functional, conceptual, instrument, and measurement equivalence.

Measurement bias is the area most critical to the majority of cross-cultural studies conducted in empirical tourism research. For example, a survey of visitors to Austria is not confronted with the typical problems listed above, and sampling is based on representatives of each country or cultural subgroup who visit Austria. Representativity of the home country's population, or matching of individuals across countries of origin, is consequently not a relevant criterion. Rather, the representativity of each country or cultural group for the visiting pattern to Austria is of importance. Measurement equivalence, however, is relevant in all contexts and for all constructs measured in typical empirical tourism research.

Smith and Reynolds (2002) further break down the aspect of measurement bias, and differentiate between *response sets* and *response styles*. Response sets describe differences in responses that are due to how respondents from different cultures would like to be perceived. In contrast, response styles is used for differences in responses that are systematically related to the response format. Smith and Reynolds conclude that 'Failure… to detect differences in cross-national response bias will… affect data comparability, may invalidate the research results and could therefore lead to incorrect inferences about attitudes and behaviors across national groups' (2002, p. 450).

This section focuses on the discussion and investigation of emerged standards in the area of cross-cultural response styles. Several empirical studies have been conducted that aim to detect cross-cultural response styles. Chun, Campbell, and Yoo (1974) tested differences in extreme response styles between US and Korean students, and concluded that a significant difference exists, and that US students were more prone to demonstrate an extreme response style. Bachman and O'Malley (1984) investigated differences between colored and Caucasian high school seniors in responding to Likert questionnaire items, and found that colored students were more likely to use extreme response options. Hui and Triandis (1989) concluded from their study that Hispanic respondents use extreme scores more often than non-Hispanic respondents, and this is supported by the Marin, Gamba and Marin (1992) study. Watkins and Cheung (1992) found differences in response styles between survey participants across countries, and also detected that the variation is higher among women. Clarke III (2000) found that Hispanics and colored respondents exhibited higher levels of extreme response styles than the other groups, and that the French used more extreme responses than Australians. Van Herk, Poortinga, and Verhallen (2004) identified response style biases in countries within the EU, with respondents from the Mediterranean showing higher levels of both extreme and acquiescence response styles than respondents from north-western Europe.

These findings indicate that cross-cultural response styles do exist, and that they represent a major threat to empirical tourism research based on data collected from respondents from different countries or cultural backgrounds.

3.1 Are we 'following the recommendation of…' in conducting cross-cultural comparisons? An empirical investigation

Based on the review of prior work and the nature of empirical tourism research, the following hypotheses are formulated regarding the accepted standards with regard to cross-cultural response bias, and the extent to which cross-cultural empirical tourism research findings are endangered by response styles:

H2.1 The majority of empirical tourism studies (more than 50 percent) are based on multicultural samples (that is, samples including respondents from more than one cultural group or country).

H2.2 The majority of empirical tourism studies using multicultural samples (more than 50 percent) draw comparisons between respondents from countries and/or cultural backgrounds.

H2.3 The majority of empirical tourism studies using multicultural samples (more than 50 percent) use the multi-category ordinal response format.

H2.4 The majority of empirical tourism studies using multicultural samples (more than 50 percent) do not mention potential problems resulting from cross-cultural response styles.

H2.5 The majority of empirical tourism studies using multicultural samples (more than 50 percent) do not assess the extent of cross-cultural response style contamination.

H2.6 The majority of empirical tourism studies using contaminated multicultural samples (more than 50 percent) do not correct for cross-cultural response style contamination.

The same review procedure as outlined in section 2.1 was used. Answers to the following questions were coded into a data set:

(1) Did the sample include respondents from different countries or cultural backgrounds? For all studies for which this was the case, additional data was coded.

(2) Was a comparison across countries or cultural backgrounds undertaken?

(3) Which response format was used?

(4) Was any aspect related to problems with cross-cultural studies mentioned?

(5) Was the extent of the contamination assessed?

(6) Was data corrected for the contamination?

The frequency counts for these variables are included in Table 2.

Table 2 here

At least one-third of articles published in 2005 used samples that included respondents from more than one country of origin or cultural background. Given that the sample is not described sufficiently in many studies, this proportion could be as

high as half of all studies. Of those studies that use multicultural data, 36 percent actually draw cross-cultural conclusions. Hypotheses H2.1 and H2.2 therefore must be rejected. The proportion of studies endangered by cross-cultural response styles among all empirical studies published in 2005 is not more than 50 percent, based on 2005 publications between 34 and 49 percent of studies include multicultural samples.

Hypothesis H2.3 cannot be rejected, because 86 percent of respondents used the multi-category ordinal answer format. Hypotheses H2.4 to H2.6 cannot be rejected either, because the vast majority (91 percent) of empirical studies based on multicultural data do not discuss the dangers associated with this approach, do not assess the danger of response styles and do not correct for response style contamination.

3.2 A few things to consider regarding cross-cultural response styles

Essentially there are three ways to avoid problems with cross-cultural response styles: (1) not to conduct them (which admittedly is not much help for those who do); (2) to choose a response format that is less susceptible to cross-cultural response styles; and (3) to assess the existence/extent of the response bias and correct for it.

With respect to the second recommendation (to use response formats less susceptible to response styles), there has been little empirical research undertaken to determine which response formats would be suitable. Clarke III (2000; 2001) and Roster, Rodgers, and Albaum (in press) found that lower numbers of response options in multi-category ordinal scales lead to more extreme answers. This is not surprising, given that the number of options is lower. However, neither study identified

differences between cultural groups with respect to this shift towards extreme answer options. This means that a certain kind of response style (extreme response style) is more prominent the fewer the scale point. However, it does not seem to be the case that different cultures shift to extreme answers more or less frequently. Cronbach (1950, p. 21) recommended the use of binary format instead of multi-category ordinal scales, as well as the following recommendation: 'Since response sets are a nuisance, test designers should avoid forms of items which response sets infest.' Given the unambiguous findings reported above regarding the way that multi-category ordinal scales are highly prone to cross-cultural response styles, binary answer formats should be seriously considered as an alternative if ordinal or metric level data is not essential. Lower-level, high-quality data may be preferable to higher-level, contaminated data.

Regarding option 3 (to assess the existence/extent of the response bias and correct for it), several authors have made suggestions how cross-cultural response styles can be corrected for (Cheung and Rensvold, 2000; Byrne and Campbell, 1999; Greenleaf, 1992a and 1992b; Van de Vijver and Poortinga, 2002; Welkenhuysen-Gybels, Billiet, and Cambre, 2003). Their recommendations range from very simple approaches, for example, investigating if systematic response patterns can be detected for the same cultural group; to modeling approaches that try to extract the extreme response and acquiescence bias from the actual information content and then correct the data accordingly. These approaches have one thing in common: they assume to know the extent to which data is contaminated and then be able to correct for this contamination. This assumption has the disadvantage that exactly which type of contamination occurs, and its extent, are unlikely to be evident. The decision to make

a particular correction, therefore, carries the danger of transforming the data incorrectly, and by doing so, introducing new contaminations.

Dolnicar and Grün (under review) took a different approach, which takes precisely this danger into consideration. They recommended identifying the subset of all correction methods that are theoretically appropriate for a particular data set at hand, correcting the data using all correction techniques and then computing the results for the uncorrected and all corrected data sets. Where no deviations in findings occur, a firm conclusion about cross-cultural differences can be drawn. If, however, the findings differ independently of the correction method used, conclusions must be drawn with care, and should draw the reader's attention to the possibility that response styles may be causing detected differences (or no differences) between respondents from different cultures or countries.

While this article has focussed on response style effects in cross-cultural comparisons, the equivalence dimensions discussed in the literature review should be considered and discussed in any study that involves cross-cultural comparisons.

4 MARKET SEGMENTATION

It is now widely accepted among tourism researchers that tourists are not one homogeneous group of people who seek the same benefits from a destination, have the same expectations, undertake the same vacation activities and perceive the same vacation components as attractive. Tourists are highly heterogeneous. In the optimal case, the tourism industry should therefore cater for individuals and their specific vacation needs. While this approach may be feasible in online interfaces (for example,

by supporting individual people in their destination choice), it is not feasible to modify the entire marketing mix of a tourism business or destination to suit individual needs. The next best option to individual customization is the identification or definition of groups of similar tourists: market segments.

Because of its benefits, the concept of market segmentation has been embraced both by the tourism industry and tourism researchers. The aim of market segmentation studies is to identify, construct or define market segments, and profile their characteristics in sufficient detail to make them an actionable target market for tourism industry. Every market could be segmented in myriad different ways, and each of these possible segmentations of the market is not equally attractive. Ideal segments would contain tourists with similar tourism needs and behaviors, and similar socio-demographic profiles. These are targets who are profitable, who could be easily reached with marketing communication messages, who match the strengths of the tourism destination or business, and whose needs are not catered for by major competitors. Such ideal segments would be highly attractive from the tourism industry point of view, because they would have the most potential for profit increase through more targeted marketing activities, with a higher effect on market demand within the targeted segment.

Consequently, it is the tourism researcher's aim to explore markets, and suggest market segments to the tourism industry that are as ideal as possible. The researcher must choose between large numbers of possible segmentation solutions; a decision which is better made based on the structure of the data, rather than on the subjective opinion of the tourism researcher or manager. The segmentation research approach,

which aims to investigate data structure systematically, therefore represents the key to successful data-driven market segmentation. Failing to explore the market (data) in such a way as to identify or construct ideal segments can lead to an irreversible competitive disadvantage for the tourism business or destination that uses the segmentation as their basis for marketing action. Consequently, it is crucial to ensure that the research approach to market segmentation is rigorous and avoids potential misinterpretations.

This burden of responsibility is different for different kinds of segmentation studies. In the case of *a priori* (Mazanec, 2000) or *commonsense* segmentation (Dolnicar, 2004) and extensions thereof (concepts 1, 3, 4, and 5, according to the classification of segmentation studies proposed by Dolnicar, 2004) the crucial decision is the selection of the segmentation criteria. For example, a destination might choose to target young tourists using age as the commonsense criterion. On closer evaluation, however, it might be that using the stage in the family lifecycle would have been a better choice, because the destination's strength lies in providing optimal services to young families, rather than young singles or groups of young tourists. In the case of *post-hoc* (Myers and Tauber, 1977), *a posteriori* (Mazanec, 2000) or *data-driven* segmentation (Dolnicar, 2004) and extensions thereof (segmentation concepts 2, 4, 5, and 6), this burden of responsibility rests on the research approach of the data-driven segmentation study undertaken. Because the process of data-driven segmentation consists of numerous components, most of them requiring the researcher's decision, it is more difficult to avoid potential misinterpretations or suboptimal procedural decisions than is so for *a priori* segmentation studies.

