J. of the Acad. Mark. Sci.
DOI 10.1007/s11747-017-0532-y

METHODOLOGICAL PAPER

Marketing survey research best practices: evidence and recommendations from a review of *JAMS* articles

John Hulland[1] · **Hans Baumgartner**[2] · **Keith Marion Smith**[3]

Received: 19 August 2016 / Accepted: 29 March 2017

Abstract Survey research methodology is widely used in marketing, and it is important for both the field and individual researchers to follow stringent guidelines to ensure that meaningful insights are attained. To assess the extent to which marketing researchers are utilizing best practices in designing, administering, and analyzing surveys, we review the prevalence of published empirical survey work during the 2006–2015 period in three top marketing journals—*Journal of the Academy of Marketing Science* (*JAMS*), *Journal of Marketing* (*JM*), and *Journal of Marketing Research* (*JMR*)—and then conduct an in-depth analysis of 202 survey-based studies published in *JAMS*. We focus on key issues in two broad areas of survey research (issues related to the choice of the object of measurement and selection of raters, and issues related to the measurement of the constructs of interest), and we describe conceptual considerations related to each specific issue, review how marketing researchers have attended to these issues in their published work, and identify appropriate best practices.

Keywords Survey research · Best practices · Literature review · Survey error · Common method variance · Non-response error

Aric Rindfleisch served as Guest Editor for this article.

✉ John Hulland
jhulland@uga.edu

[1] Terry College of Business, University of Georgia, 104 Brooks Hall, Athens, GA 30602, USA

[2] Smeal College of Business, Penn State University, State College, PA, USA

[3] D'Amore-McKim School of Business, Northeastern University, Boston, MA, USA

Surveys are ubiquitous, used to inform decision making in every walk of life. Surveys are also popular in academic marketing research, in part because it is difficult to imagine how certain topics could be studied without directly asking people questions, rather than, say, observing their behaviors, possibly in response to different experimental conditions manipulated by the researcher. In their review of academic marketing research published in the *Journal of Marketing* (*JM*) and the *Journal of Marketing Research* (*JMR*) between 1996 and 2005, Rindfleisch et al. (2008) found that roughly 30% of the articles—representing 178 published papers—used survey methods. In this research, we conduct a follow-up investigation of the use of surveys during the decade since their review (i.e., 2006 to 2015), adding the *Journal of the Academy of Marketing Science* (*JAMS*) to the set of journals studied since (1) many articles published in *JAMS* rely on surveys and (2) *JAMS* has an impact factor comparable to *JM* and *JMR*. We classify each article as either survey-based or non-survey-based empirical work, as a conceptual paper, or as something else (most typically an editorial or commentary). A summary of these results (for survey work), by year and by journal, is shown in Fig. 1.

The total numbers of papers published in *JAMS*, *JM*, and *JMR* over the 10-year period were 436, 489, and 636, respectively; the corresponding numbers of empirical research papers based on surveys were 202, 212, and 108. Overall, we found that about a third of the papers published in this period were survey-based (33.4%), a figure very similar to the one reported by Rindfleisch et al. Both *JAMS* (46.3% of all papers) and *JM* (43.4%) contained more survey work than *JMR* (17.0%). Although surveys were less common in *JMR*, our findings indicate overall that survey techniques continue to play an important role in academic research.

While it is easy to assemble and administer a survey, there are many sources of error that can contaminate survey results

Published online: 10 April 2017

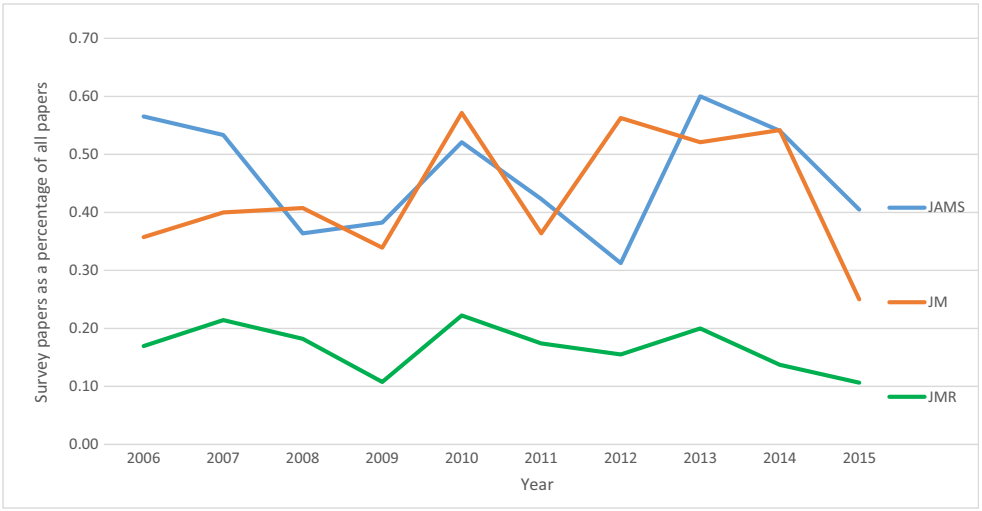

Fig. 1 Survey-based research as a percentage of all papers published in three journals (2006–2015)

and limit the usefulness of survey findings. This has led some researchers to distrust survey results (e.g., Kamakura 2001; Wittink 2004; see also Huber et al. 2014; Rindfleisch 2012).[1] Although we believe that surveys can provide novel and revelatory insights into both the minds of individuals (e.g., consumers, employees) and the practices of organizations, researchers have to pay careful attention to survey design and administration so that the results they obtain are meaningful. Rindfleisch et al. (2008) discuss several key threats to survey validity and provide valuable guidelines to survey researchers, but their focus is on the relative merits of cross-sectional and longitudinal surveys with regard to the control of common method bias and the validity of causal inferences.

Our aim in this paper is broader, with our review covering two sets of decisions that researchers have to make when designing and analyzing a survey. First, whom does the survey describe, and if the survey does not involve self-reports, who provides the desired data? Groves et al. (2004) call this "representation," because the question deals with whom the survey represents. Second, what is the survey about or, more specifically, what does the survey measure, and do the observed responses actually measure what the survey researcher is interested in? Groves et al. (2004) refer to this as "measurement." Errors that can invalidate survey results may arise in both areas, and we discuss a variety of important issues related to representation and measurement in more detail below.

We also perform a comprehensive empirical review of published survey research in an effort to assess the state-of-the-art and identify potential weaknesses. In particular, we analyze all survey-based empirical articles published in *JAMS* between 2006 and 2015. We chose *JAMS* because it is a premier outlet for survey-based research; we restricted our review to *JAMS* to keep the task manageable. We specifically focus on important issues related to representation and measurement for which we were able to gather relevant empirical data based on our analysis of published research. We developed a coding scheme based on the survey literature and a preliminary analysis of a subset of the articles to be analyzed, content-analyzed all articles, and determined in which areas researchers use best practices and in which areas there is room for improvement. In the remainder of the paper, we summarize our findings and derive recommendations with an eye toward improving survey practices.

Our review shows that survey researchers publishing in *JAMS* seem to have a deep appreciation for the importance of managing measurement error (especially random measurement error) in their surveys. Constructs are conceptualized carefully, usually multiple items are used to operationalize abstract constructs, and extensive measurement analyses are conducted before the hypotheses of interest are tested. At the same time, we also identify two broad weaknesses frequently encountered in survey papers.

First, survey researchers follow certain "established" survey practices ritualistically, even when they are not directly relevant to the research in question or when they are not particularly meaningful. For example, a focus on coverage, sampling, and non-response error seems misplaced when there is no real target population to which the researcher wants to generalize the findings. As another example, tests of non-response bias are often conducted in an almost automated fashion, with little apparent real desire to reveal any selection bias that may exist.

Second, survey researchers continue to struggle with certain issues commonly acknowledged to constitute a serious

[1] In conducting their topical review of publications in *JMR*, Huber et al. (2014) show evidence that the incidence of survey work has declined, particularly as new editors more skeptical of the survey method have emerged. They conclude (p. 88)—in looking at the results of their correspondence analysis—that survey research is more of a peripheral than a core topic in marketing. This perspective seems to be more prevalent in *JMR* than in *JM* and *JAMS*, as we note above.

threat to the validity of surveys. In particular, although survey researchers recognize the importance of assessing and controlling common method bias, they do not always deal effectively with this threat. In the pages that follow, we describe the aforementioned issues in greater detail and offer recommendations for improving survey practices, after first describing in more detail how we conducted our empirical review.

Empirical review

In order to assess the extent to which marketing survey researchers are utilizing best practices in deciding whom to ask and in designing their survey instruments, we conducted an in-depth analysis of all survey-based studies published in *JAMS* from 2006 through 2015. We first developed a preliminary coding protocol of issues related to both representation (unit of analysis, source of data, sampling frame and type of sampling, response rate, method used to check for nonresponse bias, etc.) and measurement (pretesting of instrument, extent of use of multi-item scales and assessment of the quality of construct measurement, investigation of common method variance via various a priori and post hoc methods, etc.) that we felt were important for survey validity and that could be coded meaningfully and efficiently for a large collection of articles.

