Journal of Travel Research
52(5) 551–574
© 2013 SAGE Publications
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0047287513479842
jtr.sagepub.com

Asking Good Survey Questions

Sara Dolnicar[1]

Abstract

Surveys are the main instrument of data collection in empirical tourism research. The quality of the collected data depends on the quality of survey questions asked. This paper provides theory- and evidence-based guidance on designing good survey questions to increase the validity of findings resulting from survey research in tourism.

Keywords

survey research, questionnaire design, survey questions, measurement, scale development, Likert scale, DLF IIST, BIT

> "We are the ones who will hear," said Phouchg, "the answer to the great question of Life . . ., The Universe . . . And Everything!" . . .
> "An answer for you?" interrupted Deep Thought majestically. "Yes. I have."
> The two men shivered with expectancy. Their waiting had not been in vain. . . . Both of the men had been trained for this moment, their lives has been a preparation for it, they had been selected at birth as those who would witness the answer, but even so they found themselves gasping and squirming like excited children.
> "And you're ready to give it to us?" urged Loonquawl.
> "I am."
> "Now?"
> "Now," said Deep Thought.
> They both licked their dry lips. . . . The two men fidgeted. The tension was unbearable. . . .
> "Forty-two," said Deep Thought, with infinite majesty and calm. . . .
> It was a long time before anyone spoke. . . .
> "Forty-two!" yelled Loonquawl. "Is that all you've got to show for seven and a half million years' work?"
> "I checked it thoroughly," said the computer, "and that quite definitely is the answer. **I think the problem, to be quite honest with you, is that you've never actually known what the question is**."
> "But it was the Great Question! The Ultimate Question of Life, the Universe and Everything," howled Loonquawl.
> "Yes," said Deep Thought with the air of one who suffers fools gladly, "but what actually *is* it?"
> A slow stupefied silence crept over the men as they stared at the computer and then at each other.
> "Well, you know, it's just Everything . . . everything . . .," offered Phouchg weakly.
> "Exactly!" said Deep Thought. "**So, once you do know what the question actually is, you'll know what the answer means**."
>
> Douglas Adams, *The Hitchhiker's Guide to the Galaxy* (1979, pp. 127–29)

Introduction

Deep Thought is not called Deep Thought without a reason. This smart computer from the *Hitchhiker's Guide to the Galaxy* expressed perfectly the problem of not putting deep enough thought into formulating survey questions, and that answers to carelessly formulated questions are useless, or as Jacoby (1978, p. 87) put it: "meaningless and potentially misleading junk." So: are we doing our best to ask good survey questions or are we just creating meaningless and potentially misleading junk? Below are three responses to this question provided by measurement experts in the social sciences:

> We are being strangled by our bad measures. Let's identify them and get rid of them. (Jacoby 1978, p. 92)

> I allege that all the findings in the social sciences based on Likert items and Semantic Differential items are suspect—and this means the majority of findings! (Rossiter 2011, p. 79)

> It would not be surprising if 90% of the findings and lack of findings proved to be wrong. (Kollat, Blackwell, and Engel 1972, p. 577)

Why are we still using bad measures? One possible reason is that social scientists are not provided with clear guidance on how to develop good survey measures. Instead, recommendations about measurement in the social sciences are scattered across disciplines. Even if successfully retrieved,

[1] University of Queensland, Australia

Corresponding Author:
Sara Dolnicar, University of Queensland, Australia.
Email: s.dolnicar@uq.edu.au

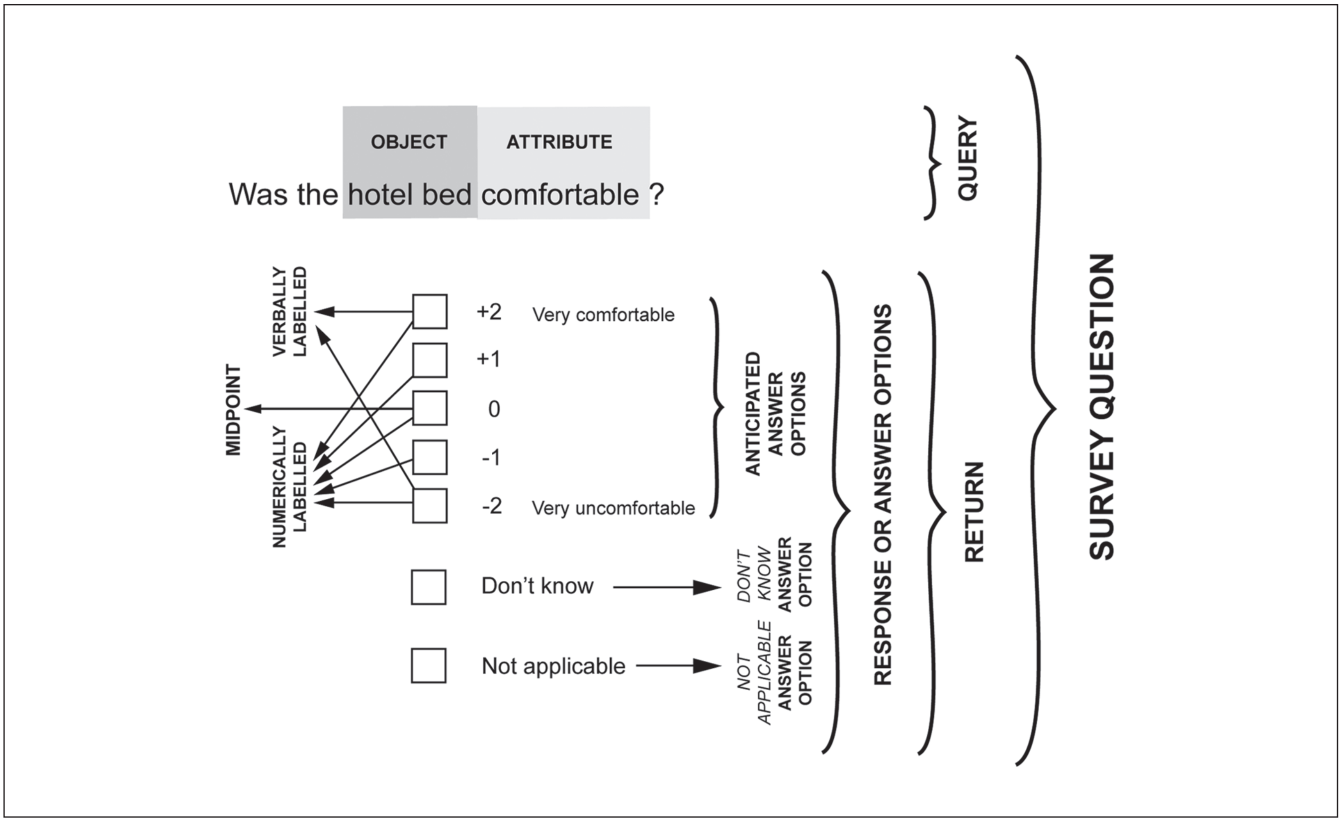


Figure 1. Elements of survey questions.

recommendations are often contradictory, making it difficult to derive clear conclusions about various measurement challenges. The present paper provides theory- and evidence-based guidance relating to the survey researcher's typical challenges, specifically:

1. How to define what is being measured?
2. How many questions to ask?
3. How to ask a question (the query)?
4. How to allow respondents to answer (the return)?

For each of these areas, current practice in tourism research is assessed by reviewing 78 survey studies recently published in the leading generalist tourism journals (*Annals of Tourism Research, Journal of Travel Research*, and *Tourism Management*[1]).

Figure 1 provides an overview of the key elements of survey questions discussed in this article. As can be seen, a *survey question* consists of two parts: the *query* and the *return*. The *query* can be a statement (e.g., "Wollongong is a perfect destination for a family holiday") or a question (in Figure 1: "Was the hotel bed comfortable?"). The query typically contains information about the *object* that is being measured (in Figure 1: the hotel bed) and the *attribute* of the object that is of interest (in Figure 1: comfortable). Both object and attribute can be abstract or concrete; their nature affects how they need to be measured.

The *return* is what respondents return in response to the query. The return can be open, requiring respondents to use their own words when responding. Alternatively, the return can consist of a check or tick to one or more *response options* or *answer options* listed in the questionnaire. In some instances, the attribute may be included in the return rather than the query, for example, when the image of a destination is measured and a number of attributes are listed as part of the return: Is Sydney . . . (query) 1. Safe, 2. Expensive, 3. Fun (return).

Response options can be *anticipated*, meaning that they contain the kind of answer the survey designer anticipates. The number of anticipated answer options provided to respondents can range from one only to many. Anticipated answer options can be verbally labeled, in which case words are used to express the exact meaning of answer options (in Figure 1: "Very comfortable" and "Very uncomfortable" are the endpoint labels for the answer format). Alternatively, they can be labeled by using little pictures or icons or numbers (in Figure 1: –2 to +2).

Answer options that are frequently offered in surveys but do not represent the main, anticipated responses participants will return are *"Don't know" answer options*. The intention of such options is to prevent respondents from guessing when they do not know the answer. *Not applicable answer options* can be offered when questions may not be relevant to all respondents.

All these elements are discussed in detail in the following sections. Bad choices with respect to any of those elements can substantially reduce the quality of survey measures which, in turn, will decrease the validity of knowledge developed or market intelligence gained from a survey study. Selected examples will be discussed. Note, however, that most of us have made some of these mistakes at some point and every journal has published papers with measurement deficiencies. The purpose of presenting examples from published work is to improve our collective understanding of the exact nature of key pitfalls to be able to avoid them.

How to Define What Is Being Measured?

> The first half of the battle consists of putting the issue in a form that we can understand ourselves . . . After we are sure that the issue is fully defined . . . we can begin to translate it into simple words for public consumption. (Payne 1980, pp. 26–27)

The key to developing good survey questions is to precisely define what is being measured. Although there is broad agreement that whatever is being measured needs to be clearly defined, virtually no recommendations on how this should be done are provided in textbooks on (tourism) marketing research.[2]

Most measurement experts discuss the requirements for a good construct definition, but do not provide practical instructions as how to develop a definition that complies with the criteria stated. For example, Bagozzi (2011) argues that the theoretical meaning of a construct should be defined in terms of well-formedness, specificity, scope, ambiguity, vagueness, and transcendence versus immanence and include its antecedents and consequences. Bagozzi does not explain *how* to achieve that, especially if antecedents and consequences are unknown at the time of investigation. Kahane (1982) discusses different kinds of definitions and outlines criteria of good definitions: they should not be vague, ambiguous, obscure or circular; they should state the essential properties of the thing named by the term, make clear the context in which the term is to be used, and make it possible to eliminate the defined term and replace it with its definition. Kahane does not provide an instruction of how to ensure this.

To the author's knowledge, the first theory-based guide to construct definition has been proposed by Rossiter (2011), who specifies three elements of the construct that need to be spelled out:

1. the rater (the person being asked),
2. the object (the object under study), and
3. the attribute (what exactly about the object will be studied).

If the object and the attribute are clear, it is easy to follow Rossiter's instructions. For example, if a hotel wants to know if Mr. Smith found the hotel bed comfortable, the construct would be defined as follows:

MR. SMITH'S (rater)

EVALUATIONS OF HOW COMFORTABLE (attribute)

THE HOTEL BED (object) IS

In this case there is no ambiguity about any of the elements in the definition. *Mr. Smith* is Mr. Smith, and *comfortable* is a term that is clear to everyone, does not require further explanation, and its meaning is shared across respondents. Finally, the object (the hotel bed Mr. Smith slept on) is also unambiguous. Rossiter refers to such attributes and objects as *concrete* and lists as their key characteristics that they are clearly understood by respondents and unambiguous, and that respondents agree on their meaning.

Abstract objects or attributes, on the other hand, are "something that scientists put together from their own imaginations, something that does not exist in an isolated, observable dimension of behavior" (Nunnally 1978, p. 96). Abstract objects or attributes consists of a number of components, each of which has to be included in the definition and captured in the empirical measure, making the task substantially more complex. In cases where both the object and attribute are abstract, the following elements need to be specified (Rossiter 2011):

1. the rater,
2. the object,
3. the components of the object,
4. the attribute,
5. the components of the attribute.

For example, if the national tourism organization of Australia wants to assess

US TOURISTS' (rater)

KNOWLEDGE (abstract attribute)

ABOUT AUSTRALIAN CITY DESTINATIONS (abstract object)

it is necessary to specify the components of both the abstract attribute KNOWLEDGE and the abstract object AUSTRALIAN CITY DESTINATIONS, which is abstract because it contains more than one entity (referred to as *abstract collective* by Rossiter) and therefore by definition is not unambiguous. So, one possible definition of this construct is

US TOURISTS' (rater)

KNOWLEDGE (abstract attribute)

- about travel time from the USA
- about travel cost from the USA
- about temperature during summer holidays in the USA at

AUSTRALIAN CITY DESTINATIONS (abstract object)

- Sydney
- Melbourne
- Brisbane
- Canberra
- Perth

Such definitions make it clear to the researcher developing survey questions what precisely is being measured, but it also clarifies to readers of publications reporting study results what exactly has been measured and to which body of knowledge it contributes.