In addition to grouping segmentation studies in *a priori* (commonsense) and data-driven (*a posteriori*, *post-hoc*) studies, data-driven studies can be further classified as either *response-based* or *step-wise*. The step-wise procedure aims to group respondents according to certain variables in the first step, and describing them in the second. Typical descriptors, or background variables, are variables relevant to marketing, for example, the media behavior of segments, their purchasing frequency, and the amount of money spent on holidays per year. In the step-wise procedure the grouping is based purely on the variables selected, for example, travel motives. The background variables do not interfere with the determination of the segments. Response-based segmentation uses one or more variables, which are relevant from a marketing perspective as the dependent variable in the segmentation process, thus confounding, for example, the travel motive segments with their media behavior. The segments are consequently not pure travel motivation segments.

This section discusses step-wise, data-driven procedures because they dominate the area of segmentation studies in tourism research. All stages (discussed in detail below) are depicted in Figure 2.

Figure 2 here

The **study design** stage is mentioned in Figure 2 because many segmentation studies have design requirements different from other studies. Segmentation studies aim to identify all market segments in the market. Sometimes small niche markets are of

particular interest because they may provide a distinct competitive advantage. It is therefore not necessarily desirable for a segmentation study to include a representative sample of respondents. If identification of segments is the primary aim, the sample must be as heterogeneous as possible — it should include the widest variety of respondents in sufficient number to enable the cluster algorithm to identify niches.

At the **data collection** stage, variables must be carefully developed before being included in a questionnaire. It is not good practice either to include all items that seem interesting without pre-testing or theoretical justification, or to include as many redundant items as possible to achieve a high alpha value (Cronbach, 1951), if this happens at the expense of the conceptualization of the construct of interest or respondent fatigue, while not providing any information of additional value. Evidence demonstrates the negative effect of only one or two variables that are unrelated to the segmentation (Milligan, 1980; Milligan, 1996), which leads to the strong advice both from methodological researchers (for example, Milligan) and marketing scientists (Punj and Stewart, 1983) to exercise extreme care in selecting variables.

The approach typically used for scale development in tourism research follows Churchill's early recommendations (Churchill, 1979). Since then, Churchill himself (1998, p. 30), while pleased with the improvement of **measurement** in the area of marketing, expressed criticism about the way in which his paradigm was frequently misinterpreted: 'The bad news is that measurement seems to almost have become a rote process, with the Paradigm article serving as backdrop for the drill, thereby supposedly lending legitimacy to what seems to be at times thoughtless, rather than thoughtful, efforts.'

Two (groups of) authors have recently proposed alternatives to Churchill's ruling paradigm of scale development. As Finn and Kayande (2005, p. 12) expressed in a review of the Churchill scale development procedure: 'Step-by-step applications overemphasize validation numbers at the expense of conceptual rigor. Numbers [are] often misleading due to misidentification of relevant objects of measurement.' Recommendations for improvement of the present scale development paradigm have been proposed by Rossiter (2002), who criticizes the lack of conceptualization and questions the need for multi-items in particular instances; and Finn and Kayande (1997), who criticize the limitation of the scaling of characteristics of individuals, although typically, researchers are not interested in single individuals, because, for example, psychological measurements are.

Given that the data set is the most fundamental basis for good market segmentation, careless selection of items to be included in the questionnaire can critically affect the quality of results. Whenever a tourism researcher conducts a segmentation study, data collection should be planned as an integral part of the study, to ensure that all important pieces of information are obtained without burdening respondents with unnecessary items. If a segmentation researcher is, however, confronted with a secondary data set in which redundant items are included, the direct inclusion of such items into the segmentation process should be critically questioned.

In sum, the dangers of uncritically following measurement paradigms when the segmentation base is collected include: (1) the construct that is of central interest could be badly conceptualized; (2) respondents could be confronted with large numbers of questions that are highly redundant, which is likely to lead to lower data

quality due to respondent fatigue; (3) items that are not redundant and thus measure a different dimension of a construct (and might well be the most important items for identifying niche markets) may be eliminated because they reduce the values of reliability measures; and (4) by including many redundant items, the segmentation researcher has only shifted the variable selection problem from the pre-survey to the pre-segmentation phase. The process by which the items were derived has to be described in detail to clarify that the construct of interest is best captured by the selected variables, and that the variables can be expected to differentiate between segments.

If the construct that was measured for the purpose of segmenting is used for segmentation, the next step (the **selection of variables** to be used in the segmentation process) is unnecessary. However, segmentation studies are sometimes conducted based on data that was collected for an entirely different reason. The biggest danger with respect to the choice of variables to be included in the segmentation process in this case is the uncritical inclusion of as many variables as possible, in the hope that some structure will emerge (Aldenderfer and Blashfield, 1984; Everitt, 1979). Or, as Milligan (1996, p. 348) puts it: 'Far too many analyses have been conducted by including every variable available… Most researchers do not appreciate the fact that a variable should be included only if a strong justification exists that that variable helps to define the underlying clustering.'

The question of how many respondents are required to group them, based on a certain number of variables, cannot be answered easily. **Sample** size requirements essentially depend on two factors: the methodological approach chosen to analyze the data

(parametric approaches require minimum sample sizes, whereas non-parametric explorative analyses do not), and the structure of the data (if the data set is very well structured, only a few variables may be needed to group individuals correctly; if, however, the data set is not at all well structured, a lot of information from many respondents is required to determine the best grouping). While the choice of method is under the researcher's control, the data structure is not.

The only recommendation that has been published (to the author's knowledge) has been Formann's (1984), in the context of latent class analysis (a parametric procedure). He states that a sample of at least 2^k is needed to segment the respondents based on k variables; preferably $5*2^k$ should be available. The number 2 indicates that Formann assumes that a binary answer format will be used. If an ordinal format is used, the number 2 has to be exchanged by the number of ordinal scale categories chosen. Imagine, for example, a block of 20 travel motives, which respondents are asked either to agree or disagree (binary scale) with. If these 20 items are to be used as a segmentation base using a parametric procedure, the required sample size is 1,048,576 respondents. If 15 items are used, 'only' 32,768 respondents are needed, and with 10 items, 1,024 completed surveys are sufficient to segment the market, based on the travel motives.

The next stage, the **selection of answer format**, will not be discussed in detail in this section because it has been dealt with in detail in the section on answer formats. However, it is worth noting here that all cluster algorithms used to segment markets are based on distance computations, as illustrated in Figure 3, in which depicts a very simple case of three respondents and three variables. Respondents answered on a

binary scale. If the match of answers is used as a measure of similarity, respondent 1 and respondent 2 reach a value of 2, because they both want excitement during their holiday and both do not want to rest. Respondent 3 achieves a value of 1 with respondent 1, because neither cares about security, and a value of 0 with respondent 2. In this case (using the proposed distance measure and a hierarchical algorithm), respondents 1 and 2 would be assigned to one market segment. If the absolute Euclidean distance were used, the distance between respondent 1 and 2 would be $|(0–0)+(1–1)+(0–1)| = 1$. For respondent 1 and 3 it would amount to 2, and for respondents 2 and 3 it would be 3; again indicating that respondents 1 and 2 are the least dissimilar.

Figure 3 here

Given that similarity or dissimilarity of response vectors is used in the cluster step, the choice of distance measure is very important, especially regarding its suitability for the answer format selected. As noted in the section on answer formats, the use of multi-category ordinal scales is most complicated, because equidistance cannot be assumed, and therefore, the most common distance measure, Euclidean distance, is not an appropriate choice.

In Figure 2 the box representing **pre-processing** of data is depicted in light grey in order to indicate that (while it appears that pre-processing of data has developed to become an accepted standard in empirical tourism research) pre-processing is not an

essential component of the step-wise, data-driven segmentation process. Aldenderfer and Blashfield (1984) discuss the issue of data pre-processing through standardization and other forms of transformations extensively. They review several studies which came to different conclusions in respect to the effect of data standardization on results. In sum, the dangers of pre-processing are that: (1) the relations of variables to each other could be changed; (2) differences between segments could be reduced; and (3) segments identified are done so in a space different from originally postulated (Ketchen and Shook, 1996).

The most frequently used method of pre-processing in market segmentation is factor analysis. While factor analysis can help eliminate variables that measure the same construct, and by doing so prevent one construct being weighted higher in the segmentation solution, the danger associated with this procedure is that differences between segments that are not clearly separated from each other cannot be detected as easily (Aldenderfer and Blashfield, 1984). However, Aldenderfer and Blashfield found no evidence of negative impact if the data contained well-separated segments.

Arabie and Hubert (1994) take a clearer position on the use of factor analysis in the context of clustering. They state that '"tandem" clustering is an outmoded and statistically insupportable practice,' because data is transformed, thus the nature of the data is changed before segments are searched for. This is supported by Milligan (1996), who, based on experimental findings that clusters in variable space are not well represented by clusters in component space, states that the researcher has to address in which space the segments are postulated to exist.

In tourism research, the typical reason stated for using factor analysis is the need to reduce the number of variables. This argument poses two questions:

(1) Why was the number of items not reduced in the variable measurement stage to retain a reasonable number of relevant, non-redundant questions that are expected to discriminate between segments?

(2) If the researcher did not have any influence on the data collection, and is faced with a data set with too many variables, why is factor analysis preferred over simpler ways of variable selection, which avoid data transformation?

The most illustrative argument against the uncritical use of factor-cluster analysis in tourism research is provided by Sheppard (1996). He explains the paradox that homogeneity has to be assumed for factor analysis, whereas heterogeneity is explored by cluster analysis. He also demonstrates in an empirical example that the results derived from factor-cluster analysis, cluster-factor analysis and cluster analysis based on raw data lead to totally different conclusions. In his example, the factor-cluster approach led to results different from cluster analysis on its own, and effectively failed to identify the true segment structure in the data. Furthermore, he demonstrated how the exclusion of items based on low loadings with factors can undermine the aim of the entire segmentation study if the low loading item actually represents a relevant discriminating variable between segments. When ‘accurate and detailed’ segmentation results are the aim of the study (the case for most tourism segmentation studies), Sheppard recommends clustering of raw data directly. Sheppard’s study shows that assumption (5), discussed in the section on the standard research approach, is not

appropriate, because factor-cluster analysis not only leads to different, but inferior, results if the aim is the identification of market segments.

In sum, there are number of problems associated with the practise of using factor analysis in the pre-processing stage of a segmentation study to reduce variables: (1) the data is transformed and segments are identified based on the transformed space not the original information respondents gave, which leads to different results; (2) with a typical explained variance of between 50 and 60 percent, up to half of the information that was collected from respondents is discarded before segments are identified or constructed; (3) eliminating variables that do not load highly on factors with an Eigenvalue of more than 1 means that potentially the most important pieces of information for the identification of niche segments are discarded, thus making it impossible ever to identify such groups; and (4) interpretations of segments based on the original variables are not possible — segments can only be interpreted with respect to their factor score values.