We then randomly selected an initial set of 30 survey-based papers from the journal (with roughly equal representation across all ten years). One paper was subsequently identified as non-survey related and dropped from further consideration. Furthermore, several papers employed more than a single study, resulting in a total of 32 initial survey studies published in 29 papers. All three authors individually coded all 32 studies based on the preliminary coding protocol. Across 62 coding categories, Fleiss' extension of Cohen's kappa to multiple raters yielded 48 coefficients above .8 (almost perfect agreement), three between .6 and .8 (substantial agreement), and ten between .4 and .6 (moderate agreement); only one coefficient was between .2 and .4 (fair agreement).[2] All kappa coefficients indicated significant inter-rater agreement. Coding discrepancies were discussed and resolved, and the coding protocol was refined slightly to simplify the coding of all remaining studies, which were coded by one of the three authors (with all three authors coding roughly the same number of papers).

As noted earlier, a total of 202 papers published in *JAMS* were identified as survey-based. While some papers reported multiple studies, these additional studies were typically either not relevant (e.g., a paper might report both an experiment and a survey, but our focus here is solely on surveys) or they used very similar survey procedures, thus yielding limited incremental information. For simplicity (e.g., to avoid dependencies introduced by using multiple studies from the same paper), we therefore treated the paper as the unit of analysis. If multiple surveys were reported in the same paper, we used the most representative study. All of our subsequent analyses and discussion are based on the full set of 202 papers (including the initial subset of 29).

The vast majority of survey papers (n = 186; 92.1%) published in *JAMS* during this period employed a cross-sectional approach to survey design and analysis. This result is very consistent with the 94% figure found by Rindfleisch et al. (2008). The few papers using longitudinal data collection designs generally included multi-wave analyses of the data. Because the number of longitudinal studies is small, we do not further discuss longitudinal design issues here.

We categorized the 202 papers by mode of data collection and found that mail (44.6%) or electronically-based (30.7%) surveys were most common. Fewer researchers used phone (7.4%) or in-person (13.9%) administration techniques, while a small number (about 2%) used some other type of data collection (e.g., fax surveys, questionnaires distributed at churches or meetings).[3] For 12.4% of the studies, the mode employed was not clear. While this mix of data collection methods appears to be generally appropriate (i.e., different research questions, respondents, and contexts presumably require use of different data collection modes), we encourage all survey researchers to clearly describe how their data were collected because different data collection modes have distinct, particular weaknesses (e.g., in-person interviews may be more susceptible to social desirability biases).

Using logistic regression, we analyzed whether the use of different data collection modes changed over the 10-year period under review. This analysis shows that the use of mail surveys decreased over time (the linear trend was negative and significant, $\chi^2(1) = 15.4$, $p < .0001$), but there were no other significant effects. Although use of electronic surveys has gone up somewhat, as might be expected, this effect was not reliable ($\chi^2(1) = 2.3$, $p = .13$). The trend toward increased academic use of online surveys and decreased use of mail mirrors a similar shift in mode choice by commercial survey researchers.[4]

Below, we more fully report the findings from our review of survey research. The specific issues investigated are grouped into two broad categories: issues related to the choice of the object of measurement and the selection of raters (survey unit representation) and issues related to the design of the survey instrument used to obtain construct measures (measurement of constructs). For each of the issues identified, we provide an overview of relevant considerations and report the

[2] A copy of the coding scheme used is available from the first author.

[3] Several studies used more than one mode.

[4] Traditionally, commercial researchers used phone as their primary collection mode. Today, 60% of commercial studies are conducted online (CASRO 2015), growing at a rate of roughly 8% per year.

findings from our review. On the basis of these findings, we also provide recommendations on how to improve survey practices and point to papers that utilize good practices.

Issues related to survey unit representation

Survey unit representation deals with the question of whom the survey describes and who provides the desired data (in cases where the survey does not involve self-reports). When the researcher tries to generalize the findings from the particular units represented in the study to some broader population, this includes issues related to the definition of the target population, sampling, and non-response bias. However, in typical academic marketing surveys there is often no obvious target population to which the researcher wants to generalize the findings, the sample studied is arbitrary (e.g., chosen based on ease of access), and it is difficult to talk about selection bias when the sample is one of convenience (although non-response will lead to loss of power).

If the goal is to test theoretical hypotheses of interest, as is usually the case, the most important consideration is to select measurement objects and a research context in which the hypotheses can be meaningfully tested, using sources of data (both primary and secondary) that yield accurate information about the units studied. In cases in which the researcher is interested in making inferences to some underlying population, the sampling procedure has to be described in detail, and the calculation of response rates and the assessment of potential non-response bias (including the elimination of respondents by the researcher) assume greater importance. These are the issues discussed in this section.

Choice of measurement object and selection of raters to provide measurements

The research question will generally determine the object of measurement. In marketing surveys, the object of measurement is usually either an individual (e.g., consumer, salesperson) or a firm (although other units are possible, such as ads). If individuals serve as the unit of analysis, survey measures are often based on self-report, although it is sometimes desirable to collect ratings about an individual from observers (e.g., a salesperson's job performance is rated by his or her supervisor). When the firm is the unit of analysis, the researcher needs to take care to select appropriate informants (Phillips 1981). In some cases this may mean using multiple informants for each unit of analysis, or relying on different respondents to provide information for different constructs (depending on who is most knowledgeable about a construct). At times, no specific rater is involved and secondary data are used to measure constructs of interest (e.g., publicly available secondary data).

Casual observation indicates that consumer researchers generally pay little attention to selecting appropriate respondents, typically assuming that everyone is a consumer and that undergraduate students are just as representative as other consumers (but see Wells 1993 for a critique of this position). In recent years, online surveys based on professional survey takers who receive minimal pay for completing brief questionnaires have become popular (e.g., Amazon's Mechanical Turk [or MTurk] panel). Research shows that MTurk samples are not necessarily worse than other samples for some purposes and may in fact be more representative of the general population (e.g., compared to participants from subject pools), although MTurk respondents may be less attentive and less suitable for more complex surveys (e.g., Goodman et al. 2013; Paolacci et al. 2010). In particular, the presence of "professional" respondents in online panels is a concern (Baker et al. 2010; Hillygus et al. 2014). Professional respondents are experienced survey-takers who actively participate in a large number of surveys, primarily to receive cash or other incentives offered by the researchers (although the pay is usually minimal).

In contrast, managerial researchers usually pay much more attention to describing the population of objects with which the research is concerned (e.g., firms in certain SIC codes) and devote considerable care to identifying suitable respondents. However, it is still necessary to justify why a certain population was chosen for study (e.g., why those particular SIC codes?) and why the sources of information used in the study (e.g., key informants, secondary financial data) are most appropriate to serve as measures of the underlying constructs (which are often more complex than constructs investigated using self-report measures).

For the papers in our review, the unit of analysis was typically either an individual (52.0%) or a firm (40.1%), with the balance being something else (e.g., SBUs, teams). Mirroring this, slightly over half of the studies (54.5%) involved self-reports, and about half of the studies relied on responses from key informants (50.0%). A subset of papers (n = 43; 21.3%) also reported use of additional information sources such as secondary data. As one might expect, the unit of analysis used (individual versus other) and the source of reporting (self versus other) are highly inter-related (Pearson χ^2 (1) = 153.4, $p < .001$). When key informants were used, the median number of respondents was 1 (mean = 2.68; max =50), with most studies relying on one (66.0%) or two (19.2%) informants.

Description of sampling

When the primary aim of the research is to test the veracity of proposed theoretical effects, the use of a convenience sample may suffice. However, if the researcher is interested in a particular target population to which generalizations are to be made, an explicit sampling frame has to be specified and the

manner in which the sample was drawn needs to be clearly and explicitly described. Generalizations to the target population require the choice of sample to be justified, and when the sampling scheme is complex particular care is required because the analysis has to take into account these complexities (among other things).

Based on our review of published survey research, we classified 111 studies (55.0%) as using an explicit sampling frame, and 86 studies (42.6%) as being based on convenience samples.[5] The use of a convenience sample is significantly associated with both the individual unit of analysis (Pearson χ^2 (1) = 36.8, $p < .001$) and self-reporting (Pearson χ^2 (1) = 32.6, $p < .001$). For those studies employing a frame, 58.6% used simple random sampling to form their sample and 16.2% used a more complex sampling scheme (e.g., stratified or clustered samples). A few studies were based on a census of the sampling frame. In many cases we found the descriptions provided for the more complex sampling procedures to be confusing and incomplete. When an explicit sampling frame is used, it should be described clearly and the choice should be justified in the context of the research questions. Complex sampling schemes magnify the need for clarity and transparency in sampling procedures, as these designs often impact the selection and specification of analysis techniques.