In the 78 recently published tourism survey studies, not one instance was identified where rater, object, and attribute were specified. For the constructs studied most frequently (attitude, behavioral intention, satisfaction, behavior, image, and loyalty), definitions were only provided in 37% of cases. When provided, definitions differed substantially. For example, four definitions of the construct *loyalty* read as follows:

- "the brand attitude of consumers in their intentions to repatronize or repurchase and willingness to pay a premium price for a preferred product or service" (Horng et al. 2012, p. 817)
- "a measure of the attachment that a customer has to a brand" (Aaker 1991, p. 39, quoted in Hsu, Oh, and Assaf 2012, p. 84)
- "the consumer's intention to visit or willingness to recommend the hotel or restaurant brand" (Nam, Ekinci, and Whyatt 2011, p. 1015)
- "a deeply held commitment to re-buy or re-patronize a preferred product/service consistently in the future, thereby causing repetitive same-brand or same brand set purchasing, despite situational influences and marketing efforts having the potential to cause switching behaviour" (Oliver 1999, p. 34, quoted in Prayag and Ryan 2012, p. 345)

These definitions do not have much in common. Although they use the same rater (the guest at a tourist destination or restaurant), they differ in the object under study (hotels, restaurants, destinations), and the components of the attribute *loyal* differ substantially: Horng et al. use the components (1) intention to repurchase and (2) willingness to pay a price premium; Hsu et al. use attachment; Nam, Ekinci, and Whyatt use (1) intention to visit and (2) willingness to recommend; and Prayag and Ryan use commitment to consistently repurchase. If these four studies arrive at different conclusions about the antecedents and consequences of loyalty, it may well be because different aspects of loyalty have been investigated, making cumulative development of knowledge about the construct difficult. Table 1 provides a similar analysis for the most frequently studied constructs in the 78 recent empirical studies.

Table 1. Frequently studied constructs in tourism and their definitions.

Construct	No. of Studies in Sample	% Definition Provided	Are the Definitions the Same?	% Unambiguous Definitions
Attitude	17	12%	No	0%
Behavioral intention	13	23%	No	0%
Satisfaction	12	17%	No	0%
Behavior	10	20%	No	0%
Image	8	50%	No	25%
Loyalty	4	100%	No	50%

The last column in Table 1 specifies the percentage of unambiguous definitions provided. For the purpose of this review, definitions were classified as unambiguous if it can be assumed that handing the definition to two survey researchers and asking them to develop measures based on this definition would lead to the same measures. While this assessment is subjective, it can still be concluded with some confidence that definitions are rarely unambiguous.

Below are examples of how some of the frequently studied constructs could be defined:

Destination image:

CHINESE TOURISTS' (rater)

BELIEFS (attributes, specific beliefs need to be elicited from consumers) ABOUT

WOLLONGONG (object)

Satisfaction:

FIRST TIME VISITORS' (rater)

OVERALL SATISFACTION (attribute)

WITH THE HOTEL (object)

Intention to return:

DOMESTIC TOURISTS’ (rater)

INTENTION TO RETURN (attribute)

TO WOLLONGONG (object)

To sum up: the key to developing high quality measures is to have a clear definition of what is being measured. This requires identifying whether what is being measured is concrete or abstract in nature and then spelling out clearly the rater whose perspective is studied, the object and the attributes of the object. If either the object or the attribute or both are abstract, it is also necessary to spell out which components the object or attribute consists of (Rossiter 2011).

How Many Questions to Ask?

In 1979 Churchill wrote an article on how to develop better measures for marketing constructs and proposed a psychometric framework for the development of scales. His procedure includes the following steps: (1) defining the construct; (2) developing survey questions; (3) collecting primary data using the original set of variables; (4) purifying the measure using coefficient alpha, item-to-total correlation, and confirmatory factor analysis; and (5) reporting item quality using convergent validity (correlation with other measures thought to measure the same thing) and nomological validity (significant correlations with theorized antecedents and consequent variables). Content validity in Churchill’s proposed framework is ensured by following the scale development procedure step by step.

Churchill’s article led to a measurement revolution in marketing and many other social sciences including empirical tourism research. In many instances the Churchill procedure has improved measures; but some unintended negative consequences have also resulted from the wide adoption of the procedure.

The Use of the Churchill Procedure as a Mindless Drill

As Churchill himself (1998, p. 30) noted: “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 . . . the rote approach falls far short of what we should be striving for.”

Survey researchers in tourism largely follow Churchill’s scale development paradigm, although they rarely explicitly say so: of the 279 constructs measured empirically in the reviewed articles, only 20% explicitly state the measurement paradigm under which they operate. However, given the scale quality criteria provided by the authors, it can be assumed that at least 60% of studies using multiple items to measure a construct use parts of the Churchill procedure to develop scales. Whether or not tourism researchers have a tendency to use it as a “rote process” or whether they use it as originally intended is hard to assess because the actual items are rarely provided. Survey items are provided for only 30% of constructs; a justification of the choice of scale development paradigm has never been provided.

The Uncritical Use of the Quality Criteria Proposed by Churchill

To quote Churchill again (1998, p. 35), the “concern is not with these calculations, but with their blind use as decision criteria. I have seen any number of instances where people calculate alpha and the item-to-total correlations and automatically use low item-to-total correlations to throw out items, thereby improving alpha. The danger in doing so is that while that procedure invariably improves alpha, it just as invariably lowers the quality of the measure of the construct.” The criteria Churchill proposed to evaluate the quality of scales have become so widely accepted that journal reviewers and editors appear to judge the measure merely by the size of coefficient alpha (although we all know that size does not really matter). What matters is whether the items are capturing the meaning of the construct. This can never be assessed by looking at coefficient alpha, it can only be assessed by looking at the survey questions people were asked.

The review of recent survey studies in tourism reveals that coefficient alpha values were provided for 56% of the constructs measured using multiple items. Factor analysis was used in 17% of constructs to reduce the number of items, and in 45% of cases to determine underlying dimensions of the constructs. For example, one study investigated, among others, job satisfaction of Scottish chefs. Factor analysis led to a factor that was labeled “promotion,” but the item “people get ahead as fast as they do in other places,” which is clearly conceptually related to promotion, was assigned to a different factor. Similarly, the item “I like the people I work with” was not assigned to the factor labeled “co-workers.” In this instance, factor analysis clearly did not group the items into meaningful components defining the job satisfaction construct for chefs. Similarly, a study aimed at understanding how the needs of disabled air passengers can be met used factor analysis to conceptualize the dimensions, leading to a number of illogical assignments of items to factors, such as the items “barrier-free environment” and “reservation phone number” being assigned to the factor “compensation and improvement schemes.”

Cases such as those illustrated above are not uncommon. Components of constructs are rarely developed conceptually in tourism research, often leaving it up to a statistical method (typically factor analysis) to perform this task which by nature is conceptual, not statistical. Churchill (1979) points out that factor analysis tends to produce more factors than conceptually postulated and thus should not be used to circumvent conceptual construct definition (but does have its place in confirming conceptually postulated dimensions). Rossiter (2011) argues that there is no role for factor analysis at all in item and scale development.

It is also surprising how often survey researchers who have been in control of data collection argue that they used factor analysis to reduce the number of variables they have initially chosen to use in the survey. A group of authors, for example, developed survey questions with the aim to learn more about prospective volunteer tourists, including pretrip expectations. They developed and asked people 18 questions and factor analyzed them to "make the number of variables more manageable." In so doing, the researchers lost, in this instance, 37% of the information they collected (the variance explained was 63%), but more critically, they can no longer draw conclusions about the very pieces of information they originally decided were critical.

An implicit C-OAR-SE perspective (Rossiter 2002, 2011) can only be found in one single survey study in tourism: Choi et al. (2012) state that the "questionnaire was then reviewed by three tourism researchers to check content validity" (p. 29).

Overall, it can be concluded that the understanding of alternative measurement paradigms is relatively low. The incidence of using items or scales taken or derived from previously published studies is high (63% of constructs in the reviewed articles), risking reduced validity of the measure.

The Assumption That Measuring Every Construct Requires Multiple Survey Questions

The issue of single- versus multi-item measures is hotly debated. Opposing views are due to differences in core beliefs about measurement in the social sciences held by two schools of thought, Domain Sampling Theory which underlies the Churchill procedure and C-OAR-SE theory. Both agree that if the object being measured (e.g., a hotel) has multiple components, or if a single object is being assessed using attributes that have multiple components (e.g., knowledge about online booking of plane tickets), multiple survey questions are required. For example, to investigate

> HOTEL GUESTS' (rater)
>
> SATISFACTION (attributes)
>
> WITH THE HOTELS PERFORMANCE IN DIFFERENT ASPECTS (components of the object)

multiple questions are required because each aspect of the hotel (e.g., room size, room cleanliness, comfort of the bed, size of bathroom) has to be assessed separately. The two schools of thought disagree about how many questions are needed to measure one component. For example, when measuring

> HOTEL GUESTS' (rater) ASSESSMENT
>
> ASSESSMENT OF THE COMFORT (attribute)
>
> OF THE HOTEL BED (object)

C-OAR-SE theory suggests that it is sufficient to ask only one good question (e.g., "Was the bed comfortable?"). Domain Sampling Theory suggests that multiple questions should be asked (e.g., "Was the bed comfortable?" "Did you sleep well on the bed?" "Did you wake up refreshed in the morning?").

What is the underlying reason for this difference? Followers of Domain Sampling Theory argue that the observed score (the answer given by the respondent on the survey) is a result of the true score (the real answer that is in the respondent's head) being distorted by systematic and random error (observed score = true score + systematic error + random error). Optimally, the observed score is equal to the true score. But according to Domain Sampling Theory, it is not possible to obtain an observed score that is equal to the true score because systematic error and random error distort the observed score away from the true score. Systematic error is caused by respondents' tendencies to use extreme or middle answer options. Random error is caused by factors such as respondents' mood, but also imprecise measures or variation in how the questionnaire is administered. Under Domain Sampling Theory, error is minimized by asking respondents more than one question about each component of interest assuming that any error will average out over multiple questions.

Churchill (1979), a proponent of Domain Sampling Theory, provides three arguments for asking respondents more than one question per component: (1) single items are too specific, and such specificity can be averaged out by using multiple items, (2) individual values for a construct have a finer discrimination if they are derived from multiple questions. For example, a single item with 10 answer options allows for differentiation between 10 groups of people based on their response. If, however, five items with 10 answer options are used, the minimum score is 0 and the maximum score over all items is 50, allowing discrimination between 50 groups of people. Finally, Churchill argues, (3) multi-item measures have higher reliabilities because random error averages out.

The C-OAR-SE position on Churchill's arguments would be as follows: (1) if an item is too specific, it should not be used; (2) finer discrimination among people is only useful if it is valid and validity does not increase, but rather decreases if additional bad items are added; and (3) there is no random error, errors must be prevented when the measures are developed, and they cannot be corrected retrospectively. Under C-OAR-SE, the following alternative model is proposed: observed score = true score + measure induced distortion + rater error.

Measure-induced distortion is caused by bad measures (a badly formulated query or unsuitable return options). Measure-induced distortion is not random error and can be prevented if survey questions are designed carefully. Rater error occurs when respondents make mistakes, due to fatigue or lack of motivation, etc. This can never be prevented; it is truly out of the control of the researcher and therefore represents a random form of error that leads to the observed score not being identical to the true score. Importantly, however, asking more than one question with slightly different meanings and averaging across observed scores cannot remove rater error.

According to C-OAR-SE, asking multiple questions about the same component means that bad questions will be included and will thus decrease the validity of the overall observed score. C-OAR-SE theory can therefore be seen as the stricter measurement theory in terms of researcher responsibility: it forces the survey researcher to think about every possible error and prevent it before it occurs.

Converse and Presser (1986) raise two practical concerns about multi-item scales: (1) if questions are not well worded or suboptimal answer options are provided, using multiple such items in a scale will increase, rather than cancel out, measurement error; and (2) the validity of scales changes over time, so it is unwise to assume it is still valid a few decades later. Converse and Presser acknowledge that single-item measures play a key role in surveys, but believe that this is primarily the case because they keep surveys short.