The broad term used to subsume all algorithms used in step-wise, data-driven market segmentation is cluster analysis. This term describes a large number of algorithms for grouping observations based on similarity or dissimilarity.

Extensive Monte Carlo simulations have shown that most algorithms can identify the correct segmentation solution if the data is highly structured (Buchta et al., 1997). However, if this is not the case, the algorithm chosen does not act as a neutral tool in the segmentation exercise; rather, it creates a segmentation solution. Also, each algorithm has different tendencies regarding which kind of segments it creates. Or, as

Aldenderfer and Blashfield (1984, p. 16) put it: 'Although the strategy of clustering may be structure-seeking, its operation is one that is structure-imposing.'

This has two consequences for the researcher aiming to segment a market: (1) it is important to know whether the data used as a segmentation base is well structured; and (2) the solution is likely to depend on the algorithm chosen, which makes the selection of an appropriate segmentation algorithm a crucial step in the process.

The limitations of algorithms and the ways they influence the nature of the solution are well known for most algorithms. For example, hierarchical procedures are not suitable for very large data sets because the hierarchical clustering algorithm requires the computation of all pair-wise distances at each stage of grouping, as in the example above with three respondents. Within the group of hierarchical procedures single linkage procedures create chain formations in the final segmentation solution (Everitt, 1993); self-organizing neural networks not only partition the data, but also render a topological map of the segmentation solution that indicates the neighborhood relations of segments to one another (Kohonen, 1997; Martinetz and Schulten, 1994); fuzzy clustering approaches relax the assumption of exclusiveness (for example, Everitt, 1993); and ensemble methods use the principle of systematic repetition to arrive at more stable solutions (for example, Leisch, 1998 and 1999; Dolnicar and Leisch, 2000 and 2003). These are just a few of the distinct properties that different techniques have.

One example of an ensemble technique is bagged clustering, which has only recently been introduced into tourism research (Dolnicar and Leisch, 2003). Its major advantage is that it investigates the structure of the data while simultaneously

producing a segmentation solution. Comparative studies (Dolnicar and Leisch, 2004) have found bagged clustering to produce more stable and therefore more reliable segmentation solutions.

One component of the segmentation algorithm is the **measure of association** used, as mentioned above. The measure of association chosen must be suitable for the answer format, which means that it must be able to deal with binary, ordinal or metric-level data. Euclidean distance, the most widely used measure, is suitable for binary data and metric data to determine a particular kind of distance. Strictly speaking it is not suitable for ordinal data unless it has been shown that the distances between the categories are perceived as equidistant by all respondents.

Arguably, the most critical decision in the process of step-wise, data-driven segmentation is the decision of **how many clusters** to choose, a problem that remains unsolved since the wider adoption of clustering techniques (Thorndike, 1953). Similar to the decision about which algorithm to use, this decision depends on the nature of the data being analyzed. If the data is very highly structured in terms of density structure (that is, clear market segments exist), every algorithm can recommend the correct number of clusters (Buchta et al., 1997). If, however, the data is not highly structured (which, based on the author's experience, is the typical case in the social sciences), deciding on the number of clusters is very difficult. Many different approaches and indexes have been proposed in the past (for comparative studies see Milligan, 1981; Milligan and Cooper, 1985; Dimitriadou, Dolnicar and Weingessel, 2002; Mazanec and Strasser, 2000).

The issue of **validity** of market segmentation solutions cannot be discussed independently of the aim of the segmentation exercise, and the aim is not independent of the available data. The aim of detecting natural clusters that exist in the data (Aldenderfer and Blashfield, 1984) is only suitable if the data is highly structured and actually contains natural density clusters. This can be assessed by investigating the stability of solutions if computed repeatedly for the same number of clusters. If the stability results indicate that natural segments do not exist in the data — which is the implicit assumption made by Mazanec (1997) and Wedel and Kamakura (1998) — the aim of the market segmentation exercise is to identify the most managerially useful segments. The most common case in market segmentation is to construct artificial groupings, even though this aim is counterintuitive. Such solutions are valuable because the segments constructed are more homogeneous groups of individuals, which can be targeted with customized messages. The degree of managerial usefulness can be evaluated by inspecting the segment profiles and assessing the match with organizational strengths — or by assessing stability and choosing the most stable solution. Stability is a major issue in data-driven market segmentation as compared to the *a priori* approach (Myers and Tauber, 1977).

While stability is one condition of validity, if naturally occurring segments are the focus, another aspect of validity is independent of whether natural groups are identified or whether artificial groups are constructed. Segments should be distinctly different from one another. Given that the clustering algorithm produces a solution where segments are distinctly different with respect to the variables used in the segmentation process (the segmentation base), testing for significance of difference in the segmentation base is not a legitimate test for distinctness. However, additional

information that is available about the respondents can be used to test whether segments are distinctly different. Depending on the number and data scale of the additional variables, and the number of segments, different approaches can be used to assess the distinctness of the segments. Options include discriminant analysis, analysis of variance, chi-square tests, and binary logistic regression.

4.1 Are we 'following the recommendation of…' in market segmentation? An empirical investigation

The following hypotheses were formulated regarding emerged standards for step-wise, data-driven market segmentation in empirical tourism research.

H3.1 In the majority of segmentation studies (more than 50 percent) data is not specifically collected for the purpose of segmentation.

H3.2 In the majority of segmentation studies (more than 50 percent) no explanation is provided for the measurement of variables.

H3.3 In the majority of segmentation studies (more than 50 percent) no explanation is provided for the selection of variables.

H3.4 In the majority of segmentation studies (more than 50 percent) the sampling strategy is not developed in view of the segmentation study.

H3.5 In the majority of segmentation studies (more than 50 percent) the segmentation base is of multi-category ordinal nature.

H3.6 In the majority of segmentation studies (more than 50 percent) data is pre-processed using factor analysis.

H3.7 In the majority of segmentation studies that pre-process data (more than 50 percent) no explanation for pre-processing is provided.

H3.8 In the majority of segmentation studies (more than 50 percent) the measure of association used is not stated.

H3.9 In the majority of segmentation studies (more than 50 percent) data structure is not investigated.

H3.10 In the majority of segmentation studies (more than 50 percent) the choice of the number of clusters is based — at the most — on one run per number of clusters.

H3.11 In the majority of segmentation studies (more than 50 percent) the segmentation solution is not validated.

Of the 65 empirical tourism studies reviewed in 2005, only eight were segmentation studies. Table 3 includes the frequency counts of relevance to test the above hypotheses.

Table 3 here

The information contained in the reviewed articles is insufficient to allow testing of H3.1. The classification of whether data was collected in view of the segmentation has proven to be very subjective, and the two coders involved instead chose to report this hypothesis as not testable. Both H3.2 and H3.3.cannot be rejected because 75 percent of studied articles do not contain an explanation for the measurement of variables, or explain why the variables used as segmentation base were chosen.

Hypothesis H3.4 generated a situation similar to that faced with H3.1: the articles in which the study was reported do not clearly indicate whether the sampling strategy took into account the fact that the study aim was segmentation.

Hypothesis H3.5 cannot be rejected, because 88 percent of studies use multi-category ordinal answer formats. Hypothesis H3.6 cannot be rejected, because 63 percent of studies use factor analysis before segmenting respondents. Hypothesis H3.7 should be rejected, because all of the studies that pre-process data explain why they do so. Unfortunately, the explanation is typically that the number of items has to be reduced. As mentioned above, this problem should have been addressed earlier in the study, and not at the analytic stage, where elimination of items comes at a high price with regard to information loss.

Hypotheses H3.8 and H3.9 cannot be rejected, because no study mentioned the measure of association or investigated data structure before grouping the individuals. Strongly associated with the fact that structure is not investigated before segmenting is the fact that all segmentation studies decided on the number of clusters by using only one computation of each number of clusters in the appropriate range. This indicates that it is likely that many of the findings will have a strong random

component driving the results. Hypothesis H3.10 therefore cannot be rejected. Fifty percent of all segmentation studies used the k-means clustering algorithm, and 25 percent used Ward's clustering.

Hypothesis H.11 cannot be supported. Seven out of eight studies did validate the segmentation solution, although the predominant method is the use of chi-square tests and analysis of variance, which are not corrected for multiple testing (six studies).

While recent studies provide sufficient empirical evidence to validate the claim that the standard research approach in data-driven market segmentation is still prevalent in tourism, several, more general conclusions can also be drawn:

(1) There is a lack of conceptual transparency of segmentation studies generally (do clusters actually exist in the data, or does the solution merely represent one of many possible groupings?).

(2) The explorative and structure-imposing nature of segmentation studies (one computation with one algorithm of a cluster analytic procedure is assumed to deliver the true results) is generally not acknowledged.

(3) The dangers of some of the emerged standards in components of this standard research procedure are not discussed, thus leading to potential misinterpretations of results.

In sum, the review of recently published segmentation studies indicates that the standard approach hypotheses above do exist. Also, there is a clear pattern of

repeating designs that have been published before without explaining why this design is suitable or preferable for the research problem at hand.

Prior reviews in the field support these findings. For example, Frochot and Morrison (2000) review 14 data-driven benefit segmentation studies, and their findings support some of the standard components this paper covers (although they explicitly state that they do not perceive that a common standard has emerged).

(1) Items in surveys are generally not pre-tested (which leads the chosen segmentation base to include large numbers of possibly redundant items).

(2) Data are typically of ordinal format, and use five- or seven-scale points.

(3) Nine out of 14 studies used the so-called factor-cluster approach, mostly Varimax, and rotated the factor solution.

Baumann (2000) reviews 243 segmentation studies from the literature prior to 2000 in the broader area of business studies. Dolnicar (2002) analyzes the tourism-focus subset. According to these reviews, two-thirds of market segmentation studies in tourism use some kind of ordinal data scale, about one-fifth uses binary data, and metric data is virtually not used at all for the variables selected as segmentation bases. The majority of studies factor analyze data sets (43 percent) and use factor scores as segmentation base instead of the original data. Only 38 percent do not pre-process data at all, and about six percent standardize the data.

With respect to the clustering algorithms chosen, 40 percent use k-means clustering, and another 40 percent use Ward's clustering. The reports do not include

reasons and procedures for selecting a particular number of segments in one-third of the studies, more than two-thirds use heuristics and/or a subjective judgment to make this decision, which results in one-third selecting three clusters and another third choosing a four-cluster solution. Approximately half of the studies examine some form of validity, in which 15 percent used discriminant analysis, nine percent compared results with external variables and two percent investigated the match with theories or prior findings. Furthermore, 80 percent of the studies do not mention which similarity measure is used to group respondents, or that the number of variables used is typically not harmonized with the available sample size. Samples sizes (ranging from 46 to 7,996, with a median value of 461) and number of variables in the segmentation base (ranging between three and 56) are uncorrelated.