Calculation of response rate and assessment of potential non-response bias

Non-response problems can arise at both the overall response unit and individual measurement item levels (Berinsky 2008). Unit non-response occurs when a potential observation unit included in the initial sample (e.g., an individual) is missing entirely from the final achieved sample (Lohr 1999). Item non-response occurs when a respondent does not provide data for all items. While both sources of non-response are important, we focus here on unit non-response. (For an overview of approaches used to deal with missing individual item data, see Graham 2009).

Groves and Couper (2012) suggest that unit non-response is affected by two distinct processes: (1) an inability to reach the potential respondent (i.e., a contact failure) and (2) declined survey participation (i.e., a cooperation failure). The choice of mode used for survey data collection (e.g., mail, electronic, phone) can affect the extent to which potential respondents are both able and willing to participate in a study. For example, face-to-face surveying techniques are often compromised by the inability to reach all members of the sampling frame, and consumers are frequently unwilling to answer phone calls from unknown third parties. On the other hand, consumers appear to be both more open and available to online surveys. Published evidence suggests that non-response is more the result of respondent refusal than an inability to reach potential respondents (Curtin et al. 2005; Weisberg 2005).

Two issues arise in connection with the problem of unit non-response. The first issue is how the incidence of non-response is calculated. Although different formulas for computing a response rate exist (see *The American Association for Public Opinion Research* 2016), in general a response rate refers to the ratio of the number of completed surveys over the number of eligible reporting units. Response rates may differ depending on what is counted as a completed interview (e.g., returned surveys vs. surveys with complete data) and how the denominator of eligible reporting units is defined (e.g., whether only pre-qualified respondents were counted as eligible reporting units).

Our review of survey-based studies in *JAMS* shows that there is considerable variation in how response rates are calculated (making it difficult to compare response rates across studies). Often, researchers seem to be motivated to make their response rates look as good as possible. The reason for this is presumably that low response rates are regarded with suspicion by reviewers. Based on this finding, we have two recommendations. On the one hand, we urge researchers to be honest about the effective response rate in their studies. On the other hand, reviewers should not automatically discount research based on lower response rates, because recent evidence has shown that the presumed positive relationship between response rates and survey quality does not always hold (*The American Association for Public Opinion Research* 2016). In fact, low response rates are not necessarily a problem for theory-testing unless there is reasonable concern that the responding sample is systematically different from the sample that did not respond with respect to the hypotheses investigated in the study. Of course, lower response rates also reduce the power of statistical tests.

One key aspect that should be discussed by researchers where relevant is the extent to which any pre-qualification procedures or preliminary contacts were used to winnow down the initial sample, prior to survey distribution. When pre-qualification—particularly participant pre-commitment to participate in the survey—has been obtained, non-response to the initial request may in fact be driven by a systematic difference between the initially targeted sample and the eventual sample (e.g., due to the sensitive nature of the issues covered, differences in interest in the topic). In other words, the higher response rate associated with use of a pre-qualified sample may actually be a misleading indicator of lack of bias. In general, researchers should disclose all relevant samples sizes, including the original sample contacted, the number agreeing to participate, the number receiving the survey, and the ultimate number of respondents who completed the survey.

[5] Although the two categories are not necessarily mutually exclusive, the overlap was small ($n = 4$).

Keeping in mind the caveat about comparing response rates across studies (because of the differences in how response rates are calculated), the average reported response rate in our review was 37.7% (sd = 21.6), with a median value of 33.3% and with a small number of studies reporting much higher response rates. Some researchers went to considerable lengths to achieve higher response rates, including repeated follow-up attempts, careful design of the survey instrument to minimize completion costs, and use of multiple response modes.

The second issue related to unit non-response is whether those individuals completing the survey differ systematically from those who do not. If a difference exists, it hampers the researcher's ability to infer generalizable findings. In marketing, the most commonly used assessment of nonresponse bias involves comparing early versus late respondents, with the key assumption being that later respondents more nearly match non-respondents (Armstrong and Overton 1977). Unfortunately, it is doubtful that this assumption generally holds (researchers rarely provide a cogent reason to presume that late respondents might have more in common with non-respondents than early respondents), and even if it were true it is often not clear why the variables on which early and late respondents are compared are relevant. Over time, use of the Armstrong and Overton test appears to have become somewhat ritualistic, with researchers looking at differences across readily available variables (e.g., comparison of early versus late respondents on variables that are more or less irrelevant to the constructs being studied, testing for differences on variables that might never be expected to vary over time) rather than focusing on something more meaningful.

Researchers should consider (1) using analysis alternatives that are best suited to assessing the nature of non-response in the specific context of the study, and (2) making use of more than one assessment technique. For example, researchers could compare the characteristics of the respondent group (e.g., demographics) to those of the sampling frame or another known referent group. More sophisticated modeling approaches—such as a Heckman selection model—can also be used to ensure that no significant biases are being introduced into the study as a result of non-response (e.g., Heckman 1979; Winship and Mare 1992). Finally, some researchers suggest conducting limited follow-up surveys with non-respondents using short questionnaires and a different mode of data collection (e.g., phone contact where initial mail requests have not succeeded) to determine the reason for non-response (e.g., Groves 2006).

Across all the papers included in our review, roughly half (n = 99; 49.0%) investigated whether non-respondents differed from respondents in a significant way.[6] The frequency with which various tests were used to assess potential non-response bias is reported in Table 1. The most commonly used test (reported in 62.6% of the papers describing at least one non-response test) compared early versus late respondents, as suggested by Armstrong and Overton (1977). Characteristics of the final sample were compared to those of the sampling frame in 32.3% of the papers that investigated non-response, while some other approach (e.g., comparison of responders and non-responders) was reported in 32.3% of the papers. The majority of papers (72.7%) systematically investigating non-response relied on a single test, while 27.3% used two tests. Our qualitative impression is that the Armstrong and Overton approach often does not represent a meaningful assessment of non-response bias. For example, consider a researcher interested in studying the determinants of a firm's market orientation. Non-response would only be detrimental to the findings of the study if the results for responders yielded a biased picture of the determinants of market orientation. However, it is not clear why a comparison of early and late responders in terms of firm size (for example) should provide relevant information to alleviate concerns about non-response bias.

A special case of unit non-response occurs when the researcher decides to eliminate respondents from the final sample. The increased use of online data collection and the lack of control over the survey setting in this situation have led to concerns that respondents may not pay sufficient attention to the instructions, questions, and response scales (Goodman et al. 2013). Various techniques have been suggested to identify careless responders (Meade and Craig 2012), including the use of so-called instructional manipulation checks (Oppenheimer et al. 2009), where respondents are asked to choose a certain response option (i.e., if they choose the wrong response, they presuably did not pay attention). Because of concerns about the transparency and reproducibility of research findings, some journals now require detailed information about whether data were discarded. In certain situations it is perfectly legitimate to exclude cases (e.g., if they do not meet certain screening criteria, such as sufficient background experience), but it is necessary that researchers provide full disclosure of and a clear justification for any data screening (and possibly a statement about what the results would have been if the respondents had been retained).

Overall, 41.1% of the published studies involved the elimination of at least some respondents. This was done using screens for relevant knowledge, past experience, and other study-specific criteria for inclusion (or exclusion). The reasons for eliminating cases were not always clear, however. Further, in some of the papers cases were eliminated because of missing data. While this may be necessary in extreme cases where information for most measures is absent, in general researchers should try to retain cases where possible by using more advanced missing data techniques to infer appropriate values (e.g., see Graham 2009).

[6] This is close to the number of studies in which an explicit sampling frame was employed, which makes sense (i.e., one would not expect a check for non-response bias when a convenience sample is used).

Table 1 Frequency of use: non-response assessment

Test	Papers using	Proportion using
Armstrong and Overton (early versus late)	62	62.6%
Sample versus sampling frame	32	32.3%
Other techniques	32	32.3%
Armstrong and Overton		
• Alone	39	39.4%
• With another test	23	23.2%

(1) The proportions in the final column are calculated only for the 99 papers that reported some type of non-response assessment

(2) Some papers report use of multiple techniques

Recommendations concerning survey unit representation issues

In the most general sense, consumer researchers should probably concern themselves more with choosing appropriate respondents for their research and de-emphasize ease of access and minimal costs as selection criteria. While managerial researchers tend to follow the traditional survey research script of defining a sampling frame, drawing a sample from the frame, calculating response rates, and investigating non-response bias, this practice becomes a mere ritual when there is no inherent interest in the target population to begin with. There are advantages to having relatively homogenous samples that exhibit sufficient variation on the study constructs (e.g., choosing firms from certain SIC codes), but a pre-occupation with coverage, sampling, and non-response error is only meaningful when generalizability to a specific target population is of interest. It seems more important to discuss why certain objects of measurement and contexts are appropriate for the study and why certain respondents (or, more generally, certain sources of data) are suitable for particular measurements (for good examples, see Arnold et al. 2011; Hughes et al. 2013). Even when generalizability is not of primary concern, the manner in which the sample was drawn has to be described clearly, because this has implications for how the data need to be analyzed (see Bell et al. 2010, and De Jong et al. 2006, for good examples). In addition, researchers should report response rates accurately (e.g., see Hughes et al. 2013; Stock and Zacharias 2011) and reviewers should not equate a low response rate with low survey validity. Finally, if non-response bias is a concern, meaningful tests of selection bias should be reported (not simply a token Armstrong and Overton comparison of early versus late respondents on irrelevant variables). A summary of our recommendations with regard to the representation of survey units is contained in Table 2.