Recently, Rossiter and Bergkvist (2009) have investigated under which circumstances it is sufficient to ask one single question. Rossiter (2011) identifies three different kinds of objects (concrete, abstract collective, abstract formed) and four kinds of attributes (concrete perceptual, concrete psychological, abstract achieved, and abstract dispositional). Of particular interest in terms of single-item measures are *concrete objects*, meaning that the respondents understand clearly and unambiguously what they are being asked about (e.g., their hotel bed) AND a *concrete perceptual attribute*, meaning that the respondents understand clearly and unambiguously what attribute is being assessed (e.g., comfortable). It is in such cases—which Bergkvist and Rossiter (2007) refer to as *doubly concrete*—that single-item measures are sufficient and, as they argue, superior to multi-item measures because adding any additional items would blur what exactly is being asked. Using the hotel bed example: Asking three questions ("Was the bed comfortable?" "Did you sleep well on the bed?" "Did you wake up refreshed in the morning?") reduces the quality of the measure because the additional questions measure slightly different things, but are weighted equally as the main meaning. That this effect indeed occurs has been shown empirically by Bergkvist and Rossiter (2009).

A summary of when single and multiple items are required (based on Rossiter 2011) is provided in Table 2.

As can be seen, in most cases (all light-gray cells in Table 2) multiple items are required for a range of different reasons. Only in one situation (dark-gray cell in Table 2) is a single item sufficient to capture the construct, that is, when both the object and the attribute under study are concrete.

Of the 78 recent articles reviewed, only 7% use single-item measures. For example, Bojanic and Warnick (2012) measured both satisfaction with and likelihood to return to an event with a single item each. Many instances can be found where doubly concrete constructs are measured using multiple items. For example, in one study, travel intention was measured with four items, only one of which actually reflected travel intention, the other three read as follows: "I'll say positive things about cruising to other people"; I'll recommend cruising to others"; and "I'll encourage friends and relatives to go on a cruise." Given that three of four measures do not actually measure intention to travel at all, it may well be that the proposed model of explaining intention to travel actually explains intention to recommend instead.

It can be concluded that the scale development paradigm proposed by Churchill dominates survey research in tourism, although the actual measures are rarely made publicly available, thus making it impossible to assess whether following the Churchill procedure led to good measures. Five key precautionary steps can be taken to avoid pitfalls relating to scale development:

1. Identification of the nature of the construct to determine whether one or multiple items are required.
2. If multiple items are required, a scale development paradigm has to be chosen. One option is to follow the Churchill paradigm (outlined in detail in his 1979 article, and discussed further in Churchill 1998) and to avoid the unintended consequences outlined above. C-OAR-SE is, to the best of the author's knowledge, the only comprehensive measurement theory for the social sciences which represents a true alternative to the Churchill procedure. C-OAR-SE has been developed only recently (Rossiter 2002) and is described in detail in Rossiter (2011).
   Both procedures have in common that the construct under study has to be clearly defined. Rossiter requires rater, object and attribute to be defined. Churchill emphasizes that the researchers must clearly conceptualize constructs and specify domains, but does not provide guidance on what precisely a definition should include. In both cases it is acknowledged that it may be necessary to conduct qualitative research in advance of developing the items in order, for example, to determine whether the meaning is unambiguous, to identify

Table 2. When to use single or multiple items?

		Object Type		
		Single concrete object • Examples: • country, city, town • airline • accommodation option	*Multiple concrete objects* Examples: • ecotourism destinations • backpacker hostels	*Complex objects with multiple components* Examples: • important product amenities • important service elements
Attribute type	*Single concrete attribute* Examples: • demographic characteristics • recall, recognition • perceptual attribute • evaluative attribute • specific emotion • overall evaluation or satisfaction • intention • behavior	*Single-item measure* Examples: • Will you visit this city again as a tourist? • Is Qantas Australian? • Will you recommend this hotel to your friends?	*Multi-item measure* (category-representative objects) Examples: • Are ecotourism destinations clean? • Are backpacker hostels inexpensive?	*Multi-item measure* (main component objects) Examples: • Overall, how satisfied were you with the ○ Hotel staff ○ Your room ○ The breakfast
	Complex achieved attribute Examples: • multiattribute image • accumulated knowledge	*Multi-item measure* (main component attributes) Examples: • Is Australia ○ Clean ○ Safe ○ Expensive ○ Friendly	NOT APPLICABLE because abstract achieved attributes about abstract collective objects are not of interest	*Multi-item measure* (main component objects) Examples: • During your stay at the hotel, were you satisfied with ○ Staff ○ Friendliness ○ Effectiveness ○ Room ○ Size ○ Cleanliness
	Complex dispositional attribute Examples: • personality trait • motive or goal	*Multi-item measure* (main component attributes) Examples: • Is Qantas ○ Innovative ○ Honest ○ Reliable	NOT APPLICABLE because complex dispositional attributes about abstract collective objects are not of interest	NOT APPLICABLE because complex dispositional attributes about abstract formed objects do not need to be measured

multiple meanings of objects or attributes from the consumer or expert perspective. It is in the next stage where the approaches differ: according to C-OAR-SE theory, only one question is required per component. For example, if values are being measured, only one survey question is required per value (e.g., sense of belonging). Churchill, in line with psychometric theory more broadly, argues that multiple survey questions should be used for each component because each survey question is assumed to have a "slightly different shade of meaning" (Churchill 1979, p. 68) and the average across all those different shades of meaning are assumed to best reflect the component of the construct. As a consequence, Churchill requires the original set of items to be put to field and the resulting data to be used for purifying the measure using coefficient alpha, item-to-total correlation, and confirmatory factor analysis. In C-OAR-SE this step is not required because C-OAR-SE postulates that validity of an item cannot be empirically tested, it has to be ensured by expert assessment. Rossiter is highly critical of empirical purification of items, stating that "statistical finagling after the fact cannot fix problems of low item-content validity or low answer-content validity" (Rossiter 2011, p. 99). So, while—under the Churchill paradigm—purification of items after the initial round of data collection is undertaken with the assistance of measures such as coefficient alpha, item-to-total correlation and confirmatory factor analysis and, for the final items, convergent validity and nomological validity have to be reported, C-OAR-SE requires content validity to be (1) rationally argued by demonstrating both item-content validity (how clearly and unambiguously the question is asked) and answer scale validity (whether the response options offered are the main ones for the rater, no more and no less) and (2) confirmed as content valid by experts.

3. Justification of why a certain measurement theory was used.
4. Genuine adherence to the recommendations (and the spirit) of that scale development paradigm.
5. Making available the final items to readers. Ultimately, it does not matter how items are developed; what matters is that the items capture the construct. Whether or not this is the case can only be judged by comparing the items with the construct definition. Such an assessment can be made retrospectively and readers of study results should be able to convince themselves of the validity of the empirical measure used.

How to Ask a Question (the Query)?

Developmental Qualitative Research

Development of questions may or may not require input from people other than the researcher. For example, if a researcher wishes to measure

TOURISTS' (rater)

INTENTION TO RETURN (attribute)

TO WOLLONGONG (object),

they can easily formulate a valid survey question that will capture the construct because neither the attribute (intention to return) nor the object (Wollongong) requires further specification by experts or tourists. If, however, the construct under study is

RESIDENTS' (rater)

BELIEFS (attribute)

ABOUT THE EFFECTS OF TOURISM ON THE LOCAL AREA (object),

it is unlikely that the researcher, through introspection or reading of the literature alone, will be able to capture every important belief local residents may hold. Residents must be asked about their beliefs by means of developmental qualitative research. The term *developmental qualitative research* is used to describe qualitative research conducted for the sole purpose of informing the development of questionnaires. This term is derived from Converse and Presser's (1986) term *developmental pretest* in the context of questionnaire pretesting. Developmental qualitative research can use a range of valid sources of insight. Typical sources include interviews and focus groups, but valuable information could also come, for example, from reading the Letters to the Editor section of the local newspaper or going to the annual general meeting of the local tourism organization.

In cases where the attribute under study is not perceptual, consumers cannot assist in the development of survey questions. For example, if the construct is

RESIDENTS' (rater)

KNOWLEDGE (attribute)

ABOUT THE ECONOMIC IMPACT OF TOURISM TO THE LOCAL ECONOMY (object),

experts on economic impacts of tourism to communities need to be involved in item generation.

The Wording of the Query

Payne (1980) reports on a large number of split-ballot experiments investigating effects of small wording differences. He found, for example, that using the word *should* versus the word *might* in the following question: "Do you think anything should/might be done to make it easier for people to pay doctor or hospital bills?" leads to a 19% difference in agreement: 82% say it "should," but only 63% say it "might." In a similar experiment, testing the effects of different wording of answer options, the difference was 28%. Payne compares these differences with sampling errors: the biggest documented historical disasters in opinion polling due to sampling errors occurred in the 1936 and 1948 election polls in the United States. The errors made were *only* 19% and 4%, respectively. It is worth keeping this in mind when developing a survey study: question wording matters a lot and can affect results more than other survey design factors, including sampling and statistical analysis.

The key challenge in formulating survey questions is to ensure that respondents interpret them the same way. Cantril (1940), Payne (1980), and Converse and Presser (1986) derive from their experiments and experience a number of practical recommendations on how to reduce variability of interpretation:

1. *Use plain, everyday language*: It is critical to formulate queries in plain, simple language that can be understood by every respondent. Optimally both

queries (and response options) are written in the language people would use every day—plain *spoken* language.

2. *Use short queries*: The longer the query, the harder it is to understand for a respondent. Payne (1980) uses the term *marathon questions* and warns against their use, suggesting no more than 20 words as a rule of thumb. Also, the number of difficult words as a proportion of all words should be as low as possible.
3. *Avoid acronyms and technical terms*: Acronyms and technical terms complicate queries unnecessarily as they either assume that people understand them or expect people to put additional cognitive effort into interpreting them.
4. *Make queries specific*: Optimally, the query is specific enough that no respondent requires clarification on what is being asked.
5. *Avoid double-barreled queries*: Queries must not be double barreled, meaning that they include two objects or attributes, thus confusing respondents about what is being asked (e.g., "How satisfied were you with food and wine at the restaurant?"). Responses to such queries are meaningless because it is not clear to which part of the query their response relates.
6. *Avoid double negatives*: Double negatives (e.g., "Do you agree that people should not waste water?" No Yes) should be avoided because they confuse respondents.
7. *Avoid "Strongly agree" to "Strongly disagree" answer scales*: Converse and Presser (1986) and Schuman and Presser (1981) advise against asking respondents to indicate their level of agreement to statements. The agreement–disagreement format is not optimal because it is prone to capturing response bias. Specifically, as Schuman and Presser (1981) demonstrate: acquiescence bias or yes-saying bias, a finding replicated many times subsequently (for a review, see Krosnick 1999). Such bias cannot be retrospectively corrected and, consequently, biases study results. The issue of response bias will be discussed in detail later.

   Multicategory agreement–disagreement statements are heavily used in tourism survey research: of 279 constructs in the reviewed articles, agreement–disagreement answer options were used in 41% of cases.
8. *Pretest, pretest, pretest*: Pretesting is critical to developing good survey questions. It is particularly valuable to request feedback from other survey experts, to ask pretest respondents to talk out loud when they complete the survey to see if they misunderstand or struggle with any aspect of the survey (Rossiter 2011) and to ask pretest respondents to explain what they mean by the answer they have given.

Order Effects

The order of presentation of questions within the survey as well as answer options for each individual query affects responses. Such effects are referred to as order effects. In an extensive review of experimental work on response order effects, Krosnick (1999) concludes that respondents—when asked categorical questions—tend to choose the first of the response options (primacy effect) when they are presented with options visually (written questionnaire or card with options as part of an interview or online), but they tend to choose the last option (recency effect) when response options are presented to them orally, for example, in phone interviews. Furthermore, Krosnick identifies the following factors to increase response order effects: lower levels of cognitive skills of respondents, fatigued respondents, and difficult and longer questions.

How to Allow Respondents to Answer (the Return)?