Tourism researchers strongly adhere to the standard research approach in step-wise, data-driven market segmentation, and empirical tourism research follows this emerging standard procedure much more consistently than other disciplines, for example, marketing. Therefore, the question regarding the origin of this standard approach arises. An attempt to find the roots of the standard procedure can take two approaches: (1) review the pioneering publications in data-driven market segmentation in tourism published in the early 80s; and (2) study in detail articles that are frequently cited as justification for the use of the standard approach.

The pioneering publications review leads to the conclusion that many of the standard components have indeed been used by authors who originally introduced step-wise, data-driven market segmentation into tourism research, and were consequently setting the benchmark for future work. However, many of these pioneering studies did not

provide detailed reasoning or a methodological discussion of the problems associated with this particular approach. For example, Calantone, Schewe, and Allen (1980) use 20 importance attributes from 1,498 respondents, using a six-point response scale. These attributes were first factor analyzed, and then cluster analyzed. The authors referenced Haley (1968) as the methodological source for their work, who represents the original source for benefit segmentation. Haley does not recommend the use of factor analysis for pre-processing in his paper. He mentions that Q-sort factor analysis could be applied as a grouping algorithm, not as a pre-processing tool, but does not discuss many other methodological issues of data-driven segmentation.

Goodrich (1980) segmented 230 respondents based on 11 benefit attributes, which were collected using a seven-point answer format. He pre-processed the 11 benefits using factor analysis, and cluster analyzed the factor scores. He did not provide an explanation or reference for adopting this procedure. Crask (1981) clustered tourists based on factor scores (explaining 57 percent of the variance of the original ordinal data). The stated aim was to determine underlying dimensions based on the 15 variables included in the questionnaire, which measured the importance tourists assigned to certain vacation attributes. He did not provide an explanation of how the 15 motivational variables were derived or why they might be expected to capture the construct adequately. The author did not cite any methodological/statistical source supporting the chosen procedure. Mazanec (1984) used raw data to segment tourists based on benefits. The author used a binary data format, provided a detailed explanation why binary data was deemed preferential to ordinal data and did not compute factor analysis before clustering the data.

The second approach taken to investigate the roots of the standard data-driven segmentation research approach in tourism leads to similar conclusions. At least one author of a data-driven segmentation study in tourism cited the articles discussed below as a justification of the methodology used. Park, Yang, Lee, Jang and Stokowski (2001, p. 58) provides a typical example: 'The factor-cluster combination for segmentation used in this study is a basic type of segmentation methodology (Dimanche et al., 1993) and is widely used in tourism.' However, Dimanche, Havitz and Howard (1993) do not postulate the use of the factor-cluster approach uncritically. They segment tourists based on a particular construct (involvement), for which a scale had been developed and which has repeatedly been shown to have a specific underlying factor structure. The reasoning for using factor analysis before clustering is consequently not because it has any methodological advantages or to follow an established procedure, but because it is a natural result of the structure of the construct as it was found to be more easily measurable. Typically this is not the case in data-driven segmentation studies in tourism, however. Dimanche et al. provide justifications for each step of their analysis, including the choice of the clustering algorithm, which is atypical of most segmentation studies conducted in the last decade. They cite Aldenderfer and Blashfield (1984) and Smith (1989) as sources for using factor-cluster analysis. They also cite Smith (1989) as the source of classifying market segmentation in tourism into *a priori* and factor-cluster, rather than proposing this classification themselves, as indicated in the above citation. Tracing further by following the references used by Dimanche (1983, et al.) requires the study of Aldenderfer and Blashfield (1984) and Smith (1989), with the former representing a

general social sciences handbook on cluster analysis and the latter a tourism-specific analysis handbook.

Aldenderfer and Blashfield do not recommend factor-cluster analysis as a suitable tool for pre-processing. They mention factor analysis as an alternative to cluster analysis for the purpose of developing numerical taxonomies, as do Sokal and Sneath (1963). They refer, however, to Q-sort factor analysis, which is based on the correlation matrix of units (respondents), rather than characteristics (variables, questions), a procedure that (to the author's knowledge) so far has not been applied in tourism. It also does not appear to be particularly suited for data analytic situations in which large numbers of respondents answer only a few questions, as opposed to uses in biology, where a few specimens are classified on the basis of a large number of characteristics. Aldenderfer and Blashfield explicitly point out that there is controversy about whether one should pre-process data at all before clustering.

Smith, however, postulates the existence of two segmentation approaches in tourism research: *a priori* segmentation and factor-cluster segmentation. Factor-cluster segmentation is a term that appears to have been coined by empirical tourism researchers, because it does not occur in other disciplines. This classification is misleading, because it does not mention the vast number of other existing ways to segment respondents in an *a posteriori* or data-driven manner, and which has been described in detail by numerous experts in cluster analysis and numerical taxonomy (Sokal and Sneath, 1963; Aldenderfer and Blashfield, 1984; Everitt, 1993). Also, Smith's discussion of market segmentation analysis fails to cite a single publication of methodological nature to support the claims made and the methods proposed. The

only two references on data-driven segmentation are empirical examples of segmentation studies using the factor-cluster approach, one of which is an internal working paper, the other a study conducted by the author himself.

Frochot and Morrison (2000, p. 32) conclude from their review of benefit segmentation studies that 'it would appear that the combination of factor and cluster analysis seems to be superior due to its effectiveness in reducing sometimes large number of benefit statements to a smaller set of more understandable factors or components.' Interestingly, their conclusion contradicts their statement on page 31, that items that might help to discriminate between segments should not be eliminated. This is what factor analysis typically does: it integrates variables into factors, in which case highly discriminating variables for a particular segment may only carry a low loading value. Or, such variables may simply be dropped due to low loadings on a factor (Nunnally, 1967, recommends a value of 0.3/0.4 for loadings to factors as a criterion for retentions of variables), or because such items may well form their own factor with low explained variance that is likely to be dropped following the most commonly used Kaiser criterion, which recommends the inclusion of all factors with an Eigenvalue above 1 (Stevens, 2002).

Cha, McCleary, and Uysal (1995) also choose the factor-cluster approach, using six factor scores that explain only 50 percent of the original 30 motivational items (this is 50 percent of the information collected from respondents). They do not discuss the consequences of eliminating half of the information contained in the raw data, or the homogeneity assumption of factor analysis, which is in contradiction with the heterogeneity assumption of segmentation. Their argument for factor analyzing raw

data is to identify underlying motivational factors, but do not provide a methodological justification for this approach, nor an explanation why such a large number of motivational items (30) were originally included in the questionnaire.

Shoemaker (1994) also conducts factor analysis, but appears to use the resulting factor scores in a more critical manner. The starting points of his analysis were 39 items. Factor analysis resulted in 12 factors. Shoemaker used those 12 factor scores, but included seven additional items that were not well represented by the factor analysis. This is a sensitive approach — in line with the recommendation by Frochot and Morrison (2000) — which makes use of factor analysis to reduce the dimensionality of the problem. However, given that factor analysis assumes homogeneity and recommends eliminations of variables that are not well represented by the factor solution (but might be essential to identify market segments), he includes additional variables of relevance. Interestingly, this informed use of factor analysis as a pre-processing tool in market segmentation is not used by the authors citing Shoemaker as a reference for factor-cluster analysis.

Sheppard (1996) is a particularly interesting case. His study is cited incorrectly on numerous occasions. Authors of segmentation studies refer to his study to justify the use of factor-cluster segmentation, although Sheppard points out the inconsistency of this approach and states clearly (p. 57) that ‘Cluster analysis on raw item scores, as opposed to factor scores, may produce more accurate or detailed segmentation as it preserves a greater degree of the original data.’ Conducting factor analysis is appropriate, according to Sheppard, if a generalizable instrument is being developed, an instrument for the entire population, assuming homogeneity, not heterogeneity.

In sum, the standard research procedure for exploratory data-driven market segmentation as it is used in tourism research has developed within the field. Market segmentation studies in other disciplines do not reflect the high level of adherence to these standards, and methodological and statistical publications of general nature express reservations with regard to many of the components of the standard research approach outlined above.

4.2 A few things to consider when segmenting markets

First, data-driven segmentation is by definition an exploratory process. If it were a confirmatory process, the definition of an expected segment structure would have to be postulated based on theory, and it would have to be established whether the empirical data significantly deviates from this postulated structure. While this approach can be taken, it does not seem to be the main aim of tourism researchers, whose primary interest is the exploration and description of market information in view of deriving potentially useful segments. Furthermore, typical data-driven methodology does not provide for this option. Even model-based segmentation methods, such as finite mixture models (Wedel and Kamakura, 1998) or latent-class analyses, do not define the structure of postulated segments *ex ante*. They propose models which typically differ in the number of segments, but do not hypothesize a certain nature of segments. Consequently, it is the responsibility of the researcher not to draw to strong conclusions about the results.

The exploration of data leads to the conclusion that one particular segmentation solution should be chosen. However, this is not necessarily the better, or only,

solution. Aldenderfer and Blashfield (1984, p. 14) put it in the context of cluster analysis: 'it is important to recognize the fundamental simplicity of these methods. In doing so, the user is far less likely to make the mistake of reifying the cluster solution.'

Second, the empirical data available to tourism researchers typically does not contain true clusters. Sometimes it does not contain any clear data structure at all. However, any kind of cluster analysis searches for structure. If there is no structure, cluster analysis will impose structure, for example, when clustering 1,000 respondents based on only two variables answered on a 100-point scale. Taking this example further, the answers of the 1,000 respondents to the two questions are not correlated, and each respondent uses a different point on the 100-point scale to answer the questions. This would lead to a two-dimensional plot of the data, where respondents are essentially evenly spread across the two-dimensional plane. In this case, clusters do not exist. Yet if we use k-means clustering, we will obtain clusters that will tend to be spherical in nature and of roughly equal size. Whereas, if we use single-linkage hierarchical clustering, we are likely to find chain-like clusters.

Evaluating the actual structure in the data is important. Consider whether the segmentation aim is true clustering (when density clusters actually do exist in the data), stable clustering (if there is data structure, but not of the density cluster type), or constructive clustering (if no structure exists in the data, as illustrated in the above example). Dolnicar and Leisch (2001) illustrate these options with artificial data of different structures, and provide recommendations of how to assess data structure, which is essentially based on replications of computations, to determine compliance

between independently derived solutions. Constructive clustering is a perfectly legitimate approach, because managerially, it might still be better to focus on a more homogeneous part of the market, even if no true segments exist. Clarity about the nature of the segmentation study undertaken is important because it has major implications on the kind of conclusions that can be drawn, and it provides conceptual transparency to the reader and user of such a study.