Issues related to the measurement of constructs

Measurement error is intimately related to the design of the measurement instrument used in the survey (although measurement error can also be due either to the respondent and/or how the respondent interacts with the measurement

Table 2 Major recommendations: survey unit representation

Issue	Recommendations
Choice of measurement object and selection of raters	• Describe and justify the object of measurement (this is particulalry important when the unit of analysis is not the individual, but it is relevant even when, say, consumers are studied). • Determine the most appropriate rater for the chosen object of measurement (this is particularly important when the unit of analysis is the firm, but even when the unit of analsis is the individual, self-reports may not necessarily be best).
Description of sampling	• Explicate and justify an explicit sampling frame when the research aims to generalize to a specific target population. • Describe fully the process used to sample from the frame.
Response rate	• Fully disclose all relevant sample sizes on the basis of which the response rate is calculated (i.e., what is counted as a completed survey and what is counted as an eligible reporting unit). • Responding elements that are subsequently eliminated from the final sample should be noted and explained. • Reviewers should refrain from rejecting studies simply because the response rate is low. Low response rates do not necessarily imply nonrepresentative samples.
Assessment of potential non-response bias	• Avoid ritualistic assessments of non-response bias that are uninformative (e.g., an Armstong and Overton comparison based on questionable variables). If selection bias is a concern, compare respondents to the sampling frame or non-respondents on relevant variables that may distort the hypotheses tested. • More than one formal test of non-response should be used whenever possible.

instrument), and the design of the measurement instrument is under the direct control of the researcher. Measurement error, which refers to a difference between the observed response and the "true" response, can be systematic or random (Nunnally 1978), but systematic error is thought to be a particularly serious problem. In contrast to random error, which affects the reliability of measurement, systematic error biases measurements non-randomly and thus introduces systematic distortions. While random errors can be effectively controlled through the use of multiple items to measure each theoretical construct of interest, systematic errors are potentially more serious, yet also more difficult to deal with and generally more poorly handled. Researchers often assume that respondents' answers are only determined by substantive considerations related to the construct of interest (and possibly small random influences); in fact, observed responses are frequently affected by other factors that can introduce systematic errors (such as response styles; see Baumgartner and Steenkamp 2001).

The literature on the psychology of survey response is vast. Entire books have been written on how to design survey instruments, including formulating questions that are readily understood by respondents, using appropriate response scales, and arranging the questions in a questionnaire so that undesired order effects are avoided (e.g., Groves et al. 2004; Schuman and Presser 1981; Sudman et al. 1996; Tourangeau et al. 2000). It is impossible to do justice to these issues within the confines of this article. Instead, we focus on four issues that have been identified as important determinants of survey validity and that we can address empirically using the data from our review of recent survey studies in *JAMS*. These are (1) the use of pretests to ensure that high quality survey instruments are used, (2) the use of multi-item scales with strong psychometric properties to control measure unreliability, (3) the use of research design elements that a priori help to reduce common method bias, and (4) the use of statistical techniques to account post hoc for common method bias.

In making a distinction between these last two topics, we draw on Podsakoff et al. (2012), who distinguish between a priori and post hoc strategies for dealing with common method variance (CMV), suggesting that both are important considerations for survey design and analysis. Our emphasis on CMV reflects the recognition (particularly in the managerial literature) that systematic non-substantive influences on observed responses caused by the method of measurement and the resulting distortion of relationships between variables of interest are among the most serious threats to the validity of empirical findings (Bagozzi and Yi 1990). While the extent of method variance varies widely by study, it can be substantial. In summarizing results from a number of published studies, Podsakoff et al. (2003) note that the relationships between measures tend to be overstated when common method variance is not accounted for (and thus contributes erroneously to the shared variance between constructs). Other studies have determined that CMV accounts for as much as a third of total variance across study contexts (e.g., Cote and Buckley 1987; Doty and Glick 1998; Lance et al. 2010; Ostroff et al. 2002). In fact, Cote and Buckley found the degree of method variance to exceed the amount of trait variance in some survey-based contexts.[7] For these reasons, we focus our discussion particularly on the issue of common method bias because our empirical review suggests that survey researchers continue to struggle with assessing and addressing this potentially serious problem.

Pretests

In order to answer questions accurately, respondents have to understand the literal meaning of a question, but in addition they will usually also make inferences about the implied meaning of a question, based on various cues available in the survey or the survey setting (Schwarz et al. 1998). Several authors have suggested lists of common comprehension problems due to poorly formulated items, including Tourangeau et al. (2000), Graesser et al. (2000, 2006), and Lenzner and colleagues (e.g., Lenzner 2012; Lenzner et al. 2011; Lenzner et al. 2010). Although there is a good deal of advice available to researchers on how to properly formulate questions and, more generally, construct questionnaires, it is frequently difficult even for experts to anticipate all the potential difficulties that may arise during the administration of the survey. It is therefore crucial that surveys be pretested thoroughly before they are launched. Although a variety of pretest methods are available (e.g., informal pretests with small samples of respondents, cognitive interviews, expert panels), researchers frequently fail to pretest their questionnaires sufficiently. When researchers cannot use established scales and have to develop their own, pretesting is particularly critical for developing sound measures.

In our review of survey papers published in *JAMS* over the past decade, we observe that 58.4% (n = 118) reported some level of pretest activity. In many cases the pretests were qualitative interviews (e.g., experts or potential survey takers were asked to comment on the survey instrument) or small-scale quantitative studies, providing initial feedback to the researchers on the quality of the survey design (and possibly the tenability of the hypotheses). This appears to be a healthy level of pretest activity (given the use of established measures and scales by many researchers). However, we encourage *all* researchers to consider using pretests prior to running their main studies. For papers focused on scale development, representing 5% (n = 10) of our review set, and papers in which non-validated measurement scales are used for some constructs, pretesting is essential.

[7] It is interesting to note that Cote and Buckley examined the extent of CMV present in papers published across a variety of disciplines, and found that CMV was lowest for marketing (16%) and highest for the field of education (> 30%). This does not mean, however, that marketers do a consistently good job of accounting for CMV.

Multi-item scales

In some special circumstances a single measure may validly capture a construct without too much random noise. For example, Rossiter (2002) argues that when both the object of measurement (e.g., a specific brand) and the attribute being measured (e.g., liking for a brand) are "concrete," single items are sufficient (see also Bergkvist and Rossiter 2007). Most marketing constructs of theoretical interest are not concrete in this sense, so multiple items are needed to ensure adequate content validity, construct reliability, and convergent validity (Hinkin 1995; Nunnally 1978).[8]

Virtually all of the survey papers reviewed (97.5%) used multiple measurement items to operationalize at least some of the focal (i.e., non-control variable) constructs. In fact, the percentage of focal constructs measured using multiple items was typically over 90% (mean = 91.6%; median = 100%). Thus, in general researchers appear to be taking appropriate care in operationalizing their key constructs with multiple measures.

In line with this extensive use of multi-item measures, most papers (96%) reported at least some degree of measurement analysis, most typically using SEM techniques. Confirmatory factor analysis results were reported for 86.6% of the studies, and exploratory factor analysis results for 22.8%. (For papers not focused on scale development, use of both CFA and EFA is likely unnecessary. In general, use of CFA is preferred.)

It is important to distinguish between reflective and formative measures (or indicators) of constructs (e.g., Diamantopoulos et al. 2008; Hulland 1999; MacKenzie et al. 2005). For reflective indicators, the observed measures are seen as manifestations of the underlying construct; the direction of causality runs from the construct to the indicators. In contrast, for formative indicators the observed measures are defining characteristics of the construct; the direction of causality goes from the indicators to the construct. Jarvis et al. (2003) showed that indicators are frequently misspecified (in particular, researchers routinely specify indicators as reflective when they should be treated as formative), which has various undesirable consequences (e.g., estimates of structural relationships will be biased). For the present purposes, it is important to note that assessing the quality of construct measurement differs substantially between reflective and formative measurement models (see MacKenzie et al. 2011 for an extensive discussion). The usual guidelines for ascertaining reliability and convergent validity (discussed below) apply only to reflective constructs; different criteria have to be fulfilled for formative constructs (because the measurement models are entirely different in their flow of causality). We were not able to explicitly code whether indicators were correctly specified as reflective or formative, and thus cannot be sure that the appropriate measurement model was used in a given study.[9] However, since most of the studies likely used indicators that were reflective, the subsequent discussion focuses on the use of reflective measures.