Open or Closed Returns

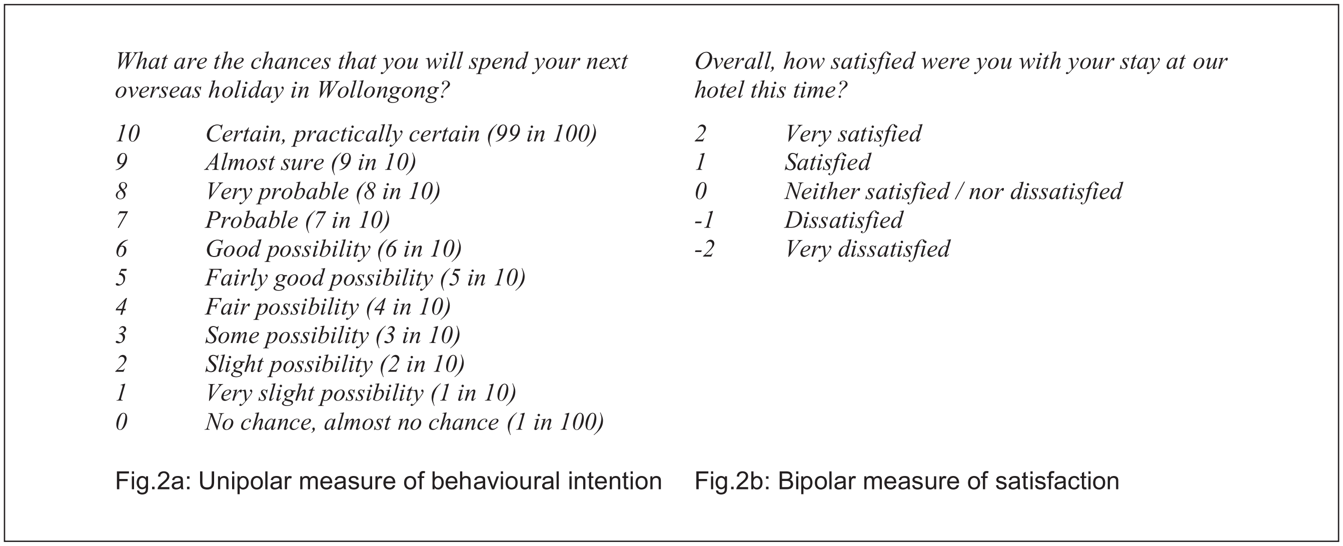
Open questions play an important role in survey research (Payne 1980; Krosnick 1999), especially for the assessment of salience and the collection of a wide range of responses, such as suggestions for improvement, because respondents are not influenced. The main advantage of closed questions is that they are more specific because all participants see the same responses. If the set of responses has been carefully developed (which often requires developmental qualitative research), closed questions should not lead to the omission of important responses.

Unipolar and Bipolar Response Options

Unipolar answer options go in one direction. Intention to visit a tourism destination, for example, demands offering unipolar answer options to respondents because the chances of visiting a destination lie between zero and 100%, they can never be negative (see Figure 2a). Bipolar answer options allow respondents two directions, with one positive and one negative endpoint. For example, tourists can be highly satisfied (extremely positive evaluation) or highly dissatisfied (extremely negative evaluation) with their stay at the hotel (see Figure 2b).

Compared to other aspects of questionnaire design, the issue of unipolar versus bipolar response options has received little attention. The overwhelming popularity of the Likert-type scale in the social sciences may be partially to blame. Using Likert-type scales (Likert 1932) automatically implies bipolar answer options, with the endpoints representing "strongly agree" and "strongly disagree."

Rossiter (2011) emphasizes the importance of formulating response options correctly in terms of their direction. It is important to note that whether or not to use unipolar or



What are the chances that you will spend your next overseas holiday in Wollongong?

10	*Certain, practically certain (99 in 100)*
9	*Almost sure (9 in 10)*
8	*Very probable (8 in 10)*
7	*Probable (7 in 10)*
6	*Good possibility (6 in 10)*
5	*Fairly good possibility (5 in 10)*
4	*Fair possibility (4 in 10)*
3	*Some possibility (3 in 10)*
2	*Slight possibility (2 in 10)*
1	*Very slight possibility (1 in 10)*
0	*No chance, almost no chance (1 in 100)*

Fig.2a: Unipolar measure of behavioural intention

Overall, how satisfied were you with your stay at our hotel this time?

2	*Very satisfied*
1	*Satisfied*
0	*Neither satisfied / nor dissatisfied*
-1	*Dissatisfied*
-2	*Very dissatisfied*

Fig.2b: Bipolar measure of satisfaction

Figure 2. Unipolar and bipolar response options.

bipolar response options is not a question of researcher preference, it is determined by the nature of the constructs. The researcher has to determine whether the construct under study conceptually allows positive and negative responses (as is the case for beliefs, attitudes, satisfaction) or whether responses are logically limited by the zero point and a positive extreme (as is the case for number of nights spent at a destination, intention to return to a destination, intention to share the satisfaction about a trip with friends). Importantly, numeric coding of the answer options must reflect the directional nature of response options. For unipolar answer options, numeric coding starts with "0" and increases in numbers; for bipolar constructs, the middle point (if a middle point is offered) is coded as 0, with negative options coded as –1, –2, etc. and positive options coded as 1, 2, etc. Correct coding ensures that results from data analysis are not misinterpreted.

Of the constructs studied in recent tourism studies using nonbinary rating scales, 93% offered respondents bipolar answer formats (although only 7% state explicitly that they have done so). Of all answer scales that were identifiable as bipolar to the reader, 89% incorrectly numerically label the answer options as unipolar.

One of the survey studies reviewed for this article used "bipolar adjective scales, where 1 = negative and 7 = positive" (correctly this should be scored as –3 to +3 to account for the bipolar nature of response options), concluding that "the country's affective country scores are relatively weak, as low as 3.4 for 'safe.'" In this study, the value 3 was the middle point, so an average value of 3.4 indicates that the average respondent perceives the country neither as safe nor as unsafe. In another example, a seven-point unipolar answer format ranging from "not at all" to "a very great extent" was used. Respondents were asked six questions measuring the extent to which they felt cheated as tourists. The first five asked whether they felt staff were cheating, duping, acting dishonestly, being deceitful, and swindling. The last item asked if respondents felt they had been *under*charged. This item is subsequently reverse coded. So if a respondent stated that they have "not at all" been undercharged this is, incorrectly, interpreted as the highest level of overcharging.

To ensure that the correct direction of the answer scale is chosen, the conceptual nature of the construct needs to be understood. Common unipolar constructs measured in tourism include destination image, perceptions, intention to visit/return, intention to recommend, loyalty, and travel motivations; these should be numerically coded from 0 to the maximum number on the scale. Common bipolar constructs measured in tourism include satisfaction, overall evaluations, and overall attitude; these should be coded in the data set from the maximum negative to the maximum positive number on the scale, with or without a zero in the middle, depending whether a neutral middle option is included in the answer options or not.

The Number of Response Options

Response options can be *anticipated answer options, not applicable answer options*, and *no response options* (see Figure 1). *Anticipated answer options* represent the kinds of responses the researchers expect, *not applicable answer options* can logically not be answered by the respondent because the question is not applicable to him or her (e.g., if you are asked if you liked your stay in Wollongong and you have never been to Wollongong before), and *no response* or *"Don't know" options* allow respondents to not answer the question although there is no logical reason that they could not answer it (E.g., you are asked what your image of Sydney is. Even if you never visited Sydney you may have a perception of it. You may, however, choose not to answer all questions because you do not feel competent to assess each attribute of Sydney).

Is Wollongong ...... ?

Expensive	□
Fun	□
Clean	□

Fig. 3a: Pick any binary

Is Wollongong ...... ?

Pretty expensive	□*Yes* □*No*
Lots of fun	□*Yes* □*No*
Squeaky clean	□*Yes* □*No*

Fig. 3b: Forced choice binary

Is Wollongong ...... ?

Expensive	□*Yes* □*No*
Fun	□*Yes* □*No*
Clean	□*Yes* □*No*

Fig. 3c: Doubly level free with individual satisfaction thresholds (DLF IIST) or Binary Internal Threshold (BIT)

Figure 3. Binary answer options.

In this section, the focus is on *anticipated answer options*. With respect to *anticipated options*, three choices are available to survey designers: offering only one answer option, two answer options, or more than two answer options. Examples of a destination image measure with one and two anticipated answer options are provided in Figure 3. They are all referred to as binary because the numerical coding of responses in all cases is 1 if the respondent agrees and 0 otherwise.

In the case of the *pick any* or *pick any out of n* answer format (Figure 3a), respondents tick the box next to the attribute if they believe that it applies (in this case, to Wollongong). Alternatively, they do not tick a box, in which case it is not possible for the researcher to determine whether they believe that the attribute does not apply to Wollongong, or whether the respondents wanted to get through the questionnaire quickly and simply skipped the question. Offering only one answer option is common in commercial brand image studies.

The same question can be asked offering two answer options, effectively forcing respondents to answer. This answer format is referred to as *forced choice binary* (Figure 3b). This answer format does not restrict the way the query is formulated or the way the attributes, in this case characteristics of a tourist destination, are formulated.

A more specific kind of the *forced choice binary* format is the *DLF IIST* format. DLF IIST stands for *Doubly Level-Free with Individually Inferred Satisfaction Threshold* (it is also referred to as the *BIT* measure, which stands for *Binary Internal Threshold*) and is illustrated in Figure 3c. This format is characterized by additional restrictions being placed on the basic forced-choice binary format: neither the description of the answer options ("Yes" and "No") nor the attribute part of the question ("expensive," "fun," "safe") may contain an indication of magnitude (e.g., "very," "extremely," "rather," . . .). The theory is that, as a consequence, respondents answer the question with respect to their personal threshold for any given attribute. For a detailed discussion of the DLF IIST measure, see Rossiter, Dolnicar, and Grün (2010) and Rossiter (2011), for empirical evidence of its performance see Dolnicar, Rossiter, and Grün (2012) and Dolnicar and Grün (2013).

Alternatively, respondents can be offered multiple anticipated answer options, ranging from three to endlessly many. As soon as three anticipated options are provided (and assuming those do not include a neutral midpoint), respondents are simultaneously providing a direction and magnitude response. This can be positive when it is essential to understand magnitude, for example, to assess the effectiveness of an advertising campaign. It does come, however, at a price: the magnitude response can be biased by people's tendencies to respond in certain ways (known as response styles; Paulhus 1991), making it impossible to separate the response from the response style.

The discussion about which number of answer options is best has led to a large number of investigations. One of the earliest studies was conducted by Peabody (1962). Peabody finds no relationship between the number of answer options and the reliability of a scale and recommends the use of two answer options that lead to equal reliability but are simpler and more convenient to use and administer. Similarly, Komorita (1963) studied whether responses to Likert scaled items could be reproduced by weighting direction information (binary agree, disagree) with magnitude information, concluding that the magnitude information "had practically no effect on total scores" (p. 332). Jacoby and Matell (1971) extended this line of research to include four assessment criteria: test–retest reliability (stability over two measurements at different points in time), internal consistency reliability (measured using Cronbach's alpha), concurrent validity (predictive validity to another measure asked in the same survey, using a graphical scale), and predictive validity (predictive validity to another measure asked in the second survey, using a graphical scale). They compare 18 different versions of the Likert-type scale ranging from 2 to 19 answer options. Results confirm Peabody's and Komorita's findings: no statistically significant difference was determined in any of the criteria used to compare the performance of the answer formats.

Similar conclusions have been drawn in later studies which demonstrate that reliability of responses increased over time if separate questions were asked to capture direction and extremity (Krosnick and Berent 1993), and that the confounding of direction and magnitude in Likert-type scales leads to a tendency to use middle answer options (Albaum 1997; Albaum et al. 2007; Yu, Albaum, and Swenson 2003). This tendency can be avoided by asking respondents to answer two questions, one about direction only (including a "yes" and "no" option or similar and, if appropriate a "Don't know" option) and the second one aimed at capturing magnitude only (including, e.g., a "very" and a "somewhat" option).

A large number of empirical studies did not use the underlying concept of direction versus magnitude to drive the investigation. Rather they used a range of different criteria for the assessment of the number of answer options. Key studies are reported in the following sections.

Reliability and Validity

Empirical evidence on whether questions with fewer or more response options are more reliable and valid are mixed; a number of studies conclude that the number of response options does not affect reliability (Bendig 1954; Peabody 1962; Komorita 1963; Komorita and Graham 1965; Matell and Jacoby 1971; Jacoby and Matell 1971; Remington et al. 1979; Preston and Colman 2000), others come to the opposite conclusion (Symonds 1924; Nunnally 1967; Jones 1968; Oaster 1989; Finn 1972). The same is the case for validity, with Matell and Jacoby (1971), Jacoby and Matell (1971), and Preston and Colman (2000) finding no difference and Loken et al. (1987) and Hancock and Klockars (1991) concluding that validity is increased if more answer options are provided.

The above studies are problematic because they vary in their definitions of reliability and validity and, as noted by Chang (1994), they do not decompose systematic method variance and trait variance, giving scales with more answer options an advantage caused purely by computational reasons, specifically the restriction of range effect (see Nunnally 1970; Martin 1973, 1978) which affects correlation-based measures such as Cronbach's alpha and test–retest measures. Using structural equation models, Chang concludes that criterion-related validity is independent of the number of answer options and reliability can be improved by reducing the number of answer options.

Stability of Responses over Repeated Measurements

When a good survey question is asked, the observed score is close to the true score and it can be expected that respondents will be able to reproduce the observed score response if asked the same question again (in the absence of any environmental changes). If responses are found not to be stable, this can be due to a bad query or bad response options. Therefore, stability is a useful criterion for the comparative assessment of answer options.