With respect to the stages outlined in Figure 2, it is important to consider the following aspects when developing and conducting a step-wise, data-driven segmentation study:

With respect to the measurement of variables for segmentation, the contributions by Churchill (1998), Finn and Kayande (1997), Rossiter (2002), and Finn and Kayande (2005) give an excellent overview of the current discussion on scale development in the social sciences. The guiding principle all these authors emphasize is that scale development is not a process that can be undertaken by following a step-by-step recipe. Clear specifications about what the construct to be measured is and what it is not are essential; item generation should be based on as many different sources as possible in order to assure that no relevant components of the construct are omitted; validity of measures should be assessed carefully; and typical measures of scale quality should be used to point to possible problems with the scale, rather than to take radical measures, possibly at the expense of the ability of the scale to capture the actual construct. For example, Churchill (1998) mentions five ways in which the coefficient alpha can easily be increased while reducing the validity of the scale. Finn and Kayande (2005, p. 13) put it like this: ‘researchers need to think

more carefully about the nature of… constructs, to work much harder up front to generate and select items for their scales, and to design answer categories.'

In terms of answer categories, a choice of three of four **answer formats** are available in the context of segmentation studies: binary, ordinal and metric, which have different advantages and disadvantages. For binary and metric data format, which can be obtained by asking respondents to respond with 'yes' or 'no' (binary) or by asking them for a percentage evaluation or to make a cross at a point on a line that best represents their response (metric), clear measures of distance exist. This is of particular importance for segmentation that is based on distance computations. Metric data (if a continuous underlying construct can reasonably be assumed) allows the use of all data analytic methods, and thus does not impose certain statistical techniques on the researcher. Binary data enables respondents to complete surveys significantly faster (about 30 percent) than ordinal or metric surveys, which reduces fatigue and non-response effects, thus improving data quality. Mazanec (1984, p. 18) states explicitly that he views it as 'preferable to economize on scale levels rather than on the number of benefit items.' Furthermore, 'Measurement of benefits is easiest for the respondent if he is asked only to evaluate a benefit item as being important or not important.' The main disadvantage of these answer format alternatives is that respondents are not very familiar with them, and that they may not reflect the nature of the underlying construct. For example, while constructs such as behavioral intentions appear highly suited for a binary scale, attitudes may require more options to allow respondents to express their views. Suitability of the scale for the construct under study should be assessed in the pre-testing phase of the questionnaire.

Thus the recommendation is to thoroughly investigate what kind of information is actually needed. Researchers should consider making more use of either binary or metric formats. By doing so, they would benefit from avoiding the dangers associated with ordinal scales discussed above.

If ordinal scales emerge as the most suitable answer format for a particular construct, one alternative to avoid measurement problems is to use the summated scale across items if there are multiple underling dimensions, which is the procedure Likert (1932) originally proposed for the scale named after him. By using summated scores, normal distribution can be assumed when analyzing the data. So, if a construct has several dimensions that are measured by subscales, the summated values over the subscales could safely be used as the segmentation base.

The main recommendation with respect to the answer format chosen by the researcher, however, is to use data analytic techniques that are suitable for the nature of the data available. For many statistical techniques that are typically used, rank-based alternative procedures exist and should be used to avoid making wrong assumptions about the data.

Regarding the **sample** of respondents, segmentation researchers must ensure that the number of variables used as segmentation base is not too high, given the number of respondents available and the fact that it is rarely known a priori how well the data is structured. Formann's (1984) recommendation for binary data sets of at least 2^k respondents provides a good guideline for the minimum requirements, where 'k' stands for the number of variables if parametric procedures are used. If non-parametric procedures are used to explore segmentation solutions, no such rule exists.

Sample size requirements increase with decreasing data structure. The appropriateness of the number of variables given a certain sample size can consequently only be assessed when data structure analysis is undertaken. Formann's rule is still a useful guideline to evaluate reasonable sample sizes for certain numbers of variables.

For segmentation purposes — the aim of which is identification or construction of market segments — the sample does not necessarily have to be representative, unless it is important to be able to state the proportion of each segment within total population. It is, however, essential to have the full range of respondents with respect to the construct under study represented in the data. If small niche segments are expected, it may even be recommendable to try to over-sample these respondents to ensure that they can be detected if they do differ from other segments in the hypothesized way.

If the number of items is too large for the available sample size and variables have to be **selected** before clustering, it is advisable to pre-analyze the data (using simple frequency counts or factor analysis) and exclude variables based on the findings, rather than transforming the space, as is the case with using factor scores resulting from factor analysis. For example, Gitelson and Kerstetter (1990) state that if more than 90 percent do not rate a benefit, it should be excluded. While this might be a very general rule that risks excluding a highly discriminating variable, as stated in Frochot and Morrison (2000), a combination of such a frequency criterion with factor analysis results might be preferable. If items have low agreement or disagreement levels and load highly on a factor that includes many other items, it is likely that little information will be lost by excluding it from cluster analysis. Of course, it would be

preferable not to include such items with little additional information value in the questionnaire in the first place. But if the researcher is faced with a data set of this nature, the suggested procedure would be preferable to reduce the number of items, as opposed to factor analyzing and using the factor scores for cluster analysis.

If factor analysis is the preferred method of pre-processing, a procedure that accounts for heterogeneity in the data should be adopted, such as mixtures of factor analyzers, which is discussed in McLachlan and Peel (2000).

However, generally the raw, untransformed data that was collected from respondents should be used for segmentation purposes, because any form of pre-processing leads to a transformation of the segmentation space in which the segments were not postulated originally and relations between variables are changed.

Some transformations may be needed if variables are not the same in nature. For example, if the annual income in dollars is one variable and agreement or disagreement with the statement that 'the natural environment of the destination is important' are both included in the segmentation base, variables would need to be standardized, because the variables with the higher range of values would otherwise have more weight in the grouping process. But this frequently does not appear to be the case, because most tourism researchers use one block of questions to measure the same construct for their analysis (for example, all motivation items, all vacation activities undertaken, or all benefits sought from their vacation). Such items are typically questioned about using the same answer options, making transformation unnecessary.

Another case where transformations may be needed occurs when response styles exist in the data and contaminate the information contained. For example, respondents from certain cultural backgrounds may tend to use extreme values, while others may prefer the middle of the scale. Such contamination may have to be eliminated by transforming the raw data.

If the number of variables is too high and variables need to be eliminated, it is preferable to eliminate them and use the raw data of the remaining items for segmentation analysis, rather than using transformations such as factor scores.

The **clustering algorithm** and underlying **measure of association** should be chosen, while taking into consideration the nature of the data and the known properties of different clustering algorithms.

Although there is still no single optimal solution for determining the best **number of clusters**, two generic approaches can be recommended: (1) clustering can be repeated numerous times with varying numbers of clusters, and the number that renders the most stable results can be chosen; or (2) multiple solutions can be computed and selection is undertaken interactively with management.

After the segmentation solution has been determined it should be **validated**. First, the reliability can be assessed by testing the stability of the solution. If the same grouping emerges when computed numerous times, the segmentation solution is reliable than if different solutions emerge from each computation. Second, the distinctness of the derived solution can be assessed by testing whether the segments are significantly different with respect to information that was not used in the original segmentation

process. For example, benefit segments of tourists would be seen as externally valid if, for example, their expenditure patterns for different vacation activities differ significantly as well. If segments differ in benefits and no other characteristics, the validity and usefulness of the segments should be questioned.

5 CONCLUSIONS

This paper aimed to highlight how several standards have emerged in empirical tourism research that have the potential to undermine the validity of results. Three aspects of empirical tourism research were selected because they are frequently studied, and researchers appear to strongly adhere to emerged standards.

The review of empirical studies published in 2005 in the top three tourism journals led to the conclusions that: (1) multi-category ordinal scales dominate survey research in tourism despite the numerous disadvantages of this answer format and the availability of alternatives; (2) a large number of studies are based on multicultural data sets and ignore the danger of response style effects, potentially distorting their findings; and (3) a standard procedure for step-wise, data-driven segmentation has developed in the field of tourism (and even given a special name: factor-cluster analysis), the major danger of which is the elimination of about half of the information contained in the original data set through factor analysis before segments are identified or constructed.

Further aims of the article were to: (1) increase empirical tourism researchers' awareness that emerged standards are not necessarily the optimal solution; (2) encourage empirical tourism researchers to reflect more critically on their choice of measurement instruments and techniques of data analysis; (3) encourage empirical

tourism researchers to dedicate a few sentences in their manuscript that explain why they have chosen a particular approach, so that an 'explain why' culture will slowly replace the current 'cite why' one; (4) encourage reviewers to request explanations and in so doing motivate researchers to engage in (2) and (3); and (5) strengthen and enhance the field of empirical tourism research so that it becomes more open to new approaches which can be shown to outperform emerged standards.

The three areas that this paper covered are almost certainly not the only ones in which standards emerge in tourism research. Rather, the selected areas reflect topics of interest to the author. To help the field move forwards, experts in other fields should provide similar reviews on topics they are intimately familiar with and share their insights with the wider tourism research community, and the top tourism journals should be open to publishing such manuscripts.

6 REFERENCES

Aldenderfer, M.S., and R.K. Blashfield (1984). *Cluster Analysis*. Beverly Hills: Sage Publications.

Arabie, P., and L. Hubert (1994). 'Cluster Analysis in Marketing Research' In: *Advanced methods of marketing research* edited by R. Bagozzi. Cambridge: Blackwell, 160–189.

Bachman, J.G., and P.M. O'Malley. (1984). 'Yea-Saying, Nay-Saying, and Going to Extremes: Black-White Differences in Response Styles.' *Public Opinion Quarterly 48(2)*: 491–509.

Baumann, R. (2000). *Marktsegmentierung in den Sozial- und Wirtschaftswissenschaften: eine Metaanalyse der Zielsetzungen und Zugänge*. Diploma thesis at Vienna University of Economics and Management Science. Vienna.

Bendig, A.W. (1954). 'Reliability and the Number of Rating Scale Categories.' *Journal of Applied Psychology 38 (1)*: 38–40.

Byrne, B.M., and T.L. Campbell. (1999) 'Cross-cultural Comparisons and the Presumption of Equivalent Measurement and Theoretical Structure - A Look Beneath the Surface.' *Journal of Cross-Cultural Psychology 30(5)*: 555–574.

Buchta, C., E. Dimitriadou, S. Dolničar, F. Leisch & A. Weingessel. (1997) 'A Comparison of Several Cluster Algorithms on Artificial Binary Data Scenarios from Travel Market Segmentation.' Working Paper # 7, SFB 'Adaptive

Information Systems and Modelling in Economics and Management Science', Vienna.

Calantone, R., C. Schewe, and C.T. Allen (1980). 'Targeting Specific Advertising Messages at Tourist Segments.' In *Tourism Marketing and Management* edited by D.E. Hawkins, E.L. Shafer, and J.M., Washington D.C: George Washington University, pp. 133–147.