When researchers conduct a measurement analysis, they typically look at individual item reliability or convergent validity (factor loadings), construct-level reliability or convergent validity, and discriminant validity (Baumgartner and Weijters 2017; Hulland 1999). For item reliabilities, a rule of thumb often employed is to accept items with standardized loadings of 0.7 or larger, which implies that there is more shared variance between the construct and its measure than error variance (e.g., Carmines and Zeller 1979). In practice, researchers may accept some items with loadings below this threshold (especially if the item in question represents an important facet of the construct), but items with loadings below 0.5 should be avoided (Hulland 1999; Nunnally 1978). A summary measure of individual item reliability is provided by average variance extracted (AVE), which represents the average individual item reliability across all measures of a given construct (Fornell and Larcker 1981).

Construct-level reliability or convergent validity is usually assessed using two different measures: Cronbach's coefficient alpha and composite reliability. Both estimate the squared correlation between a construct and an unweighted sum or average of its measures, but the latter is a generalization of coefficient alpha because it does not assume equal loadings across items. For both statistics, a common rule of thumb is that values of 0.7 or higher should be obtained.

The traditional psychometric complement to convergent validity is discriminant validity, representing the extent to which measures of a given construct differ from measures of other constructs in the same model. One test of discriminant validity is whether the correlation between two constructs is less than unity. The preferred way to conduct this test is to check whether the confidence interval around the disattenuated construct correlation does not include one. With sufficiently large samples, this is not a stringent test because even rather high correlations will have a confidence interval excluding one. Another test of discriminant validity, suggested by Fornell and Larcker (1981), requires that for any two constructs, X and Y, the average variance extracted (AVE) for both X and Y should be larger than the shared variance (squared correlation) between X and Y. The test can be conducted using standard output from a confirmatory factor analysis.

A majority of the studies we reviewed reported individual item reliabilities or factor loadings (68.8%), coefficient alpha

[8] In practice, these items need to be conceptually related yet empirically distinct from one another. Using minor variations of the same basic item just to have multiple items does not result in the advantages described here.

[9] In general, the use of PLS (which is usually employed when the measurement model is formative or mixed) was uncommon in our review, so it appears that most studies focused on using reflective measures.

(61.9%), composite reliability (60.9%), AVE (70.3%), and a test of discriminant validity (78.2%).[10] Overall, marketing survey researchers appear to appropriately understand and apply sound measurement design and analysis approaches, including regular use of multi-item scales.

A priori methods for dealing with CMV

CMV is affected by survey measurement procedures, choice of respondents, and survey context (Rindfleisch et al. 2008). As Podsakoff et al. (2003, p. 881) note, "some sources of common method biases result from the fact that the predictor and criterion variables are obtained from the same source or rater, whereas others are produced by the measurement items themselves, the context of the items within the measurement instrument, and/or the context in which the measures are obtained." Because the study context is inextricably connected to the research question addressed by the survey, it is often difficult to "design out" contextual influences (aside from measuring different constructs at different points in time, as in panel studies). In contrast, choice of survey respondents and design of the survey instrument are under the control of the researcher and can have a substantial impact on CMV. The most important respondent strategy is using different respondents (data sources) for different constructs (which also includes using objective data for the dependent variable or validating subjective dependent variables with objective measures). Questionnaire strategies include using cover stories that conceal the true purpose of the study, employing different response scales for different items, and arranging items randomly. Taking these steps prior to the administration of a survey can help to limit the potential for common method bias.

Single respondent and single source biases Traditionally, most surveys in marketing have been cross-sectional in nature, completed by a single respondent at a specific point in time. Jap and Anderson (2004) suggest that this type of research may be especially prone to common method bias. CMV is particularly likely if responses for all constructs are collected from a single respondent, which is often the case. Podsakoff et al. (2003) provide an extensive list of respondent-specific sources of method effects that could potentially induce bias, including need to maintain consistency, social desirability, acquiescence, and mood state. By collecting responses from multiple participants, survey researchers can eliminate (or at least attenuate) these systematic, person-specific effects.

Similarly, use of a single information source to obtain measures for all constructs of interest gives rise to potential measurement context effects. The most common remedy suggested for this problem is the use of two (or more) separate data sources.[11] For example, the researcher might use a cross-sectional survey to measure the predictor variables and use secondary data sources (e.g., archival data) for the outcome measures (Summers 2001; Zinkhan 2006).

Questionnaire/instrument biases If the predictor and criterion variables are measured at the same point in time using the same instrument (e.g., questionnaire), with the items located in close proximity to one another and sharing the same response scale (e.g., a 7-point Likert scale), then systematic covariation that is not due to substantive reasons is quite likely. This practice should be avoided where possible, since it provides a plausible alternative explanation for any significant findings of relationships between constructs.

Where alternative data sources are not available, collection of the predictor and outcome measures can be separated temporally (or at least by physical separation within the questionnaire), by using different response scales, and/or by otherwise altering the circumstances or conditions under which the dependent and independent sets of measures are collected.

Empirical findings Our review of papers published in *JAMS* shows that a minority of researchers employed multiple data sources (34.2%); examples include ratings provided by multiple informants for the same variables, the use of self-reports and secondary data for different constructs, and validation of subjective measures with objective data. As discussed earlier, most variables are based on self or key informant reports (54.5% and 50.0%, respectively), and the use of secondary data (15.8%) or data from other sources (6.4%, e.g., objective performance data) is relatively uncommon. Overall, most researchers relied heavily on single-source primary data from respondents. As might be expected, relatively few studies used separate sources of information for dependent versus independent measures (23.8%), and comparisons of objective versus subjective measures of the same construct were especially rare (4.5%). Subjective and objective measures of constructs (e.g., firm performance) are often assumed to be equally valid, but this assumption of equivalence usually goes untested by the researcher. As Wall et al. (2004) observe, the correlations between subjective and objective performance measures are often not particularly strong. Thus, it is important for researchers to collect—where possible, and

[10] Most of the studies discussing discriminant validity used the approach proposed by Fornell and Larcker (1981). A recent paper by Voorhees et al. (2016) suggests use of two approaches to determining discriminant validity: (1) the Fornell and Larcker test and (2) a new approach proposed by Henseler et al. (2015).

[11] This solution is not a universal panacea. For example, Kammeyer-Mueller et al. (2010) show using simulated data that under some conditions using distinct data sources can distort estimation. Their point, however, is that the researcher must think carefully about this issue and resist using easy one-size-fits-all solutions.

particularly for certain dependent constructs—both subjective and objective measures (e.g., if a researcher is interested in a firm's objective performance and if subjective measures are known to be of questionable validity, then an effort should be made to collect actual performance data, even if this is difficult).

Some researchers described procedural remedies to counter common method bias, such as using different response scales for different constructs or randomizing items or constructs in the questionnaire. However, such safeguards were mentioned infrequently and it is difficult to judge how common they are in practice, as well as how effective they are in eliminating CMV.

Post hoc methods for dealing with CMV

Thus far, we have discussed approaches to controlling CMV through the design of the survey (both the questionnaire and the procedures used for data collection) and choice of respondent. However, such remedies may not always be feasible, and even when they are used their impact may not be sufficient. In such cases, researchers must turn to statistical techniques in order to control CMV (e.g., Richardson et al. 2009; Schaller et al. 2015; Simmering et al. 2015). Various methods for diagnosing and remedying CMV have been proposed, including (1) Harman's single factor test, (2) partial correlation procedures controlling for CMV at the scale level, (3) use of a directly measured or single unmeasured (latent) method factor to account for CMV at the item level, and (4) use of multiple method factors to correct for CMV (Podsakoff et al. 2003).[12] Each of these approaches is discussed more fully below, as all are used (to varying degrees) by marketing survey researchers.[13]

Harman's single-factor test Survey researchers continue to use Harman's test (either on its own or in conjunction with other techniques) to assess whether or not CMV may be a problem. To use the technique, the researcher loads all study variables into an exploratory factor analysis and assesses the number of factors in the unrotated solution necessary to account for most of the variance in the items. The assumption is that if there is substantial CMV present, either a single factor will emerge, or one general factor will account for most of the variance. Variations of this test, which assess the fit of a one-factor model based on a confirmatory factor analysis, have also been proposed.

It is ironic that researchers employing the Harman test often cite Podsakoff et al. (2003) as support for its use. In fact, Podsakoff et al. (2003, p. 889) *explicitly* state the opposite: "Despite the fact that this procedure is widely used, we do *not* believe that it is a useful remedy to deal with the problem" of CMV (emphasis added). For a variety of reasons, it is inappropriate to use the Harman one-factor test to determine the extent of CMV bias (see the Appendix for a more extensive discussion of this issue), particularly when it is the sole reported assessment of CMV.

Partial correlation procedures Over the past decade and a half, it has become increasingly common for survey researchers both in marketing and in other disciplines to use a measure of an assumed source of method variance as a covariate. Three variations on this technique exist. Some researchers include a particular assumed source of method variance (e.g., social desirability, an individual's general affective state) as a covariate. Other researchers employ "marker" variables (i.e., variables expected to be theoretically unrelated to at least one of the focal variables in the study) to control for method variance (Lindell and Whitney 2001; Williams et al. 2010). The assumption is that if there is no theoretical reason to expect a correlation between the marker variable and a substantive construct, any correlation that does exist reflects method variance. A third approach used by researchers is to estimate a general factor score by conducting an exploratory factor analysis of all the variables in a study in order to calculate a scale score for the first (unrotated) factor (as in the Harman test), and then partial out this effect by using the general factor score as a covariate.