Only a small number of studies have investigated stability; Dall'Olmo Riley et al. (1997) and Rungie et al. (2005) examine the problem of instability in brand image measures. Dolnicar and Rossiter (2008) show that this lack of stability can be explained partly by measurement factors. Dolnicar, Grün, and Leisch (2011) find no difference in stability in dependence of the number of answer options offered. These findings are supported by the study conducted by Dolnicar and Grün (2007d) using only repeated agreement as a criterion. If, however, a stricter criterion of stability is used (the requirement to actually choose the exact same response option), the forced-choice binary answer format outperforms all other response options (five-point and six-point formats, both fully verbalized and end-point anchored only) significantly. Additional empirical evidence for the stability of forced binary answer options has been recently provided by Dolnicar, Rossiter, and Grün (2012) and Dolnicar and Grün (2013). The key conclusion from these studies is that—in the context of brand image measurement—the forced binary answer format produces the most stable results without loss of concurrent validity, and that the answer format most frequently used in commercial market research, the pick any answer format, is highly prone to evasion, thus producing substantially fewer "yes" responses than any other answer format.

Interpretation of Findings

The purpose of survey data is to be interpreted so as to advance knowledge (a typical academic aim) or gain market insight (a typical aim of commercial survey studies). Consequently, it is important to understand whether the choice of the number of anticipated answer options affects the interpretation of findings. A number of studies have investigated this question; Green and Rao (1970) constructed artificial data and compared the recovery of data structure in dependence of the number of answer options, concluding that at least six points and at least eight attributes should be used. A different design involved collapsing empirical multicategory data to dichotomous or trichotomous data, computing factor analyses on both data sets and comparing the results (Martin, Fruchter, and Mathis 1974; Percy 1976). These studies lead to the conclusion that the results do not differ significantly using an objective measure of compliance between factor solutions as the criterion.

In those early studies, data transformation was controlled by researchers, not respondents. Later studies avoided this problem by collecting data using different answer formats directly from respondents: Loken et al. (1987) compare responses given on an 11- and a 4-point scale and conclude that there was no difference in discrimination power between sociodemographic groups and capturing of relationships between variables. Similarly, Preston and Colman (2000) empirically compare responses given using 10 different scales, including dichotomous and nearly metric (101 scale points) formats. Using correlation matrices, underlying factor structure, and coefficient alpha values as criteria, they find no differences between formats. Dolnicar, Grün, and Leisch (2011) collected empirical data using different answer formats, and find that results from a positioning analysis for both answer formats were practically the same.

It can be concluded from these studies that there is no reason to believe that any particular answer format is superior with respect to findings derived from the entire study.

User-Friendliness

In a study on respondents' answer option preferences, Jones (1968) finds that respondents prefer multiple categories. Similarly, Preston and Colman (2000) find that respondents are better able to express themselves when more categories are offered, although they perceive surveys with fewer response options as quicker. Respondents in a study by Dolnicar and Grün (2007a) also perceive the binary format as being the quickest (in the context of attitude and intention measurement) but show no preference for either a binary, 7-point, or metric format; neither do respondents perceive any of these formats as simpler or as better in terms of their ability to express their feelings.

Dolnicar and Grün (2007b) compared the user-friendliness of five different answer formats (forced-choice binary, 3-point scale, 7-point scale, visual analogue scale, and percentage allocation) in the context of responding to questions about brand image and behavioral intentions. They conclude that the 3-point, 7-point, and forced binary formats are preferred over the two metric formats. Respondents indicate that the forced binary answer format was the simplest and the quickest, but there was no significant difference in how pleasant they perceived the answer formats to be, and how well they allowed them to express their feelings.

Dolnicar, Grün, and Leisch (2011) report that 70% of respondents stated that the binary survey was easy, but only 45% indicated this for the multicategory version, supporting earlier findings by Dolnicar (2003) in which respondents perceived binary answer options as easier to answer than ordinal scales.

From these studies, it is concluded that results on user-friendliness are mixed, and no clear "winner" emerges. It is important to note, however, that people certainly do perceive some formats as easier, more pleasant, and quicker. User-friendliness should, therefore, be taken into consideration when developing a survey.

Proneness to Bias

A key concern—one shared among most measurement researchers in the social sciences—is data contamination by response styles. Two positions can be taken on response styles. The first position is that response styles can be removed from the data before data analysis. A wide range of approaches of how to remove bias of such kind has been suggested in the past (Cunningham, Cunningham, and Green 1977; Greenleaf 1992a, 1992b; Heide and Gronhaug 1992; Watson 1992; Van de Vijver and Poortinga 2002; Welkenhuysen-Gybels, Billiet, and Cambré 2003; Dolnicar and Grün 2007c).

The second position is that it is impossible to remove bias from data retrospectively and that it is therefore critical to select answer options which will reduce the likelihood of capturing response styles in the first place. Possibly the first proponent of this view was Cronbach (1950), who stated: "Since response sets are a nuisance, test designers should avoid forms of items which response sets infest" (p. 21) and recommended forced-choice and paired comparison formats to address this problem. The same argument is made by Lee et al. (2006) in favor of best worst scaling, and Rossiter, Dolnicar, and Grün (2010) in favor of the DLF IIST (or BIT) forced binary measure.

Kampen and Swyngedouw (2000) classify ordinal scales into different types and raise serious concerns about Type 5, defined as unstandardized discrete variables with ordered categories such as the agreement with statements or levels of satisfaction: "in many instances the experimenter can only hope that in general respondents or experimentators attach the same meaning to the [response] categories of an ordinal variable" (p. 99). In addition, the fact that scale points are not equidistant causes problems in both data analysis and interpretation.

In multicategory answer formats, a number of response styles can occur, including—most prominently—extreme response bias, where respondents tend to use the extreme positive and negative ends of the scale, tendency to the middle, where respondents tend to choose the middle option or an option close to the middle and acquiescence bias where people tend to respond in the affirmative (yes rather than no, agree rather than disagree). While extreme response bias and tendency to the middle can logically only occur in response formats with more than two options, acquiescence bias has been shown to also affect response options including only two options, such as "agree" versus "disagree". Specifically, a review of experimental evidence on acquiescence by Krosnick (1999) reveals that "yes" versus "no" questions are less prone to this bias than "true" versus "false" and "agree" versus "disagree" questions.

A different kind of bias occurs with pick-any measures, where respondents are only given one anticipated response option which they tick if they agree but do not tick if they disagree (for an example, see Figure 3a). Such response formats cause substantial levels of evasion (Rossiter 2011; Dolnicar, Rossiter, and Grün 2012; Dolnicar and Grün 2013), probably because respondents can complete the survey without ticking much, and without investing much cognitive effort. Krosnick (1999) would refer to this as *strong satisficing*.

The forced-choice binary format is able to avoid capturing most response styles, and thus represents an attractive option, if suitable for the measurement of the construct under study.

Speed

There are at least three good reasons to keep a survey as short as possible. First of all, short questionnaires reduce fatigue effects among respondents, which lead to a reduction of data quality (Johnson, Lehmann, and Horne 1990; Drolet and Morrison 2001; Vriens, Wedel, and Sandor 2001). Second, people have become increasingly reluctant to participate in survey studies (Hardie and Kosomitis 2005; Bednell and Shaw 2003), leading to difficulties recruiting representative samples. The length of the survey is a key reason for people to decline participating in market research (Hardie and Kosomitis 2005). Shorter surveys can increase participation rates and improve sample representativity. Finally, and most pragmatically, short surveys reduce fieldwork cost. In online surveys, for example, doubling the length of a questionnaire leads to a 30% cost increase (Dolnicar, Grün, and Leisch 2011). Therefore, if data of equal quality can be collected in a quicker way, it should be.

A number of studies have shown that offering fewer response options speeds up the process of completing a survey significantly without negatively affecting data quality (Dolnicar 2003; Driesener and Romaniuk 2006; Grassi et al. 2007; Dolnicar, Grün, and Leisch 2011), thus leading to the recommendation to offer only one or two response options. Given the proneness of the pick-any format (one answer option only) to evasion, the forced-choice full binary format (two response options) is preferable when the aim is to reduce completion times without capturing systematic bias.

Across all criteria of comparison, however, it has to be concluded that there is no single best number of answer options; different answer formats are suitable for different situations (Lehmann and Hulbert 1972; Preston and Colman 2000; Dolnicar and Grün 2009). Yet, currently, commercial market research heavily relies on the pick-any answer format, which leads to substantial evasion by respondents while academics prefer ordinal answer formats (Van der Eijk 2001) which are prone to capturing response styles.

Among the 279 empirical measures reviewed, 240 offered respondents closed answer options rather than asking them to respond in an open-ended manner or rank. Of these 240, 95% offered respondents more than two answer options. Of those, 1% offered respondents three answer alternatives, 4% provided four answer alternatives, 51% offered five, 5% offered six, 38% allowed respondents to choose between seven answer options, and 1% used 11 answer options. These results show that tourism survey researchers have adopted a standard approach of offering five or seven answer options—a dangerous choice given the known susceptibility of multicategory answer formats to response styles. The problem is exacerbated in instances where respondents from different countries participate in the survey because cultural background is a known cause of response styles. In tourism survey research, respondents from different cultural background are common; 48% of recent survey studies in tourism use such samples.

An example is provided by a study that asked respondents from a wide range of cultural backgrounds about their perceptions of being cheated by tourism service workers using a seven-point scale. Responses are compared across countries of origin. It is unclear how much of the difference is due to differences in feeling cheated versus differences in the way respondents from different cultures use seven-point scales.

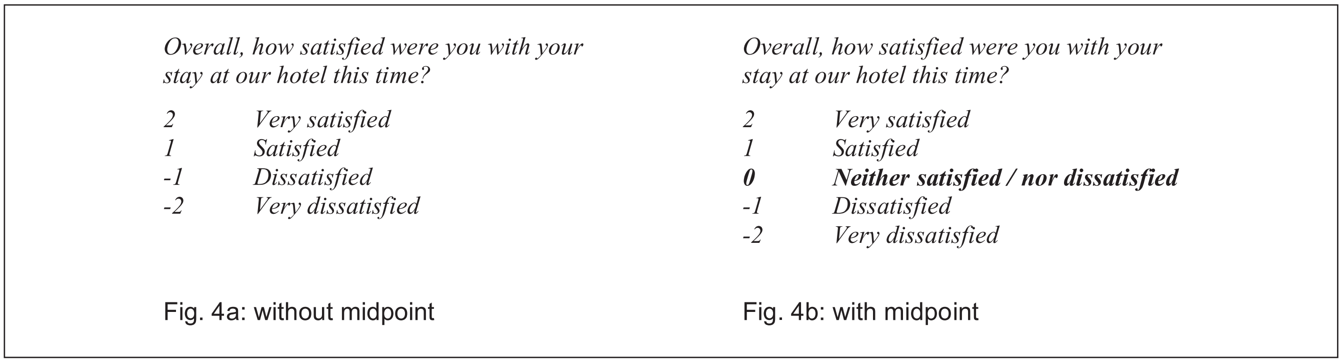
It should also be noted that a justification or explanation why a certain number of answer options is chosen has only been provided for 3 of the 240 measures. Sometimes, authors choose a certain response scale, but then report results at a lower level. In one study data on perceived severity of visitor impacts is collected on a four-point scale. The bottom two and the top two answer options are then added up for reporting, leading to the question: why were more response options used in the first place?

So, no single best number of answer options exists. Yet, randomly choosing a number of answer options is not an acceptable solution, neither is the uncritical adoption of answer options used by others in the past. Rather, researchers should be aware of all alternatives and assess whether answer options are logically meaningful given the query, whether respondents feel that they can express themselves well using the answer options provided, whether the chosen answer format is prone to biases (and if so, what will be done to minimize contamination of data), whether respondents display a high level of stability when asked the same question twice using an answer format, and whether the answer format chosen represents the most efficient option in terms of speed of completion. Sometimes, these questions cannot be answered by the researcher, making additional developmental qualitative research necessary before finalizing the decision on the number of answer options to be provided.

Midpoints

In cases where three or more anticipated bipolar answer options are offered to respondents, the researcher has to decide to include or not to include a "neutral" midpoint. Examples of a survey question with and without midpoint are provided in Figure 4.

A small number of experimental studies have been conducted over the past decades that arrive at similar conclusions: midpoints do not add much value, but have the potential to induce evasion. For example, Schuman and Presser's (1981) experimental work suggests that people who do not have strong feelings, but still lean toward one of



Overall, how satisfied were you with your stay at our hotel this time?

2 Very satisfied
1 Satisfied
-1 Dissatisfied
-2 Very dissatisfied

Fig. 4a: without midpoint

Overall, how satisfied were you with your stay at our hotel this time?

2 Very satisfied
1 Satisfied
0 Neither satisfied / nor dissatisfied
-1 Dissatisfied
-2 Very dissatisfied

Fig. 4b: with midpoint

Figure 4. Response options with and without midpoint.

the answer options, will evade by using midpoints and as a consequence the researchers will lose information. Schuman and Presser recommend, should it be critical to the study to capture the level of magnitude of the response, to first offer an answer option capturing direction only (yes, no) and then follow up with a pure intensity question to determine how strongly respondents feel about the issue (e.g.: How strongly do you feel about that?).