Cha, S., K.W. McLeary, and M. Uzsal (1995). 'Travel Motivations of Japanese Overseas Travelers: A Factor-Cluster Segmentation Approach.' *Journal of Travel Research 34(1)*: 33–39.

Chang, L. (1994). 'A Psychometric Evaluation of Four-point and Six-point Likert-type Scales in Relation to Reliability and Validity.' *Applied Psychological Measurement 18*: 205–215.

Cheung, G.W., and R.B. Rensvold. (2000). 'Assessing Extreme and Acquiescence Response Sets in Cross-Cultural Research using Structural Equation Modeling.' *Journal of Cross-Cultural Psychology 31(2)*: 187–212.

Chun, K. T., J.B. Campbell, and J.H. Yoo (1974). 'Extreme Response Style in Cross-Cultural Research – Reminder.' *Journal of Cross-Cultural Psychology 5(4)*: 465–480.

Churchill, G.A. (1979). 'A Paradigm for Developing Better Measures of Marketing Constructs.' *Journal of Marketing Research 16*: 64–73.

Churchill, G.A. (1998). 'Measurement in Marketing: Time to Refocus?' In J. D. Hess, and K. B. Monroe, eds *Proceedings of the 14th Paul D. Converse Symposium*, Chicago: American Marketing Association, pp. 25–41

Clarke III, I. (2000). 'Extreme Response Style in Cross Cultural Research: An Empirical Investigation.' *Journal of Social Behaviour and Personality 15(1)*: 137–152.

Clarke III, I. (2001) 'Extreme Response Style in Cross-Cultural Research.' *International Marketing Review 18(3)*: 301–324.

Cox, E.P. (1980). 'The Optimal Number of Response Alternatives for a Scale: A Review.' *Journal of Marketing Research 17 (4)*: 407–422.

Crask, M. (1981). 'Segmenting the Vacationer Market: Identifying the Vacation Preferences, Demographics, and Magazine Readership of Each Group.' *Journal of Travel Research 20*: 20–34.

Cronbach, L. (1950). 'Further Evidence on Response Sets and Test Design.' *Educational and Psychological Measurement 10*:3–31.

Cronbach, L. J. (1946). 'Response Sets and Test Validity.' *Educational and Psychological Measurement 6*: 475–494.

Cronbach, L.J. (1951). 'Coefficient Alpha and the Internal Structure of Tests.' *Psychometrika 16*: 297–334.

Dimanche, F., M.E. Havitz, and D.R. Howard. (1993). 'Consumer Involvement Profiles as a Tourism Segmentation Tool.' *Journal of Travel and Tourism Marketing 1(4)*: 33–52.

Dimitriadou, E, S. Dolnicar and A. Weingessel. (2002). 'An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets.' *Psychometrika 67(1)*: 137–160.

Dolnicar, S. (2003). 'Simplifying Three-way Questionnaires - Do the Advantages of Binary Answer Categories Compensate for the Loss of Information?' ANZMAC CD Proceedings.

Dolnicar, S. (2002). 'Review of Data-Driven Market Segmentation in Tourism.' *Journal of Travel and Tourism Marketing 12(1)*: 1–22.

Dolnicar, S. (2004). 'Beyond 'Commonsense Segmentation' – a Systematics of Segmentation Approaches in Tourism.' *Journal of Travel Research 42(3)*: 244–250.

Dolnicar, S. and B. Grün (in press). 'Cross-Cultural Differences in Survey Response Patterns.' *International Marketing Review.*

Dolnicar, S. and B. Grün (under review). 'Culture-Specific Response Style Effects.'

Dolnicar, S and F. Leisch. (2000). 'Behavioral Market Segmentation Using the Bagged Clustering Approach Based on Binary guest Survey Data: Exploring and Visualizing Unobserved Heterogeneity.' *Tourism Analysis 5(2–4)*: 163–170.

Dolnicar, S. and F. Leisch. (2003) 'Winter Tourist Segments in Austria- Identifying Stable Vacation Styles for Target Marketing Action.' *Journal of Travel Research 41(3)*: 281–293 (Charles Goeldner Article of Excellence Award).

Dolnicar, S and Leisch, F. (2004) 'Segmenting Markets by Bagged Clustering.' *Australasian Marketing Journal*, 12(1), 51–65.

Dolnicar, S. and F. Leisch (2001) Knowing What You Get - a Conceptual Clustering Framework for Increased Transparency of Market Segmentation Studies. Presented at the Marketing Science 2001 (Abstract Proceedings available).

Dolnicar, S., B. Grün, and F. Leisch (2004). 'Time Efficient Brand Image Measurement - Is Binary Format Sufficient to Gain the Market Insight Required?' CD Proceedings of the 33st EMAC conference.

Everitt, B. S. (1979). 'Unresolved Problems in Cluster Analysis.' *Biometrica 35*:169–81.

Everitt, B. S. (1993). *Cluster Analysis*. New York: Halsted Press.

Feldman, J.M. and J.G. Lynch Jr. (1988) 'Self-Generated Validity and Other Effects of Measurement on Belief, Attitude, Intention and Behaviour.' *Journal of Applied Psychology 73(3):* 421-435.

Finn, A., and U. Kayande (1997). 'Reliability Assessment and Optimization of Marketing Measurement.' *Journal of Marketing Research 34(2)*: 262–276.

Finn, A., and U. Kayande (2005). 'How Fine is C-OAR-SE? A Generalizability Theory Perspective on Rossiter's Procedure.' *International Journal of Research in Marketing 22*: 11–21.

Finn, R.H. (1972). 'Effects of Some Variations in Rating Scale Characteristics on the Means and Reliabilities of Ratings.' *Educational and Psychological Measurement 32 (2)*: 255–265.

Formann, A.K. (1984). *Die Latent-Class-Analyse: Einführung in die Theorie und Anwendung.* Weinheim: Beltz.

Frochot, I., and A.M. Morrison. (2000). 'Benefit Segmentation: A Review of its Application to Travel and Tourism Research.' *Journal of Travel and Tourism Marketing 9(4)*: 21–45.

Gitelson, R.J., and D.L. Kerstetter (1990). 'The Relationship Between Sociodemographic Variables, Benefits Sought and Subsequent Vacation Behavior: A Case Study.' *Journal of Travel Research 28*:24–29.

Goodrich, J. (1980). 'Benefit Segmentation of US International Travelers: An Empirical Study with American Express.' In *Tourism Marketing and Management* edited by D.E. Hawkins, E.L. Shafer, and J.M., Washington D.C: George Washington University, pp. 133–147.

Green, P.E. and V.R. Rao (1970). 'Rating Scales and Information Recovery---How Many Scales and Response Categories to Use?' *Journal of Marketing*, *34 (July)*: 33–39.

Greenleaf, E.A. (1992a). 'Improving Rating Scale Measures by Detecting and Correcting Bias Components in Some Response Styles.' *Journal of Marketing Research 29(May)*: 176–188.

Greenleaf, E.A. (1992b). 'Measuring Extreme Response Style.' *Public Opinion Quarterly 56(3)*: 328–351.

Haley, R.J. (1968). 'Benefit Segmentation: A Decision-Oriented Research Tool.' *Journal of Marketing 32*: 30–35.

Hancock, G. R. and A. J. Klockars (1991). 'The Effect of Scale Manipulations on Validity: Targeting Frequency Rating Scales for Anticipated Performance Levels.' *Applied Ergonomics 22 (3)*: 147–154.

Hui, C. H., and H.C. Triandis (1989). 'Effects of Culture and Response Format on Extreme Response Style.' *Journal of Cross-Cultural Psychology 20(3)*: 296–309.

Jacoby, J. and M.S. Matell (1971). 'Three-Point Likert Scales Are Good Enough.' *Journal of Marketing Research 8*: 495–500.

Johnson, M. D., D.R. Lehmann, & D.R. Horne. (1990). 'The Effects of Fatigue on Judgments of Interproduct Similarity.' *International Journal of Research in Marketing 7(1)*: 35–43.

Jones, R.R. (1968). *Differences in Response Consistency and Subjects' Preferences for Three Personality Inventory Response Formats*, Proceedings of the 76th Annual Convention of the American Psychological Association, 247–248.

Kampen, J. and M. Swyngedouw (2000). 'The Multi-category Controversy Revisited.' *Quality & Quantity 34 (1)*: 87–102.

Ketchen D.J. jr., and C.L. Shook (1996). 'The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique.' *Strategic Management Journal 17*: 441–458.

Komorita, S. S. (1963). 'Attitude Content, Intensity, and the Neutral Point on a Likert Scale.' *Journal of Social Psychology 61*: 327–334.

Komorita, S. S. and W.K. Graham (1965). 'Number of Scale Points and the Reliability of Scales.' *Educational and Psychological Measurement 25(4)*: 987–995.

Kozak, M, E. Bigne and L. Endreu. (2003). 'Limitations of Cross-Cultural Satisfaction Research and Recommending Alternative Methods.' *Journal of Quality Assurance in Hospitality and Tourism 4(3–4)*: 37–59

Kuhn, T.S. (1970) *The Structure of Scientific Revolution*. Second Edition. Chicago: University of Chicago Press.

Likert, R. (1932). 'A Technique for the Measurement of Attitudes.' *Archives of Psychology 140*: 44–53.

Loken, B., P. Pirie, K.A. Virnig, R.L. Hinkle, and C.T. Salmon (1987). 'The Use of 0–10 Scales in Telephone Surveys.' *Journal of the Market Research Society 29 (3)*: 353–362.

Marin, G., R.J. Gamba, & B.V. Marin. (1992). 'Extreme Response Style and Acquiescence among Hispanics - The Role of Acculturation and Education.' *Journal of Cross-Cultural Psychology 23(4)*: 498–509.

Martin, W.S., B. Fruchter and W.J. Mathis (1974). 'An Investigation of the Effect of the Number of Scale Intervals on Principal Components Factor Analysis.' *Educational and Psychological Measurement 34*: 537–545.

Matell, M.S. and J. Jacoby. (1971). 'Is There an Optimal Number of Alternatives for Likert Scale Items? Study I: Reliability and Validity.' *Educational and Psychological Measurement 31*: 657–74.

Mazanec, J. (2000). Market Segmentation. In: *Encyclopedia of Tourism*. J. Jafari, ed. London: Routledge.

Mazanec, J.A. (1984). 'How to Detect Travel Market Segments: A Clustering Approach.' *Journal of Travel Research 23(1)*: 17–21.

McLachlan, G., and D. Peel (2000). *Finite Mixture Models*. New York: John Wiley and Sons.

Milligan, G.W. (1980). 'An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms'. *Psychometrika 45*:325–342.

Milligan, G.W. (1996). Clustering Validation: Results and Implications for Applied Analyses. In: *Clustering and Classification*. P. Arabie and L.J. Hubert, eds. River Edge: World Scientific Publ.