Although all of these techniques are easy to implement, they are based on assumptions that do not hold for many survey studies. First, the effectiveness of either an individual difference variable (e.g., social desirability) or a marker variable will depend on how well the chosen variable actually captures CMV in the study (if at all). Second, the general factor score approach assumes that the effect (and only the effect) of CMV can be completely partialled out; in practice, however, this general factor will include both CMV and variance that results from a true, underlying relationship between the constructs, meaning that method and trait variances are once again confounded (Kemery and Dunlap 1986; Podsakoff and Organ 1986). Thus, the use of these partial correlation approaches to account for CMV effects is often not entirely satisfactory.

Use of a directly measured or single unmeasured method factor to account for method variance at the item level These approaches control for method variance at the item level by modeling individual items as loading both on their own theoretical construct and on a single directly measured or unobserved latent method factor. By doing so, these approaches (1) explicitly account for the measurement error in

[12] Podsakoff et al. (2003) also mention two other techniques—the correlated uniqueness model and the direct product model—but do not recommend their use. Only very limited use of either technique has been made in marketing, so we do not discuss them further in this paper.

[13] These techniques are described more extensively in Podsakoff et al. (2003), and contrasted to one another. Figure 1 (p. 898) and Table 4 (p. 891) in their paper are particularly helpful in understanding the differences across approaches.

individual measurement items, (2) allow the method effect to influence the individual measures rather than the underlying constructs of interest, and (3) permit the influence of CMV to vary by measure (rather than assuming that all constructs/measures are affected equally). Researchers using either of these approaches usually estimate two models (one with and the other without the method factor included), and then compare the fit of the two models (i.e., whether the model containing method effects yields a better fit). Furthermore, even if method variance is present, if the substantive conclusions derived from the two models are roughly the same, method variance is said to be unimportant.

For the model with a directly measured latent method factor, the researcher must first identify a method factor that is theoretically expected to influence many of the measurement items in a model (e.g., social desirability or acquiescent response style) and to then directly collect multiple measures for this method factor that can be included in the analysis. The extent to which this approach will be successful clearly depends on the proper choice of a method construct. When the choice of such a construct is not obvious, researchers sometimes use the unmeasured latent method factor approach. The problems with this technique are similar to those discussed in the context of the Harman test (i.e., it is likely that the unmeasured latent method factor confounds substantive and method variance). However, the approach may be useful under certain circumstances (Weijters et al. 2013). For example, if a scale contains both regular and reversed items, a method factor can be defined which has positive loadings for the regular items and negative loadings for the reversed items (after reversed items have been recoded to ensure a consistent coding direction for all items). If a respondent provides similar answers to both regular and reversed items (thus ignoring the items' keying direction), it is unlikely that his or her response behavior reflects substantive considerations.

Use of multiple method factors to control CMV This last approach represents an extension of the previous approach in two ways: (1) two or more method factors are included in the model, and (2) each of the method factors can be hypothesized to affect only a subset of the measures (rather than all of them). This approach is preferred to the extent that more specific sets of method influences can be theoretically anticipated by the researcher. At the same time, this technique creates a more complex model, which can increase sample size demands and lead to estimation problems. Various researchers have recommended use of this approach where feasible (e.g., Bagozzi and Yi 1990; Cote and Buckley 1987; Podsakoff et al. 2003).

Empirical findings A slight minority of the studies we reviewed (92 papers; 45.5%) made some attempt to assess potential systematic bias using one or more of the techniques described above. The number of papers using each of these approaches is reported in Table 3. The most common technique used by researchers—in more than half (56.5%) of the papers examining response bias—was the Harman test, an approach rejected by Podsakoff et al. (2003) and strongly discouraged in this paper. Further (as shown at the bottom of Table 3), in roughly 20% of the papers examining potential response bias, *only* the Harman test was used. As for the other techniques, about a third (33.7%) of the published papers described tests using a marker variable, and 28.3% used an implicit item factor approach. Use of a scale directly measuring a hypothesized source of method bias (e.g. social desirability; 6.5%), use of a general factor scale (9.8%), and use of other techniques (4.1%) were much less frequent.

Recommendations concerning measurement issues

A summary of our recommendations with regard to measurement is provided in Table 4. As noted above, our review suggests that survey researchers in marketing already understand and apply sound measurement design and analysis approaches in their studies. Thus, we do not offer recommendations in this area (nor do we note any "best practice" examples).

Table 3 Frequency of use: systematic bias assessment techniques

Technique	Papers using	Proportion using
Harman one-factor test	52	56.5%
Correction at the construct level using a direct measure of the hypothesized method bias	6	6.5%
Correction at the construct level using a marker variable	31	33.7%
Correction at the construct level using a general method factor	9	9.8%
Correction at the item level using a general method factor	26	28.3%
Other techniques	7	7.6%
Only Harman test used	18	19.6%

(1) The proportions in the last column are calculated only for the 92 papers that reported some type of CMV assessment

(2) Some papers report use of multiple techniques

Table 4 Major recommendations: measurement

Issue	Recommendations
Pretests	• Pretest the survey instrument prior to running all main studies. • For studies focused on scale development and those using non-validated measurement scales for some constructs, pretesting is essential.
Dealing with CMV – a priori methods	• Where possible, use more than a single source of information, whether this involves multiple respondents and/or multiple data sources (e.g., secondary data). • Measures of dependent and independent constructs should be separated from one another, either physically within the questionnaire or temporally. • Both subjective and objective measures of focal constructs (particularly for dependent constructs) should be collected and used whenever possible.
Dealing with CMV – post hoc methods	• Do *not* use the Harman one-factor test. • Measure the source of the bias directly, if possible (e.g., social desirability). • Control for CMV at the individual-item level. • Use implicit control of systematic error (based on a method factor that is inferred from the substantive items themselves) only in special circumstances (e.g., when reversed items are available). • Make use of multiple potential sources of systematic error whenever possible, to allow for triangulation.

While a small majority of the studies we reviewed included pretests in the form of qualitative interviews or small-scale quantitative studies, we recommend that *all* researchers should use pretests prior to running their main studies (for good examples, see Grégoire and Fisher 2008; Song et al. 2007). For studies focused on scale development, as well as those using non-validated measurement scales for some of the studied constructs, pretesting is essential.

Most of the reviewed studies used single-source data, and even when the firm was the unit of analysis and respondents had to report on complex firm-level attributes, single informants were frequently used. The use of multiple data sources, different sources for the independent and dependent variables, and the validation of subjective data with objective evidence is uncommon (for exceptions see Ahearne et al. 2013; Baker et al. 2014; Wei et al. 2014). As a result, opportunities to ensure construct validity through the process of triangulation are lost. Most researchers take some procedural a priori steps to avoid common method bias (e.g., by using different response scales for different constructs), but this is often not sufficient. By dealing more effectively with these issues, researchers can "build in" a priori remedies to common methods bias concerns (Podsakoff et al. 2012).

Following the approach used by Podsakoff et al. (2003, 2012), our review of post hoc attempts to deal with CMV distinguishes between approaches in which (1) there is explicit versus implicit control of systematic error (depending on whether or not the source of the bias can be identified and directly measured), (2) the correction occurs at the scale level or individual item level, and (3) a single source of systematic error or multiple sources are specified (e.g., one or more method factors). For implicit control and correction at the individual item level, either a method factor or correlated uniquenesses can be specified, and for explicit control, measurement error in the method factor may or may not be taken into account. If a researcher believes that the survey might be susceptible to particular response biases, the source of the bias should be measured directly. Controlling systematic errors implicitly is often dangerous and should only be done under special circumstances (e.g., when reversed items are available). In general, it is preferable to control for systematic error at the individual item level, and if necessary researchers should consider multiple error sources. Method factors are usually preferable to correlated uniquenesses, and measurement error in the method factor can be ignored if reliability is adequate.

The Harman technique has been shown to be a severely flawed method that should not be used, yet it continues to be cited by many researchers as evidence of minimal common method bias. Some of the other techniques that are sometimes used by survey researchers (i.e., those in which a method factor is inferred from the substantive items) are not much better, unless substantive variance can be clearly distinguished from method variance (e.g., if the researcher uses both regularly-worded and reverse-worded items, this can be done reasonably well). Baumgartner and Weijters (2017) recommend using dedicated items to measure method variance, but researchers often resist doing so because it increases the length of the questionnaire.

Conclusion

Surveys continue to be an important contributor to marketing knowledge, representing an indispensable source of data, particularly in the managerial marketing area. Similar to other research tools, surveys have distinct strengths, but they must be carefully designed and their data must be appropriately analyzed in order to avoid arriving at invalid conclusions. In this paper we reviewed all survey-based studies reported in

JAMS between 2006 and 2015, and considered two broad sets of issues emerging from our review, with the goal of providing actionable recommendations on how to improve survey practices in marketing.