Additional support for Schuman and Presser's results have been provided in recent experiments by Dolnicar and Grün (forthcoming) in which respondents were asked to respond to different survey questionnaire versions, some including "Don't know" response options, some including midpoints or neither of the two. Results of individual-level comparisons indicate that respondents use the midpoint as a substitute for a "Don't know" option. This makes them impossible to interpret: they could mean that a respondent cannot provide a response because they have no opinion, cannot provide a response because they do not feel comfortable answering, or even chooses not to provide a response. As a consequence, Dolnicar and Grün recommend avoiding midpoints because of lack of interpretability of their actual meaning.

Garland (1991), from an experimental study comparing responses when a midpoint was offered with responses when a midpoint was not offered, concludes, first, that this manipulation significantly influences responses, and, second, that the elimination of the midpoint removed social desirability bias. Rossiter (2011), who is generally sceptical about the use of midpoint options because of their susceptibility to evasion, recommends them only in cases when respondents are asked to provide overall evaluations of objects. In this case, it may well be that people are genuinely neutral, he argues.

Among recent tourism studies measuring constructs using non-binary rating scales, the vast majority (91%) offer midpoints, although only 8% state explicitly that they do. There is not a single case among the reviewed studies where a reason was stated for providing a midpoint. Given experimental findings relating to midpoints and the frequent use of midpoints in tourism surveys, it is possible that tourism survey data contains substantial counts of evasion bias. Evasion bias cannot be eliminated after data has been collected and, therefore, can negatively affect the validity of findings.

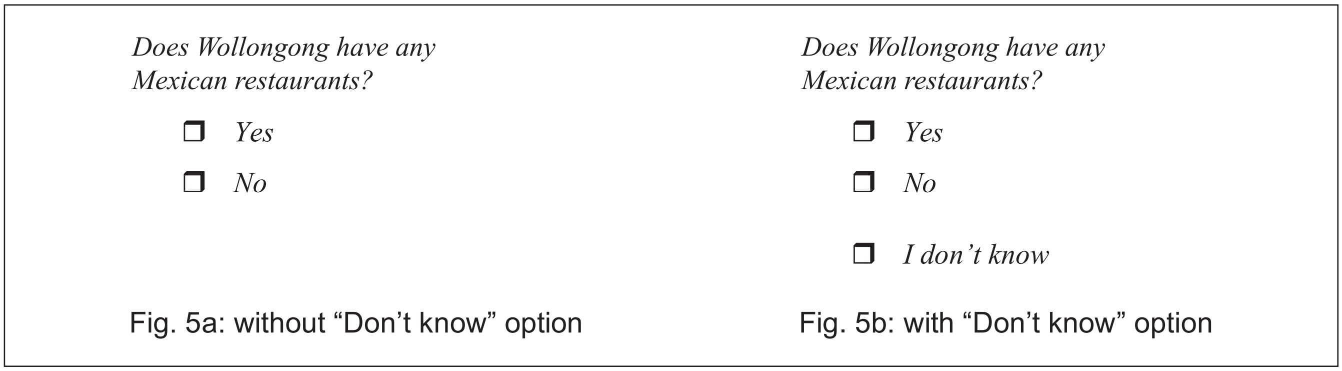
It can be concluded that midpoints should be used only if there is a specific and valid reason to do so (e.g., an overall attitude is measured which may genuinely be neutral for certain respondents). Midpoints are used in different ways by respondents. As such, they cannot be interpreted in a meaningful way; instead they have the potential of distracting respondents from using the anticipated response options provided which can be meaningfully interpreted by the researchers, even if it is "I do not know".

"Don't Know" Options

"Don't know" options in surveys, also referred to as *nonresponse options* allow respondents not to answer a question. The intention of offering such options is that people who are genuinely unable to answer are not forced to. Imagine being asked the question in Figure 5. Unless you have been to Wollongong, it is impossible for you to know and meaningless for you to guess whether or not there are any Mexican restaurants in Wollongong, making the "Don't know" option crucial in ensuring high-quality data.

Generally, two schools of thought exist with respect to offering "Don't know" options in surveys: one postulates that responding with "Don't know" is a characteristic of the respondent (a "relatively stable individual trait"; Rapoport 1982) and not related to the content of the question, much like a response style. If this assumption is correct, "Don't know" responses do not actually reflect respondents' true answers and reduce data quality.

A number of reasons are put forward by proponents of this view. For example, respondents may not feel confident in answering the question and therefore may tick the "Don't know" option unnecessarily. Some evidence for this explanation has been provided by Rapoport (1982): women gave more "Don't know" responses than men, especially when both the women and men had low education levels, so their response is not likely to be driven by knowledge of the answer but to a high extent by their confidence to give a response.



Does Wollongong have any Mexican restaurants?

❐ *Yes*

❐ *No*

Fig. 5a: without “Don’t know” option

Does Wollongong have any Mexican restaurants?

❐ *Yes*

❐ *No*

❐ *I don’t know*

Fig. 5b: with “Don’t know” option

Figure 5. Response options with and without “don’t know” option.

Another reason may be that respondents want to comply with the researcher’s request to complete the survey. Krosnick (1991) refers to this as satisficing. From a satisficing perspective, the “Don’t know” option is a way to comply with the researcher’s request while avoiding the cognitive effort associated with thinking about each question. Krosnick (1999) lists a number of other possible mechanisms that may lead to false “Don’t know” responses such as people not understanding the meaning of the questions, avoiding to answer honestly, or feeling ambivalent about the answer. Empirical evidence for the *respondents don’t want to answer, make them!* school of thought is scarce, mainly because empirical research is based on observation and subjective interpretation of survey responses by researchers rather than explanations by respondents why they did or did not use a “Don’t know” option. For example, Poe et al. (1988) find no differences in reliability of responses in dependence of offering a “Don’t know” response and base their argument against “Don’t know” options on the fact that fewer anticipated responses are given. They provide no proof, however, that the additional anticipated responses given are valid, which is the most important criterion when aiming to collect high-quality empirical data. Gilljam and Granberg (1993) show that survey answers provided by respondents who used “Don’t know” to avoid other survey questions about nuclear power were predictive of their referendum vote. They conclude that they would have been competent to answer the questions and therefore should be made to. But even if the answers they have provided were predictive, they still may have validly stated not to know the answer to the other questions.

The other school of thought argues that “Don’t know” responses are caused by respondents being unable to answer a question. If this assumption is correct, not offering a “Don’t know” answer option would force them to guess, which, in turn, would reduce data quality. Proponents thus argue for offering a “Don’t know” response option to prevent people who genuinely cannot answer the question (Are there any Mexican restaurants in Wollongong?) to guess. Empirical evidence for this argument has been provided by Hawkins and Coney (1981), Schuman and Presser (1981), and most recently, Dolnicar and Grün (forthcoming). These studies show that offering a “Don’t know” option reduces the number of uninformed or guessing responses using quantitative criteria to determine whether or not respondents were familiar enough with the content of the question to be able to answer it. Payne (1980) summarizes this school of thought’s perspective in saying: “If most respondents have no basis for opinions but in effect would have to flip a coin mentally for their answers, it would be better if their answers were not recoded” (p. 23).

In the context of tourism research, Ryan and Garland (1999) recommend the inclusion of “Don’t know” options, pointing out that the nonrandom patterns of “Don’t know” responses represent valuable information in their own right.

Note that studies on “Don’t know” options come from different disciplines: marketing (Hawkins and Coney), tourism (Ryan and Garland), but mostly public opinion research (Krosnick). In opinion research, it can be argued that everyone can have an opinion on pretty much anything. In other contexts, such as destination image measurement, this is hard to argue; if you have never ever heard of Wollongong before reading this article, how could you possibly be expected to competently answer questions about it? In the brand image context, Dolnicar and Rossiter (2008) specifically recommend using “Don’t know” options to ensure that responses relating to unfamiliar brands can be identified as such.

The key insight that emerges from the large body of work on “Don’t know” response options is, whether a “Don’t know” option is offered or not, it is impossible to avoid some contamination—either by incompetent respondents guessing or by competent respondents choosing not to answer. The key question therefore is not which school of thought is right, but how can data contamination be kept to a minimum. Dolnicar and Grün, for their brand image data set, compare the error rates for both cases empirically. They assume that people who are familiar with a fast-food brand and eat there are competent to answer questions about it. When a “Don’t know” option was not offered to respondents, 9% of respondents gave answers about fast-food restaurants they knew nothing about. This means that 9% of the data was likely to be junk, to use Jacoby's term. When a "Don't know" option was provided, 1% of respondents who were familiar with a brand chose the "Don't know" option. This means that only 1% of data was junk. For this particular data set, including a "Don't know" option was clearly better.

In recent tourism survey research studies, "Don't know" options are extremely rare (5%). A good example of using a "Don't know" option is provided by Choi et al. (2012). They offered respondents a "Don't know" response option and provided an explanation why they chose to do so: "N/A stands for a 'no response' or 'don't know' answer. Some respondents were not able to recollect the exact decision time for each decision item" (p. 31). If respondents in tourism studies can reasonably be expected to answer all questions, the low figure of 5% would be no reason for concern, but this appears to be rather unlikely. Even among the 279 constructs measured in the 78 reviewed articles, brand or destination image alone was measured ten times. It is likely that some respondents would not have felt competent to assess all the attributes for all the brands/destinations listed. As a consequence, contamination of tourism survey data by respondent guessing is likely.

It can be concluded that the researcher has to assess whether some respondents will genuinely be unable to provide an answer before finalizing the measure. This can be done logically or by conducting developmental qualitative research, asking a heterogeneous sample if they can answer the questions. If some cannot, a "Don't know" option should be provided. The "Don't know" options should be visually well separated from other response options and should not be overly strongly worded to avoid actively attracting "Don't know" responses (an effect demonstrated by Hippler and Schwarz in 1989).

Verbal Labeling of Anticipated Answer Options

Whether or not answer options should be verbally labeled appears to be one of the few uncontentious areas of questionnaire design with broad consensus that options should be verbally labeled because verbal labeling reduces the variability in the interpretation of answer options by respondents and thus leads to higher reliability of results (Peters and McCormick 1966; Krosnick and Berent 1993; Krosnick 1999). Note, however, that minor differences in wording can affect respondents' interpretations substantially and that the assumption that verbally labeled multicategory options have the same distance between them is very unlikely, as demonstrated by Worcester and Burns (1975), who provide empirically derived numerical values for different verbalizations of bipolar answer formats that can be used in the data set, rather than simply coding the values as +2, +1, 0, –1, and –2, implicitly assuming equidistance.

One of the few studies concluding that purely numerically labeled scale was as good as the verbally labeled scale was provided by Finn (1972). The study was conducted in the context of rating job complexity. Context dependence is the key argument made by Rossiter (2011), who argues theoretically for the use of verbal labels when beliefs are being measured because "beliefs are mentally represented as verbal statements" (p. 20), and for the use of (bipolar) numerical scales when respondents are asked for overall evaluations about something. Again, the reason is that it is likely such overall evaluations are mentally represented in a quantitative way.

For recently published empirical tourism studies, it is extremely difficult to assess how frequently respondents are offered a purely verbally labeled answer format, a purely numerically labeled answer format, or both. The reason is that measures are rarely provided and the description of measures tends to be rudimentary. In cases where some detail on the measure is provided, it is often unclear whether the description refers to the presentation of answer options to respondents or to the way data were coded in the data set. A typical example is illustrated by the following: "Measurements of all the constructs were carried out by the statements adopted from previous studies and a 7-point Likert type scale ranging from (1) Strongly Disagree to (7) Strongly Agree as shown in Appendix 1" with Appendix 1 stating "7-point Likert scale: '1' Disagree Strongly and '7' Agree Strongly." From both the description in the text and the items in the Appendix, it is actually not clear what exactly the respondents saw: the words, the numbers, or both. On the positive side, this is one of the few articles where the actual measures are provided to the reader.

Based on the evidence on the effects of labeling of answer option, it is recommended that answer options be verbally labeled; unless there is a specific, stated reason for not doing so (e.g., overall evaluations are requested from respondents).

Interactive Returns

Online surveys account for approximately 20% of the global research market (Poynter 2010). Of the articles reviewed for this study, 23% report that they collected data online. These figures illustrate that, while uptake in academic survey research is slower than in the commercial market research, online data collection is no longer a marginal occurrence. It is quickly developing to become the main source of survey data.

With the increase in popularity of online surveys and less concern about representativity of data collected online (Dolnicar, Laesser, and Matus 2009), entirely new ways of asking questions in surveys have become feasible. Instead of ticking a box on a piece of paper or even clicking on an icon online, respondents can now drag and drop pictorial symbols of the objects they are asked to assess in the precise order they see fit, or walk through a simulated shop and take products off shelves and place them in a shopping cart.