Myers, J.H., and E. Tauber (1977). *Market structure analysis*. American Marketing Association: Chicago.

Nunnally, J. C. (1967). *Psychometric Theory*. New York: McGraw-Hill, 1st edition

Oaster, T.R.F. (1989). 'Number of Alternatives Per Choice Point and Stability of Likert-type Scales.' *Perceptual and Motor Skills 68*: 549–550.

Park, M., X. Yang, B. Lee, H-C. Jang, and P.A. Stokowski (2002). 'Segmenting Casino Gamblers by Involvement Profiles: a Colorado Example.' *Tourism Management 23(1)*: 55–65.

Peabody, D. (1962). 'Two Components in Bipolar Scales: Direction and Extremeness.' *Psychological Review 69 (2)*: 65–73.

Percy, L. (1976). 'An Argument in Support of Ordinary Factor Analysis of Dichotomous Variables.' In *Advances in Consumer Research* Vol. III.

Preston, C.C. and A.M. Colman (2000). 'Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences.' *Acta Psychologica 104*: 1–15.

Punj, G., and D.W. Stewart (1983). 'Cluster Analysis in Marketing Research: Review and Suggestions for Application.' *Journal of Marketing Research 20*: 134–148.

Ramsay, J.O. (1973). 'The Effect of Number of Categories in Rating Scales on Precision of Estimation of Scale Values.' *Psychometrika 37*: 513–532.

Remington, M., P.J. Tyrer, J. Newson-Smith and D.V. Cicchetti (1979). 'Comparative Reliability of Categorical and Analogue Rating Scales in the Assessment of Psychiatric Symptomatology.' *Psychological Medicine 9*: 765–770.

Rossiter, J.R. (2002). 'The C-OAR-SE Procedure for Scale Development in Marketing.' *International Journal of Research in Marketing 19*: 305–335.

Roster, C.A., R. Rogers, and G. Albaum. (in press). 'A Cross-Cultural/National Study of Respondents' Use of Extreme Categories.' *Journal of Cross-Cultural Psychology.*

Rungie, C., G. Laurent, et al. (2005). 'Measuring and Modeling the (Limited) Reliability of Free Choice Attitude Question.' *International Journal of Research in Marketing 22*: 309–318

Scharf, A. (1991). *Konkurrierende Produkte aus Konsumentensicht*. Frankfurt: Verlag Harri Deutsch.

Sekaran, U. (1983). 'Methodological and Theoretical Issues and Advancements on Cross-Cultural Research.' *Journal of International Business Studies 14(2)*: 61–73.

Sheppard, A.G. (1996). 'The Sequence of Factor Analysis and Cluster Analysis: Differences in Segmentation and Dimensionality Through the Use of Raw and Factor Scores.' *Tourism Analysis 1*:49–57.

Shoemaker, S. (1994). 'Segmentation of the US Travel Market according to Benefits Realized.' *Journal of Travel Research 32(3)*: 8–21.

Smith, S.L.J. (1989). *Tourism Analysis: a Handbook*. Harlow, England: Longman.

Smith, A.M. and N.L. Reynolds. (2002). 'Measuring Cross-Cultural Service Quality: A Framework for Assessment.' *International Marketing Review 19(4–5)*: 450–481.

Sokal, R.R., and P.H.A. Sneath (1963). *Principles of numerical taxonomy*. San Franscisco: Freeman.

Stevens, J. (2002). *Applied Multivariate Statistics for the Social Sciences* (4 ed). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Symonds, P.M. (1924). 'On the Loss of Reliability in Ratings Due to Coarseness of the Scale.' *Journal of Experimental Psychology 7*: 456–461.

van de Vijver, F. J. R., & Y.H. Poortinga. (2002). 'Structural Equivalence in Multilevel Research.' *Journal of Cross-Cultural Psychology 33(2)*: 141–156.

Van der Eijk, C. (2001). 'Measuring Agreement in Ordered Rating Scales.' *Quality & Quantity 35 (3)*: 325–41.

van Herk, H., Y.H. Poortinga, and T.M.M. Verhallen (2004). 'Response Styles in Rating Scales - Evidence of Method Bias in Data From Six EU Countries.' *Journal of Cross-Cultural Psychology 35(3)*: 346–360.

Watkins, D., and S. Cheung. (1995). 'Culture, Gender and Response Bias.' *Journal of Cross-Cultural Psychology 26(5)*: 490–504.

Watson, D. (1992) 'Correcting for Acquiescent Response Bias in the Absence of a Balanced Scale: An Application to Class Consciousness'. *Sociological Methods and Research 21(1)*: 52–88.

Wedel, M., and W. Kamakura (1998). *Market Segmentation - Conceptual and Methodological Foundations*. Boston: Kluwer Academic Publishers.

Welkenhuysen-Gybels, J., J. Billiet, & B. Cambre. (2003). 'Adjustment for Acquiescence in the Assessment of the Construct Equivalence of Likert-Type Score Items.' *Journal of Cross-Cultural Psychology 34*(6): 702–722.

7 APPENDIX 1: REVIEWED LITERATURE

Alexandros, A and S. Jaffry (2005). 'Stated Preferences for Two Cretan Heritage Attractions.' *Annals of Tourism Research 32(4):*985

Andereck, K.L., K.M. Valentine, et al. (2005). 'Residents' Perceptions of Community Tourism Impacts.' *Annals of Tourism Research 32(4)*: 1056

Apostolakis, A. and S. Jaffry (2005). 'A Choice Modeling Application for Greek Heritage Attractions.' *Journal of Travel Research 43(3):* 309–318.

Ballantyne, R., N. Carr, et al. (2005). 'Between the Flags: An Assessment of Domestic and International University Students & rsquo; Knowledge of Beach Safety in Australia.' *Tourism Management 26*(4): 617.

Baloglu, S. and C. Love (2005). 'Association Meeting Planners' Perceptions and Intentions for Five Major US Convention Cities: The Structured and Unstructured Images.' *Tourism Management 26(5*): 743.

Beldona, S. (2005). 'Cohort Analysis of Online Travel Information Search Behavior: 1995–2000.' *Journal of Travel Research 44*(2): 135–142.

Bigne, J. E., L. Andreu, et al. (2005). 'The Theme Park Experience: An Analysis of Pleasure, Arousal and Satisfaction.' *Tourism Management 26*(6): 833.

Bigne, J. E., L. Andreu, et al. (2005). 'Quality Market Orientation: Tourist Agencies' Perceived Effects.' *Annals of Tourism Research 32(4)*: 1022.

Blain, C., S. E. Levy, et al. (2005). 'Destination Branding: Insights and Practices from Destination Management Organizations.' *Journal of Travel Research 43(4)*: 328–338.

Bloom, J. Z. (2005). 'Market Segmentation: A Neural Network Application.' *Annals of Tourism Research 32(1)*: 93.

Bonn, M. A., S. M. Joseph, et al. (2005). 'International versus Domestic Visitors: An Examination of Destination Image Perceptions.' J*ournal of Travel Research 43(3):* 294–301.

Brown, G. and D. Getz (2005). 'Linking Wine Preferences to the Choice of Wine Tourism Destinations.' *Journal of Travel Research 43(3)*: 266–276.

Carr, N. (2005). 'Poverty, Debt, and Conspicuous Consumption: University Students Tourism Experiences.' *Tourism Management 26(5): 797*.

Chen, H. M. and C. H. Tseng (2005). 'The Performance of Marketing Alliances Between the Tourism Industry and Credit Card Issuing Banks in Taiwan.' *Tourism Management 26(1):* 15.

Chhabra, D. (2005). 'Defining Authenticity and Its Determinants: Toward an Authenticity Flow Model.' *Journal of Travel Research 44(1)*: 64–73.

Choi, H. S. C. and E. Sirakaya (2005). 'Measuring Residents' Attitude toward Sustainable Tourism: Development of Sustainable Tourism Attitude Scale.' *Journal of Travel Research 43(4)*: 380–394.

Cole, S. T. (2005). 'Comparing Mail and Web-Based Survey Distribution Methods: Results of Surveys to Leisure Travel Retailers.' *Journal of Travel Research 43(4):* 422–430.

Connell, J. (2005). 'Toddlers, Tourism and Tobermory: Destination Marketing Issues and Television-induced Tourism.' *Tourism Management 26(5*): 763.

Daruwalla, P. and S. Darcy (2005). 'Personal and Societal Attitudes to Disability.' *Annals of Tourism Research 32(3*): 549.

Duman, T. and A. S. Mattila (2005). 'The Role of Affective Factors on Perceived Cruise Vacation Value.' *Tourism Management 26(3):* 311.

Enright, M. J. and J. Newton (2005). 'Determinants of Tourism Destination Competitiveness in Asia Pacific: Comprehensiveness and Universality.' *Journal of Travel Research 43(4)*: 339–350.

Espino-Rodriguez, T. F. and V. Padron-Robaina (2005). 'A Resource-based View of Outsourcing and its Implications for Organizational Performance in the Hotel Sector.' *Tourism Management 26(5)*: 707.

Fleischer, A. and A. Tchetchik (2005). 'Does Rural Tourism Benefit from Agriculture?' *Tourism Management 26(4*): 493.

Frochot, I. (2005). 'A Benefit Segmentation of Tourists in Rural Areas: A Scottish Perspective.' *Tourism Management 26(3)*: 335.

Haley, A. J., T. Snaith, et al. (2005). 'The Social Impacts of Tourism: A Case Study of Bath, UK.' *Annals of Tourism Research 32(3)*: 647.

Hall, C. M. (2005). 'Biosecurity and Wine Tourism.' *Tourism Management 26(6)*: 931.

Hou, J. S., C. H. Lin, et al. (2005). 'Antecedents of Attachment to a Cultural Tourism Destination: The Case of Hakka and Non-Hakka Taiwanese Visitors to Pei-Pu, Taiwan.' *Journal of Travel Research 44(2*): 221–233.

Hwang, S. N., C. Lee, et al. (2005). 'The Relationship Among Tourists & rsquo: Involvement, Place Attachment and Interpretation Satisfaction in Taiwan & rsquo's National Parks.' *Tourism Management 26(2*): 143.

Johnson, C. and M. Vanetti (2005). 'Locational Strategies of International Hotel Chains.' *Annals of Tourism Research 32(4)*: 1077.

Jones, S. (2005). 'Community-Based Ecotourism: The Significance of Social Capital.' *Annals of Tourism Research 32(2)*: 303.

Kang, I., S. Jeon, et al. (2005). 'Investigating Structural Relations Affecting the Effectiveness of Service Management.' *Tourism Management 26(3)*: 301.

Kang, S. K. and C. H. C. Hsu (2005). 'Dyadic Consensus on Family Vacation Destination Selection.' *Tourism Management 26(4)*: 571.