Overall, marketers have done a good job of adopting best practices in many areas of survey design and administration. For example, our review suggests that researchers generally are proficient in dealing with measurement issues by using multiple items to operationalize constructs, and clearly report on the reliability and validity of those measures. In other areas, there is some room for improvement. For example, researchers should take care to more clearly specify their research objective and to then link this to the choice of the object of measurement and the selection of raters to supply the required information.

A third and final set of issues needs considerably more attention in future work. First, in managing non-response, researchers should take greater care in reporting how their response rates are determined, clarifying all steps taken (e.g., some respondents dropped) to arrive at the final, reported response rate. Second, they need to consider conducting more than a single test of non-response and place less reliance solely and automatically on the approach recommended by Armstrong and Overton (1977). Third, researchers need to do a much better job of dealing with common method variance, both in designing their survey instruments and in accounting analytically for any systematic bias. In particular, use of the widely applied Harman one factor test should be discontinued. Our review of published papers points to this (i.e., poor treatment of common method variance) as the single weakest aspect of contemporary survey work in marketing.

Survey research methodology is widely used in marketing, and it is important for both the field and individual researchers to follow stringent guidelines to ensure that meaningful and valid insights are attained. In this paper, we have reviewed best practices, assessed how well marketing researchers are applying these practices, pointed to papers that do a good job of dealing with specific survey methodology issues, and provided two sets of recommendations for the further improvement of academic survey research.

Acknowledgements The constructive comments of the Editor-in-Chief, Area Editor, and three reviewers are gratefully acknowledged.

Appendix

Putting the Harman test to rest

A moment's reflection will convince most researchers that the following two assumptions about method variance are entirely unrealistic: (1) most of the variation in ratings made in response to items meant to measure substantive constructs is due to method variance, and (2) a single source of method variance is responsible for all of the non-substantive variation in ratings. No empirical evidence exists to support these assumptions. Yet when it comes to testing for the presence of unwanted method variance in data, many researchers suspend disbelief and subscribe to these implausible assumptions. The reason, presumably, is that doing so conveniently satisfies two desiderata. First, testing for method variance has become a sine qua non in certain areas of research (e.g., managerial studies), so it is essential that the research contain some evidence that method variance was evaluated. Second, basing a test of method variance on procedures that are strongly biased against detecting method variance essentially guarantees that no evidence of method variance will ever be found in the data.

Although various procedures have been proposed to examine method variance, the most popular is the so-called Harman one-factor test, which makes both of the foregoing assumptions.[14] While the logic underlying the Harman test is convoluted, it seems to go as follows: If a single factor can account for the correlation among a set of measures, then this is prima facie evidence of common method variance. In contrast, if multiple factors are necessary to account for the correlations, then the data are free of common method variance. Why one factor indicates common method variance and not substantive variance (e.g., several substantive factors that lack discriminant validity), and why several factors indicate multiple substantive factors and not multiple sources of method variance remains unexplained. Although it is true that "if a substantial amount of common method variance is present, either (a) a

[14] It is unclear why the procedure is called the Harman test, because Harman never proposed the test and it is unlikely that he would be pleased to have his name associated with it. Greene and Organ (1973) are sometimes cited as an early application of the Harman test (they specifically mention "Harman's test of the single-factor model," p. 99), but they in turn refer to an article by Brewer et al. (1970), in which Harman's one-factor test is mentioned. Brewer et al. (1970) argued that before testing the partial correlation between two variables controlling for a third variable, researchers should test whether a single-factor model can account for the correlations between the three variables, and they mentioned that one can use "a simple algebraic solution for extraction of a single factor (Harman 1960: 122)." If measurement error is present, three measures of the same underlying factor will not be perfectly correlated, and if a single-factor model is consistent with the data, there is no need to consider a multi-factor model (which is implied by the use of partial correlations). It is clear that the article by Brewer et al. does not say anything about systematic method variance, and although Greene and Organ talk about an "artifact due to measurement error" (p. 99), they do not specifically mention systematic measurement error. Schriesheim (1979), another early application of Harman's test, describes a factor analysis of 14 variables, citing Harman as a general factor-analytic reference, and concludes, "no general factor was apparent, suggesting a lack of substantial method variance to confound the interpretation of results" (p. 350). It appears that Schriesheim was the first to conflate Harman and testing for common method variance, although Harman was only cited as background for deciding how many factors to extract. Several years later, Podsakoff and Organ (1986) described Harman's one-factor test as a post-hoc method to check for the presence of common method variance (pp. 536–537), although they also mention "some problems inherent in its use" (p. 536). In sum, it appears that starting with Schriesheim, the one-factor test was interpreted as a check for the presence of common method variance, although labeling the test Harman's one-factor test seems entirely unjustified.

single factor will emerge from the factor analysis, or (b) one 'general' factor will account for the majority of the covariance in the independent and criterion variables" (Podsakoff and Organ 1986, p. 536), it is a logical fallacy (i.e., affirming the consequent) to argue that the existence of a single common factor (necessarily) implicates common method variance.

Apart from the inherent flaws of the test, several authors have pointed out various other difficulties associated with the Harman test (e.g., see Podsakoff et al. 2003). For example, it is not clear how much of the total variance a general factor has to account for before one can conclude that method variance is a problem. Furthermore, the likelihood that a general factor will account for a large portion of the variance decreases as the number of variables analyzed increases. Finally, the test only diagnoses potential problems with method variance but does not correct for them (e.g., Podsakoff and Organ 1986; Podsakoff et al. 2003). More sophisticated versions of the test have been proposed, which correct some of these shortcoming (e.g., if a confirmatory factor analysis is used, explicit tests of the tenability of a one-factor model are available), but the faulty logic of the test cannot be remedied.

In fact, the most misleading application of the Harman test occurs when the variance accounted for by a general factor is partialled from the observed variables. Since it is likely that the general factor contains not only method variance but also substantive variance, this means that partialling will not only remove common method variance but also substantive variance. Although researchers will most often argue that common method variance is not a problem since partialling a general factor does not materially affect the results, this conclusion is also misleading, because the test is usually conducted in such a way that the desired result is favored. For example, in most cases all loadings on the method factor are restricted to be equal, which makes the questionable assumption that the presumed method factor influences all observed variables equally, even though this assumption is not imposed for the trait loadings.

In summary, the Harman test is entirely non-diagnostic about the presence of common method variance in data. Researchers should stop going through the motions of conducting a Harman test and pretending that they are performing a meaningful investigation of systematic errors of measurement.

References

Ahearne, M., Haumann, T., Kraus, F., & Wieseke, J. (2013). It's a matter of congruence: How interpersonal identification between sales managers and salespersons shapes sales success. *Journal of the Academy of Marketing Science, 41*(6), 625–648.

Armstrong, J. S., & Overton, T. S. (1977). Estimating nonresponse bias in mail surveys. *Journal of Marketing Research, 14*(3), 396–402.

Arnold, T. J., Fang, E. E., & Palmatier, R. W. (2011). The effects of Customer acquisition and retention orientations on a Firm's radical and incremental innovation performance. *Journal of the Academy of Marketing Science, 39*(2), 234–251.

Bagozzi, R. P., & Yi, Y. (1990). Assessing method variance in Multitrait-Multimethod matrices: The case of self-reported affect and perceptions at work. *Journal of Applied Psychology, 75*(5), 547–560.

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., & Kennedy, C. (2010). Research synthesis AAPOR report on online panels. *Public Opinion Quarterly, 74*(4), 711–781.

Baker, T. L., Rapp, A., Meyer, T., & Mullins, R. (2014). The role of Brand Communications on front line service employee beliefs, behaviors, and performance. *Journal of the Academy of Marketing Science, 42*(6), 642–657.

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

Baumgartner, H., & Weijters, B. (2017). Measurement models for marketing constructs. In B. Wierenga & R. van der Lans (Eds.), *Springer Handbook of marketing decision models*. New York: Springer.

Bell, S. J., Mengüç, B., & Widing II, R. E. (2010). Salesperson learning, Organizational learning, and retail store performance. *Journal of the Academy of Marketing Science, 38*(2), 187–201.

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research, 44*(2), 175–184.

Berinsky, A. J. (2008). Survey non-response. In W. Donsbach & M. W. Traugott (Eds.), *The SAGE Handbook of Public Opinion research* (pp. 309–321). Thousand Oaks: SAGE Publications.

Brewer, M. B., Campbell, D. T., & Crano, W. D. (1970). Testing a single-factor model as an alternative to the misuse of partial correlations in hypothesis-testing research. *Sociometry, 33*(1), 1–11.

Carmines, E. G., and Zeller, R.A. (1979). Reliability and validity assessment. *Sage University Paper Series on Quantitative Applications in the Social Sciences*, no. 07-017. Beverly Hills: Sage.

CASRO. (2015). Annual CASRO benchmarking financial survey.

Cote, J. A., & Buckley, M. R. (1987). Estimating trait, method, and error variance: Generalizing across 70 construct validation studies. *Journal of Marketing Research, 24*(3), 315–318.

Curtin, R., Presser, S., & Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly, 69*(1), 87–98.