While market research clients are impressed by the new possibilities and online survey companies emphasize the benefits—primarily the increase in respondent engagement—of these new "funky" question formats, little research has been undertaken to investigate whether they do more good by improving respondents' engagement or more bad by potentially distracting the respondent with "special effects" from what is being asked and thus negatively affecting the validity of questions.

Research to date has largely been limited to testing whether or not interactive survey questions increase respondent engagement (Reid, Morden, and Reid 2007; Puleston and Sleep 2008; Strube et al. 2008; Sleep and Puleston 2009). Almost unanimously, they conclude that this is indeed the case. These studies do not, however, prove that the collected data are better. There is a possibility that respondents had more fun completing the survey but, maybe as a consequence of perceiving it as entertainment, provided less valid responses.

Two studies have attempted to go beyond respondent engagement as the dependent variables and assess whether data quality is in fact better if respondents are offered an interactive way of responding; Delavande and Rohwedder (2008) developed a survey question where respondents threw virtual balls into bins on the computer screen to express their opinions and compared this with a more traditional answer option. They find the results from the dynamic task to be more usable, with *usable* being defined as not violating monotonicity. Dolnicar, Grün, and Yanamandram (forthcoming) compared three response formats; a percentage allocation task, a ranking task, and a dynamic drag-and-drop task in the context of determining how much people felt vacations contributed to their quality of life. They conclude that the responses from the dynamic question format had the highest concurrent validity. In addition, respondents perceived the question format as the easiest and preferred option if they had to redo the survey.

Given the lack of experimental data on interactive questions, the key recommendation for survey researchers at this point is rather general; inspect interactive survey questions carefully to ensure that potential advantages such as increased respondent engagement are harvested without sacrifice in measurement validity.

Conclusions

Measures matter. Bad empirical measures lead to low-quality data and low-quality data, in turn, can lead to incorrect conclusions. Although the quality of empirical measures has improved over the past decades, the status quo of measurement in the social sciences in general and in empirical tourism research, in particular, is still preventing us from gaining the most valid cumulative knowledge in our field or research. Collectively, we as survey researchers can improve the quality of our measures by asking ourselves a few questions:

1. Has a clear and unambiguous definition of the construct under study been provided?

Clear and unambiguous means that, if the definition were to be handed to another researcher, they would come up with the same empirical measure for it. Working with clear and unambiguous definitions is the only way of ensuring that multiple studies investigating the same construct can build on each other, thus genuinely contributing to knowledge. If working with unclear definitions every researcher in fact studies something slightly different and the joint research effort leads to knowing a lot about nothing.

2. Has a justification for the use of a particular measurement theory or scale development paradigm been provided?

Adopting approaches used previously by other researchers in the field without a good reason and without fully understanding the advantages and problems associated with that procedure is dangerous because it may not be true to the original intentions of those who proposed the measurement theory.

3. Has the form and wording of the query been justified?

The way the query is formulated impacts the answers provided by respondents. Was plain everyday language used? Was the query short? Did it include acronyms and technical terms that may be difficult for some respondent to understand? Was the query specific enough to avoid every respondent having a different interpretation of it, etc.? Were all these details provided when study findings were reported? While it seems obvious that all measurement details should be described in detail in the manuscript or report, this is rarely the case. In fact, some of the reviewed articles do not provide one single piece of information about measurement details. For example, in one case, the description of the survey was limited to providing the sample size and indicating that a survey study was conducted. Only marginally better is the description that the "questionnaire was composed of different question types, covering dichotomous, open-ended and multiple choice questions . . . three and five Likert scale questions were also applied" without stating which measure was actually used for which construct and why. Most strikingly, not one article of those reviewed provides an explanation of why it chose to offer certain numbers of response options.

4. Have all parameters of the return been described and justified?

The precise measures used are rarely provided or described in detail. The same holds for justifications of the choice of the return format: for example, whether open or closed returns were invited, whether unipolar or bipolar answer options were offered, whether answer options were verbally labeled or not, whether or not a midpoint and a "Don't know" option were provided, etc. Each of these parameters has a big impact on the responses people will provide. If readers of the study fully understand how responses were given and why, misinterpretations of results can be avoided.

5. Have the actual measures been made available?

To build on an analogy used by Langner (reported in Rossiter 2011): ultimately it does not matter how good the chocolate cake recipe is and how strictly the cook believes they have adhered to it: if it does not look like chocolate cake, does not smell like chocolate cake, and does not taste like chocolate cake, it is not chocolate cake. The same is true with measures: no matter how convinced the researcher is that they have adhered to a certain paradigm of scale development and that their survey questions measure the construct under study, ultimately the reader should be able to assess for themselves if the survey questions are valid measures of the construct as defined by the researcher.

A detailed summary of the review of recent empirical tourism studies, textbook recommendations and recommendations emerging from the present article is provided in the Appendix.

Finally, do not forget: "One thing we most need to learn is that we must stop letting our existing methods and tools dictate and shackle our thinking and research. They are no substitute for using our heads. The brain is still the most important tool we have" (Jacoby 1978, p. 95).

Appendix

Summary of findings and recommendations

	Textbook Recommendations	Empirical Tourism Research	Recommendation
How to define what is being measured?			
	Importance of construct definition is frequently emphasized. No instructions are provided.	Only 37% of the most frequently studied constructs are defined. Definitions for constructs with the same name differ substantially. Most definitions are ambiguous or incomplete.	– Specify the rater – Specify the object (and its components if the object is abstract) – Specify the attribute (and its components if the attribute is abstract)
How many questions to ask?			
Measurement theory/paradigm	Only 20% of textbooks refer to a specific paradigm of measurement: the Churchill paradigm.	Only 20% state the measurement theory. 0% justify their choice of measurement theory. 56% appear to use the Churchill procedure.	– State and justify the measurement theory used (e.g., Churchill's scale development paradigm, Rossiter's C-OAR-SE theory) – Ensure content validity
Single-item measures Multi-item measures	13% of textbooks discuss single- versus multi-item measures, one stating that their use depends on what is being measured.	7% use single-item measures. 92% use multi-item measures. 45% determine components *empirically* using factor analysis.	– Use single-item measures if both the object and the attribute are concrete – Use multi-item measures for abstract objects or attributes – Identify components of abstract objects and attributes conceptually
How to ask a question (the query)?			
Developmental qualitative research	80% of textbooks discuss qualitative research as a source of information for survey development.		– Use developmental qualitative research if you need consumer perceptions or expert knowledge to develop items.
Wording	80% of textbooks provide detailed advice on how to word survey questions, emphasizing mainly clarity and simplicity.	Hard to assess because only 30% make survey questions available to the reader. 41% use agree–disagree answer scales.	– Provide the full questionnaire in an appendix – Use plain, everyday language – Use short queries – Make queries specific – Avoid double-barreled queries – Avoid double negatives – Do not use agree–disagree answer scales as a default – Pretest, pretest, pretest

(continued)

Appendix (continued)

	Textbook Recommendations	Empirical Tourism Research	Recommendation
How to allow respondents to answer (the return)?			
Unipolar versus bipolar	Two textbooks discuss unipolar and bipolar answer formats, 47% illustrate both without discussing in detail.	7% of nonbinary rating scales are declared to be bipolar 93% of nonbinary rating scales were bipolar 89% of bipolar scales are incorrectly numerically labeled.	– Derive logically whether a construct requires a unipolar or bipolar return – Verbalize and numerically code the response options to correctly reflect the polarity of the scale
Number of response options	40% recommend five or more answer options.	5% binary 1% three options 4% four options 51% five options 5% six options 38% seven options 1% eleven options 48% of studies include respondents from different cultural backgrounds.	– Choose the response options by asking the query open ended in a pre test. Use only the main ones, not too few and not too many. – Consider • Stability • Intended reporting • Proneness to bias • Speed – Avoid the "pick any" binary format because it leads to evasion – Consider using DLF IIST (BIT) full binary for measuring belief constructs if you have respondents from different cultural backgrounds to avoid capturing response styles. For unipolar attributes offer YES and NO response options, for bipolar attributes offer AGREE DISAGREE response options.
Midpoints	Midpoints are covered well in textbooks with only 20% not mentioning this topic. Few textbooks make clear recommendations.	91% offer midpoints, but only 10% state that they do. 0% provide an explanation or justification.	– Use a midpoint only for a bipolar attribute in a belief or attitude measure
"Don't know" option	"Don't know" response options are covered well in textbooks with most recommending its use; 20% do not mention the topic.	5% use a "Don't know" option.	– Determine whether some respondents cannot answer the question. If so, offer an explicit, visually separated "Don't know" option
Verbal labeling	Verbal labeling is discussed in 60% of textbooks. Typically advantages and disadvantages are discussed.	Hard to assess because only 30% provide the complete response scale.	– Verbally label all answer options unless the construct is an overall evaluation of the object, which can use a bipolar numerical answer scale with verbal end-anchors

Acknowledgments

The author is grateful to John Rossiter, Rajeev Sharma, Tim Coltman, Andy Carnahan, Emil Juvan, Logi Karlsson, Karen Lovasz, Catherine Frethey-Bentham, and Dan Fesenmaier for their feedback on previous versions of this manuscript and research assistants Aaron Eden and Alexander Chapple for their support with reviewing recently published empirical tourism studies and market research textbooks. This manuscript was prepared partly during the author's employment at the University of Wollongong.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author thanks the Australian Research Council for supporting the research program that has led to this publication (DP0557257, DP0878338).

Notes

1. Reviewed issues: *Journal of Travel Research*: 50(5) 2011, 50(6) 2011, 51(1) 2012, 51(2) 2012, 51(3) 2012; *Tourism Management*: 33(1) 2012, 33(3) 2012, 33(4) 2012, 33(5) 2012; *Annals of Tourism Research*: 38(2) 2011, 38(3) 2011, 38(4) 2011, 39(1) 2012, 39(2) 2012, 39(3) 2012.
2. Amazon sales data was used to determine the most popular books. The list of the 18 reviewed textbooks is available online at http://jtr.sagepub.com/supplemental.

References

Aaker, D. A. (1991). *Managing Brand Equity: Capitalizing on the Value of the Brand Nam*e. New York: Free Press.

Albaum, G. (1997). "The Likert Scale Revisited: An Alternative Version." *Journal of the Market Research Society*, 39 (2): 331-48.

Albaum, G., C. Roster, J. H. Yu, and R. D. Rogers. (2007). "Simple Rating Scale Formats." *International Journal of Market Research*, 49 (5): 633-50.

Bagozzi, R. P. (2011). "Measurement and Meaning in Information Systems and Organizational Research: Methodological and Philosophical Foundations." *MIS Quarterly*, 35 (2): 261-92.

Bednell, D. H. B., and M. Shaw. (2003). "Changing Response Rates in Australian Market Research." *Australasian Journal of Market Research*, 11 (1): 31-41.

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

Bergkvist, L., and J. R. Rossiter. (2007). "The Predictive Validity of Multiple-Item versus Single-Item Measures of the Same Constructs." *Journal of Marketing Research*, 44 (2): 175-84.

Bergkvist, L., and J. R. Rossiter. (2009). "Tailor-Made Single-Item Measures of Doubly Concrete Constructs." *International Journal of Advertising*, 28 (4): 607-21.

Bojanic, D. C., and R. B. Warnick. (2012). "The Role of Purchase Decision Involvement in a Special Event." *Journal of Travel Research*, 51 (3): 357-66.

Cantril, H. (1940). *Gauging Public Opinion*. Princeton, NJ: Princeton University Press.

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.

Choi, S., X. Y. Lehto, A. M. Morrison, and S. Jang. (2012). "Structure of Travel Planning Processes and Information Use Patterns." *Journal of Travel Research*, 51 (1): 26-40.

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

Churchill, G. A., Jr. (1998). "Measurement in Marketing: Time to Refocus?" In *Paul D. Converse Symposium*, edited by James D. Hess and Kent B. Monroe. Chicago: American Marketing Association, 25-41.

Converse, J. M., and S. Presser. (1986). Survey Questions—Handicrafting the Standardized Questionnaire. Sage series on Quantitative Applications in the Social Sciences, Number 63. Newbury Park: Sage.

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

Cunningham, W. H., I. C. M. Cunningham, and R. T. Green. (1977). "The Ipsative Process to Reduce Response Set Bias." *Public Opinion Quarterly*, 41 (3): 379-84.

Dall'Olmo Riley, F., A. S. C. Ehrenberg, S. B. Castleberry, T. P. Barwise, and N. R. Barnard. (1997). "The Variability of Attitudinal Repeat-Rates." *International Journal of Research in Marketing*, 14 (5): 437-50.

Delavande, A., and S. Rohwedder. (2008). "Eliciting Subjective Probabilities in Internet Surveys." *Public Opinion Quarterly*, 72 (5): 866-91.