Kim, D. Y., Y. H. Hwang, et al. (2005). 'Modeling Tourism Advertising Effectiveness.' *Journal of Travel Research 44(1)*: 42–49.

Kim, H. and W. G. Kim (2005). 'The Relationship Between Brand Equity and Firms&rsquo: Performance in Luxury Hotels and Chain Restaurants.' *Tourism Management 26(4)*: 549.

Kim, S. S., H. Chun, et al. (2005). 'Positioning Analysis of Overseas Golf Tour Destinations by Korean Golf Tourists.' *Tourism Management 26(6)*: 905.

Kim, S. S. and A. M. Morrsion (2005). 'Change of Images of South Korea Among Foreign Tourists After the 2002 FIFA World Cup.' *Tourism Management 26(2)*: 233.

Kim, S. S. and J. F. Petrick (2005). 'Residents & rsquo: Perceptions on Impacts of the FIFA 2002 World Cup: The Case of Seoul as a Host City.' *Tourism Management 26(1):* 25.

Kim, S. S. and B. Prideaux (2005). 'Marketing Implications Arising from a Comparative Study of International Pleasure Tourist Motivations and Other Travel-related Characteristics of Visitors to Korea.' *Tourism Management 26(3)*: 347.

Kwan, A. V. C. and G. McCartney (2005). 'Mapping Resident Perceptions of Gaming Impact.' *Journal of Travel Research 44(2*): 177–187.

Lawton, L. J. (2005). 'Resident Perceptions of Tourist Attractions on the Gold Coast of Australia.' *Journal of Travel Research 44(2*): 188–200.

Lee, C. K., Y. K. Lee, et al. (2005). 'Korea's Destination Image Formed by the 2002 World Cup.' *Annals of Tourism Research 32(4)*: 839.

Lee, C. K. and T. Taylor (2005). 'Critical Reflections on the Economic Impact Assessment of a Mega-event: The Case of 2002 FIFA World Cup.' *Tourism Management 26(4)*: 595.

Litvin, S. W. (2005). 'Streetscape Improvements in an Historic Tourist City: A Second Visit to King Street, Charleston, South Carolina.' *Tourism Management 26(3): 421*.

Mohsin, A. (2005). 'Tourist Attitudes and Destination Marketing--The Case of Australia's Northern Territory and Malaysia.' *Tourism Management 26(5):* 723.

Needham, M. D. and R. B. Rollins (2005). 'Interest Group Standards for Recreation and Tourism Impacts at Ski Areas in the Summer.' *Tourism Management 26(1)*: 1.

O'Leary, S. and J. Deegan (2005). 'Ireland's Image as a Tourism Destination in France: Attribute Importance and Performance.' *Journal of Travel Research 43(3*): 247–256.

Okumus, F., M. Altinay, et al. (2005). 'The Impact of Turkey's Economic Crisis of February 2001 on the Tourism Industry in Northern Cyprus.' *Tourism Management 26(1):* 95.

Okumus, F. and K. Karamustafa (2005). 'Impact of an Economic Crisis: Evidence from Turkey.' *Annals of Tourism Research 32(4)*: 942.

Page, S. J., T. Bentley, et al. (2005). 'Tourist Safety in New Zealand and Scotland.' *Annals of Tourism Research 32(1)*: 150.

Page, S. J., T. A. Bentley, et al. (2005). 'Scoping the Nature and Extent of Adventure Tourism Operations in Scotland: How Safe Are They?' *Tourism Management 26(3)*: 381.

Pearce, P. L. and U. I. Lee (2005). 'Developing the Travel Career Approach to Tourist Motivation.' *Journal of Travel Research 43(3)*: 226–237.

Perez, E. A. and J. R. Nadal (2005). 'Host Community Perceptions a Cluster Analysis.' *Annals of Tourism Research 32(4)*: 925.

Petrick, J. F. (2005). 'Segmenting Cruise Passengers with Price Sensitivity.' *Tourism Management 26(5)*: 753.

Petrzelka, P., R. S. Krannich, et al. (2005). 'Rural Tourism and Gendered Nuances.' *Annals of Tourism Research 32(4*): 1121.

Plummer, R., D. Telfer, et al. (2005). 'Beer Tourism in Canada Along the Waterloo-Wellington Ale Trail.' *Tourism Management 26(3)*: 447.

Pyo, S. (2005). 'Knowledge Map for Tourist Destinations--Needs and Implications.' *Tourism Management 26(4)*: 583.

Qu, R., C. Ennew, et al. (2005). 'The Impact of Regulation and Ownership Structure on Market Orientation in the Tourism Industry in China.' *Tourism Management 26(6)*: 939.

Reichel, A. and S. Haber (2005). 'A Three-sector Comparison of the Business Performance of Small Tourism Enterprises: an Exploratory Study.' *Tourism Management 26(5)*: 681.

Reisinger, Y. and F. Mavondo (2005). 'Travel Anxiety and Intentions to Travel Internationally: Implications of Travel Risk Perception.' *Journal of Travel Research 43(3)*: 212–225.

Sarigollu, E. and R. Huang (2005). 'Benefits Segmentation of Visitors to Latin America.' *Journal of Travel Research 43(3*): 277–293.

Sheehan, L. R. and J. R. B. Ritchie (2005). 'Destination Stakeholders Exploring Identity and Salience.' *Annals of Tourism Research 32(3):* 711.

Sirakaya, E., D. Delen, et al. (2005). 'Forecasting Gaming Referenda.' *Annals of Tourism Research 32(1)*: 127.

Suh, Y. K. and L. McAvoy (2005). 'Preferences and Trip Expenditures--A Conjoint Analysis of Visitors to Seoul, Korea.' *Tourism Management 26(3)*: 325.

Tsai, H. T., L. Huang, et al. (2005). 'Emerging E-commerce Development Model for Taiwanese Travel Agencies.' *Tourism Management 26(5)*: 787.

Yoon, Y. and M. Uysal (2005). 'An Examination of the Effects of Motivation and Satisfaction on Destination Loyalty: A Structural Model.' *Tourism Management 26(1)*: 45.

TABLES AND FIGURES

Figure 1: Examples of answer formats

NOMINAL	Which is your country of residence? ❒ Austria ❒ USA ❒ Australia
BINARY	Do you think Paris is expensive? ❒ Yes ❒ No
ORDINAL	Did you perceive public transportation in Paris as ❒ Very reliable ❒ Reliable ❒ Unreliable ❒ Very unreliable.
METRIC	How many days will you spend in Paris during this trip?

TABLE 1

ANSWER FORMATS USED IN EMPIRICAL TOURISM RESEARCH

Component of standard research approach	Alternatives	Frequency	Percent
Answer format used	Nominal only	5	8
	Binary only	1	2
	Ordinal only	49	75
	Metric only	0	0
	More than one of the above	9	14
Specific answer format	Likert scale	43	66
	Semantic differential	3	5
Cited prior work to justify use of response format	yes	11	17
	no	54	83
Explanation provided	yes	3	5
	no	62	95
Dangers discussed	yes	2	3
	no	63	97
Method of data analysis	Factor analysis	13	20
	Factor analysis combined with other analyses	23	35
	Descriptive statistics	9	14
	Logistic regression	4	6
	Cluster analysis	2	3
	Discrete choice modeling	2	3

TABLE 2

APPROACHES IN EMPIRICAL CROSS-CULTURAL TOURISM RESEARCH

Component of standard research approach	Alternatives	Frequency	Percent
Multicultural data set used among all empirical studies	yes	10	34
	no	0	0
	not clear	10	15
Comparison of countries/cultural backgrounds	yes	8	36
	no	14	64
Answer format used	nominal	2	9
	binary	0	0
	ordinal	19	86
	metric	1	5
Potential problems mentioned	yes	2	9
	no	20	91
Extent of contamination assessed	yes	2	9
	no	20	91
Corrected for contamination	yes	2	0
	no	20	91

Figure 2: Outline of stages of a step-wise, data-driven market segmentation study

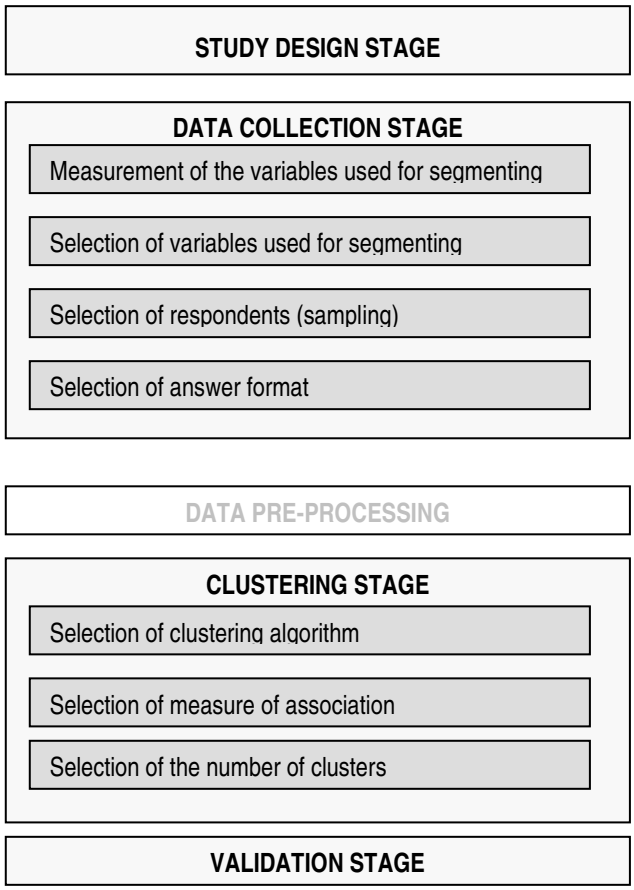

Figure 3: Data set example

	Want to rest	*Want excitement*	*Want security*
Respondent 1	*0*	*1*	*0*
Respondent 2	*0*	*1*	*1*
Respondent 3	*1*	*0*	*0*

TABLE 3

APPROACHES IN DATA-DRIVEN MARKET SEGMENTATION OF TOURISTS

Component of standard research approach	Alternatives	Frequency	Percent
Explanation for measurement of variables provided	Yes	2	25
	No	6	75
Explanation for selection of variables provided	Yes	2	25
	No	6	75
Answer format	Ordinal	7	88
	More than one	1	13
Pre-processing	No	2	25
	Yes _ factor analysis	5	63
	Yes_other	1	13
Explanation for pre-processing if pre-processed	Yes	6	100
Explanation for selection of clustering algorithm	Yes	0	0
	No	6	75
	Cited another author	2	25
Measure of association stated	Yes	0	0
	No	8	100
Data structure investigated	Yes	0	0
	No	8	100
Number of clusters selection	Based on 1 run per number of clusters	8	100
	Based on data structure investigation	0	0
validation	Yes	7	
	No	1	