De Jong, A., De Ruyter, K., & Wetzels, M. (2006). Linking employee confidence to performance: A study of self-managing service teams. *Journal of the Academy of Marketing Science, 34*(4), 576–587.

Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research, 61*(12), 1203–1218.

Doty, D. H., & Glick, W. H. (1998). Common methods bias: Does common methods variance really bias results? *Organizational Research Methods, 1*(4), 374–406.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(3), 39–50.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26*(3), 213–224.

Graesser, A. C., Wiemer-Hastings, K., Kreuz, R., Wiemer-Hastings, P., & Marquis, K. (2000). QUAID: A questionnaire evaluation aid for survey methodologists. *Behavior Research Methods, Instruments, & Computers, 32*(2), 254–262.

Graesser, A. C., Cai, Z., Louwerse, M. M., & Daniel, F. (2006). Question understanding aid (QUAID) a web facility that tests question comprehensibility. *Public Opinion Quarterly, 70*(1), 3–22.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576.

Greene, C. N., & Organ, D. W. (1973). An evaluation of causal models linking the received role with job satisfaction. *Administrative Science Quarterly*, 95-103.

Grégoire, Y., & Fisher, R. J. (2008). Customer betrayal and retaliation: When your best customers become your worst enemies. *Journal of the Academy of Marketing Science, 36*(2), 247–261.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly, 70*(5), 646–675.

Groves, R. M., & Couper, M. P. (2012). *Nonresponse in household interview surveys*. New York: Wiley.

Groves, R. M., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology* (Second ed.). New York: McGraw-Hill.

Harman, H. H. (1960). *Modern factor analysis*. Chicago: University of Chicago Press.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*, 153–161.

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science, 43*(1), 115–135.

Hillygus, D. S., Jackson, N., & Young, M. (2014). Professional respondents in non-probability online panels. In M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective* (pp. 219–237). Chichester: John Wiley & Sons.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management, 21*(5), 967–988.

Huber, J., Kamakura, W., & Mela, C. F. (2014). A topical history of JMR. *Journal of Marketing Research, 51*(1), 84–91.

Hughes, D. E., Le Bon, J., & Rapp, A. (2013). Gaining and leveraging Customer-based competitive intelligence: The pivotal role of social capital and salesperson adaptive selling skills. *Journal of the Academy of Marketing Science, 41*(1), 91–110.

Hulland, J. (1999). Use of partial least squares (PLS) in Strategic Management research: A review of four recent studies. *Strategic Management Journal, 20*(2), 195–204.

Jap, S. D., & Anderson, E. (2004). Challenges and advances in marketing strategy field research. In C. Moorman & D. R. Lehman (Eds.), *Assessing marketing strategy performance* (pp. 269–292). Cambridge: Marketing Science Institute.

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research, 30*(2), 199–218.

Kamakura, W. A. (2001). From the Editor. *Journal of Marketing Research, 38*, 1–2.

Kammeyer-Mueller, J., Steel, P. D., & Rubenstein, A. (2010). The other side of method bias: The perils of distinct source research designs. *Multivariate Behavioral Research, 45*(2), 294–321.

Kemery, E. R., & Dunlap, W. P. (1986). Partialling factor scores does not control method variance: A reply to Podsakoff and Todor. *Journal of Management, 12*(4), 525–530.

Lance, C. E., Dawson, B., Birkelbach, D., & Hoffman, B. J. (2010). Method effects, measurement error, and substantive conclusions. *Organizational Research Methods, 13*(3), 435–455.

Lenzner, T. (2012). Effects of survey question comprehensibility on response quality. *Field Methods, 24*(4), 409–428.

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology, 24*(7), 1003–1020.

Lenzner, T., Kaczmirek, L., & Galesic, M. (2011). Seeing through the eyes of the respondent: An eye-tracking study on survey question comprehension. *International Journal of Public Opinion Research, 23*(3), 361–373.

Lindell, M. K., & Whitney, D. J. (2001). Accounting for common method variance in cross-sectional research designs. *Journal of Applied Psychology, 86*(1), 114–121.

Lohr, S. (1999). *Sampling: Design and analysis*. Pacific Grove: Duxbury Press.

MacKenzie, S. B., Podsakoff, P. M., & Jarvis, C. B. (2005). The problem of measurement model misspecification in Behavioral and Organizational research and some recommended solutions. *Journal of Applied Psychology, 90*(4), 710.

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and Behavioral research: Integrating new and existing techniques. *MIS Quarterly, 35*(2), 293–334.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455.

Nunnally, J. (1978). *Psychometric methods* (Second ed.). New York: McGraw Hill.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*(4), 867–872.

Ostroff, C., Kinicki, A. J., & Clark, M. A. (2002). Substantive and operational issues of response bias across levels of analysis: An example of climate-satisfaction relationships. *Journal of Applied Psychology, 87*(2), 355–368.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision making, 5*(5), 411–419.

Phillips, L. W. (1981). Assessing measurement error in key informant reports: A methodological note on Organizational analysis in marketing. *Journal of Marketing Research, 18*, 395–415.

Podsakoff, P. M., & Organ, D. W. (1986). Self-reports in Organizational research: Problems and prospects. *Journal of Management, 12*(4), 531–544.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in Behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social Science research and recommendations on how to control it. *Annual Review of Psychology, 63*, 539–569.

Richardson, H. A., Simmering, M. J., & Sturman, M. C. (2009). A tale of three perspectives: Examining post hoc statistical techniques for detection and correction of common method variance. *Organizational Research Methods, 12*(4), 762–800.

Rindfleisch, A, & Antia, K. D. (2012). Survey research in B2B marketing: Current challenges and emerging opportunities. In G. L. Lilien, & R. Grewal (Eds.), Handbook of Business-to-Business marketing (pp 699–730). Northampton: Edward Elgar.

Rindfleisch, A., Malter, A. J., Ganesan, S., & Moorman, C. (2008). Cross-sectional versus longitudinal survey research: Concepts, findings, and guidelines. *Journal of Marketing Research, 45*(3), 261–279.

Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing, 19*(4), 305–335.

Schaller, T. K., Patil, A., & Malhotra, N. K. (2015). Alternative techniques for assessing common method variance: An analysis of the theory of planned behavior research. *Organizational Research Methods, 18*(2), 177–206.

Schriesheim, C. A. (1979). The similarity of individual directed and group directed leader behavior descriptions. *Academy of Management Journal., 22*(2), 345–355.

Schuman, H., & Presser, N. (1981). *Questions and answers in attitude surveys*. New York: Academic.

Schwarz, N., Groves, R., & Schuman, H. (1998). Survey methods. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (Vol. 1, 4th ed., pp. 143–179). New York: McGraw Hill.

Simmering, M. J., Fuller, C. M., Richardson, H. A., Ocal, Y., & Atinc, G. M. (2015). Marker variable choice, reporting, and interpretation in the detection of common method variance: A review and demonstration. *Organizational Research Methods, 18*(3), 473–511.

Song, M., Di Benedetto, C. A., & Nason, R. W. (2007). Capabilities and financial performance: The moderating effect of Strategic type. *Journal of the Academy of Marketing Science, 35*(1), 18–34.

Stock, R. M., & Zacharias, N. A. (2011). Patterns and performance outcomes of innovation orientation. *Journal of the Academy of Marketing Science, 39*(6), 870–888.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

Summers, J. O. (2001). Guidelines for conducting research and publishing in marketing: From conceptualization through the review process. *Journal of the Academy of Marketing Science, 29*(4), 405–415.

The American Association for Public Opinion Research. (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (9th ed.) AAPOR.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Voorhees, C. M., Brady, M. K., Calantone, R., & Ramirez, E. (2016). Discriminant validity testing in marketing: An analysis, causes for concern, and proposed remedies. *Journal of the Academy of Marketing Science, 44*(1), 119–134.

Wall, T. D., Michie, J., Patterson, M., Wood, S. J., Sheehan, M., Clegg, C. W., & West, M. (2004). On the validity of subjective measures of company performance. *Personnel Psychology, 57*(1), 95–118.

Wei, Y. S., Samiee, S., & Lee, R. P. (2014). The influence of organic Organizational cultures, market responsiveness, and product strategy on firm performance in an emerging market. *Journal of the Academy of Marketing Science, 42*(1), 49–70.

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods, 18*(3), 320–334.

Weisberg, H. F. (2005). *The Total survey error approach: A guide to the new Science of survey research*. Chicago: Chicago University Press.

Wells, W. D. (1993). Discovery-oriented consumer research. *Journal of Consumer Research, 19*(4), 489–504.

Williams, L. J., Hartman, N., & Cavazotte, F. (2010). Method variance and marker variables: A review and comprehensive CFA marker technique. *Organizational Research Methods, 13*(3), 477–514.

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology, 18*(1), 327–350.

Wittink, D. R. (2004). Journal of marketing research: 2 Ps. *Journal of Marketing Research, 41*(1), 1–6.

Zinkhan, G. M. (2006). From the Editor: Research traditions and patterns in marketing scholarship. *Journal of the Academy of Marketing Science, 34*, 281–283.