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., and B. Grün. (2013). "Validly Measuring Destination Image in Survey Studies." *Journal of Travel Research*, 52 (1): 3-12.

Dolnicar, S., and B. Grün. (2007a). "How Constrained a Response: A Comparison of Binary, Ordinal and Metric Answer Formats." *Journal of Retailing and Consumer Services*, 14: 108-22.

Dolnicar, S., and B. Grün. (2007b). "User-Friendliness of Answer Formats—An Empirical Comparison." *Australasian Journal of Market & Social Research*, 15 (1): 19-28.

Dolnicar, S., and B. Grün. (2007c). "Assessing Analytical Robustness in Cross-Cultural Comparisons." *International Journal of Tourism, Culture, and Hospitality Research*, 1 (2): 140-60.

Dolnicar, S., and B. Grün. (2007d). "Question Stability in Brand Image Measurement—Comparing Alternative Answer Formats and Accounting for Heterogeneity in Descriptive Models." *Australasian Marketing Journal*, 15 (2): 26-41.

Dolnicar, S., and B. Grün. (2009). "Does One Size Fit All? The Suitability of Answer Formats for Different Constructs Measured." *Australasian Marketing Journal*, 17 (1): 58-64.

Dolnicar, S., and B. Grün. (Forthcoming). "Including Don't Know Answer Options in Brand Image Surveys Improves Data Quality." *International Journal of Market Research*.

Dolnicar, S., B. Grün, and F. Leisch. (2011). "Quick, Simple and Reliable: Forced Binary Survey Questions." *International Journal of Market Research*, 53 (2): 231-52.

Dolnicar, S., C. Laesser, and K. Matus. (2009). "Online Versus Paper—Format Effects in Tourism Surveys." *Journal of Travel Research*, 47 (3): 295-316.

Dolnicar, S., and J. R. Rossiter. (2008). "The Low Stability of Brand-Attribute Associations Is Partly Due to Measurement Factors." *International Journal of Research in Marketing*, 25 (2): 104-8.

Dolnicar, S., B. Grün, and V. Yanamandram. (Forthcoming). "New Ways of Asking Survey Questions—More Fun, More Valid Data." *Journal of Travel & Tourism Marketing*.

Dolnicar, S., J. R. Rossiter, and B. Grün. (2012). "Pick-Any Measures Contaminate Brand Image Studies." *International Journal of Market Research*, 54 (6): 821-34.

Driesener, C., and J. Romaniuk. (2006). "*Comparing Methods of Brand Image Measurement*." *International Journal of Market Research*, 48 (6): 681-98.

Drolet, A. L., and D. G. Morrison. (2001). "Do We Really Need Multiple-Item Measures in Service Research?" *Journal of Service Research*, 3 (3): 196-204.

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): 255-65.

Garland, R. (1991). "The Mid-Point on a Rating Scale: Is It Desirable?" *Marketing Bulletin*, 2: 66-70.

Gilljam, M., and D. Granberg. (1993). "Should We Take Don't Know for an Answer?" *Public Opinion Quarterly*, 57:348-57.

Grassi, M., A. Nucera, E. Zanolin, E. Omenaas, J. M. Anto, and B. Leynaert. (2007). "Performance Comparison of Likert and Binary Formats of SF-36 Version 1.6 across ECRHS II Adult Populations." *Value in Health*, 10 (6): 478-88.

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: 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 (2): 176-88.

Greenleaf, E. A. (1992b). "Measuring Extreme Response Style." *Public Opinion Quarterly*, 56 (3): 328-51.

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): 147-54.

Hardie, T., and N. Kosomitis. (2005). The F Word in Market Research: High Impact Public Relations in the 21st Century. CD Proceedings from the AMSRS Conference 2005: Impact. Sydney: Australian Market & Social Research Society.

Hawkins, D. I., and K. A. Coney. (1981). "Uninformed Response Error in Survey Research." *Journal of Marketing Research*, 18 (3): 370-74.

Heide, M., and K. Gronhaug. (1992). "The Impact of Response Styles in Surveys: A Stimulation Study." *Journal of the Market Research Society*, 34 (3): 215-30.

Hippler, H. J., and N. Schwarz. (1989). "'No Opinion'-Filters: A Cognitive Perspective." *International Journal of Public Opinion Research*, 1 (1): 77-87.

Horng, J.-S., C.-H. Liu, H.-Y. Chou, and C.-Y. Tsai. (2012). "Understanding the Impact of Culinary Brand Equity and Destination Familiarity on Travel Intentions." *Tourism Management*, 33 (4): 815-24.

Hsu, C. H. C., H. Oh, and A. G. Assaf. (2012). "A Customer-Based Brand Equity Model for Upscale Hotels." *Journal of Travel Research*, 51 (1): 81-93.

Jacoby, J. (1978). "Consumer Research: How Valid and Useful Are All Our Consumer Behaviour Research Findings? A State of the Art Review." *Journal of Marketing*, 42 (2): 87-95.

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, and 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, 247-48.

Kahane, H. (1982). *Logic and Philosophy: A Modern Introduction*. 4th edition. Belmont, CA: Wadsworth.

Kampen, J., and M. Swyngedouw. (2000). "The Ordinal Controversy Revisited." *Quality and Quantity*, 34 (1): 87-102.

Kollat, D. T., R. D. Blackwell, and J. E. Engel. (1972). "The Current Status of Consumer Behaviour Research: Development during the 1968-1972 Period." In *Proceedings of the 3rd Annual Conference of the Association of Consumer Research*, edited by M. Venkatesan, pp. 576-84.

Komorita, S. S. (1963). "Attitude Content, Intensity, and the Neutral Point on a Likert Scale." *Journal of Social Psychology*, 61: 327-34.

Komorita, S. S., and W. K. Graham. (1965). "Number of Scale Points and the Reliability of Scales." *Educational and Psychological Measurement*, 25 (4): 987-95.

Krosnick, J. A. (1991). "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology*, 5: 213-36.

Krosnick, J. A. (1999). "Survey Research." *Annual Review of Psychology*, 50: 537-67.

Krosnick, J. A., and M. K. Berent. (1993). "Comparisons of Party Identification and Policy Preferences: The Impact of Survey Question Format." *American Journal of Political Science*, 37 (3): 941-64.

Lee, J. A., G. N. Soutar, J. Louviere, and T. M. Daly. (2006). An Examination of the Relationship between Values and Holiday Benefits across Cultures Using Ratings Scales and Best-Worst Scaling. ANZMAC CD Proceedings.

Lehmann, D. R., and J. Hulbert. (1972). "Are Three Point Scales Always Good Enough?" *Journal of Marketing Research*, 9 (4): 444-46.

Likert, R. (1932). "A Technique for the Measurement of Attitudes." *Archives of Psychology*, 36 (2): 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): 353-62.

Martin, W. S. (1973). "The Effects of Scaling on the Correlation Coefficient: A Test of Validity." *Journal of Marketing Research*, 10 (3): 316-18.

Martin, W. S. (1978). "Effects of Scaling on the Correlation Coefficient: Additional Considerations." *Journal of Marketing Research*, 15 (2): 304-8.

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: 537-45.

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.

Nam, J., Yuksel Ekinci, and G. Whyatt. (2011). "Brand Equity, Brand Loyalty and Consumer Satisfaction." *Annals of Tourism Research*, 38 (3): 1009-30.

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

Nunnally, J. C. (1978). *Psychometric Theory*. 3rd edition. New York: McGraw-Hill.

Nunnally, J. C. (1970). *Introduction to Psychological Measurement*. New York: McGraw-Hill.

Oaster, T. R. F. (1989). "Number of Alternatives per Choice Point and Stability of Likert-Type Scales." *Perceptual and Motor Skills*, 68: 549-50.

Oliver, R. L. (1999). "Whence Consumer Loyalty?" *Journal of Marketing*, 63:33-44.

Paulhus, D. L. (1991). "Measurement and Control of Response Bias." In *Measures of Personality and Social Psychological Attitudes*, edited by J. P. Robinson, P. R. Shaver, and L. S. Wrightsman. San Diego: Academic Press, pp. 17-59.

Payne, S. L. (1980). *The Art of Asking Questions, 13th edition*. Princeton: Princeton University Press.

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*, edited by B. Anderson. Ann Arbor, MI: Association for Consumer Research, pp. 143-48.

Peters, D. L., and E. J. McCormick. (1966). "Comparative Reliability of Numerically Anchored versus Job-Task Anchored Rating Scales." *Journal of Applied Psychology*, 50 (1): 92-96.

Poe, G. S., I. Seeman, J. McLaughlin, E. Mehl, and M. Dietz. (1988). "Don't Know Boxes in Factual Questions in a Mail Questionnaire." *Public Opinion Quarterly*, 52: 212-22.

Poynter, R. (2010). *The Handbook of Online and Social Media Research. Tools and Techniques for Market Researchers*. Chichester, UK: Wiley.

Prayag, G. and C. Ryan. (2012). "Antecedents of Tourists' loyalty to Mauritius: the role and influence of destination image, place attachment, personal involvement and satisfaction." *Journal of Travel Research*, 51(3): 342-356.

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.

Puleston, J., and D. Sleep. (2008). Measuring the Value of Respondent Engagement: Innovative Techniques to Improve Panel Quality, *ESOMAR Congress* 2008.

Rapoport, R. B. (1982). "Sex Differences in Attitude Expression: A Generational Explanation." *Public Opinion Quarterly*, 46: 86-96.

Reid, J., M. Morden, and A. Reid. (2007). "Maximizing Respondents Engagement." In *Proceedings of the ESOMAR Congress 2007*. Amsterdam: Esomar.

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: 765-70.

Rossiter, J. R. (2002). "The C-OAR-SE Procedure for Scale Development in Marketing." *International Journal of Research in Marketing*, 19 (4): 305-35.

Rossiter, J. R. (2011). Measurement for the Social Sciences: The C-OAR-SE Method and Why It Must Replace Psychometrics. New York: Springer.

Rossiter, J. R., and L. Bergkvist. (2009). "Tailor-Made Single-Item Measures of Doubly Concrete Constructs." *International Journal of Advertising*, 28 (4): 607-21.

Rossiter, J. R., S. Dolnicar, and B. Grün. (2010). "The LFFB Comparative Judgment Measure of Brand Attribute Beliefs*."* Working Paper, Faculty of Commerce, University of Wollongong, Australia.

Rungie, C., G. Laurent, F. Dall'Olmo Riley, D. G. Riley Morrison, and T. Roy. (2005). "Measuring and Modeling the (Limited) Reliability of Free Choice Attitude Questions." *International Journal of Research in Marketing*, 22 (3): 309-18.

Ryan, C., and R. Garland. (1999). "The Use of a Specific Non-response Option on Likert-Type Scales." *Tourism Management*, 20: 107-13.

Schuman, H., and S. Presser. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. New York: Academic Press.

Sleep, D., and J. Puleston. (2009). "Panel Quality: Leveraging Interactive Techniques to Engage Online Respondents." ARF Convention & Expo 2009.

Strube, S. N., Y. Zdanowicz, C. Ryan, and K. Tough. (2008). *Maximizing Respondent Engagement through Survey Design*. *CASRO Panel Conference Miami*.

Symonds, P. M. (1924). "On the Loss of Reliability in Ratings Due to Coarseness of the Scale." *Journal of Experimental Psychology*, 7: 456-61.

Van de Vijver, F. J. R., and Y. H. Poortinga. (2002). "On the Study of Culture in Developmental Science." *Human Development*, 45 (4): 246-56.

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

Vriens, M., M. Wedel, and Z. Sandor. (2001). "Split-Questionnaire Designs: A New Tool in Survey Design and Panel Management." *Marketing Research*, 13 (1): 14-19.

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

Welkenhuysen-Gybels, J., J. Billiet, and B. Cambré. (2003). "Adjustment for Acquiescence in the Assessment of the Construct Equivalence of Likert-Type Score Items." *Journal of Cross-Cultural Psychology*, 34 (6): 702-22.

Worcester, R. M., and T. R. Burns. (1975). "A Statistical Examination of the Relative Precision of Verbal Scales." *Journal of the Market Research Society*, 17 (3): 181-97.

Yu, J. H., G. Albaum, and M. Swenson. (2003). "Is a Central Tendency Error Inherent in the Use of Semantic Differential Scales in Different Cultures?" *International Journal of Market Research*, 45 (2): 213-28.

Author Biography

Sara Dolnicar's key research interests include understanding and improving the validity of quantitative measures in the social sciences and improving data-driven market segmentation studies both in terms of their conceptual foundation and methodology. While much of Sara's applied research focused on sustainable tourism, destination image and tourism marketing more generally, she has also studied topics outside of the tourism discipline, including marketing approaches to the effective recruitment of successful foster parents and ways of increasing people's engagement in water conservation and their acceptance of recycled water.