An Assessment of Equivalence Between Online and Mail Surveys in Service Research

Elisabeth Deutskens, Ko de Ruyter and Martin Wetzels

Journal of Service Research 2006 8: 346

DOI: 10.1177/1094670506286323

An Assessment of Equivalence Between Online and Mail Surveys in Service Research

Elisabeth Deutskens
Ko de Ruyter
Martin Wetzels
University of Maastricht

This article examines whether online and mail surveys produce convergent results, which would allow them to be used in mixed-mode service quality studies. In the context of a large business-to-business service quality assessment, an analysis of the accuracy and completeness of respondent answers to both open and closed questions suggests that online and mail surveys produce equivalent results. Composite reliability shows consistently high levels for both groups, and the means and variance-covariance matrices are equal across modes. However, minor differences occur between the two survey methods; online respondents provide more improvement suggestions, indicate more often to which competitor they want to switch, and provide lengthier answers in response to requests for examples of positive experiences with the company. This research provides important findings regarding the process for, and results of, comparing two survey modes.

Keywords: *service quality research; online surveys; measurement invariance; response quality*

Both academics and practitioners identify service excellence as a key factor in today's business environment (e.g., Zeithaml and Bitner 2003). To evaluate their services, most companies turn to their customers and collect customer-perceived service quality data on an ongoing basis, but the continuous measurement of service quality is both costly and time-consuming. As a result, firms increasingly make use of online surveys to collect data about service performance.

Online surveys offer great advantages over traditional mail surveys, such as lower costs and faster responses (e.g., Illieva, Baron, and Healey 2002; Schuldt and Totten 1994). However, critics have questioned the quality of responses gathered through online surveys and suggested that the completeness and accuracy of the data they provide may not match those of traditional mail surveys.

In addition, because different survey modes often produce different results (Dillman 2000), comparable survey results are especially important for mixed-mode surveys, in which companies use both online and mail surveys in combination to reduce both their costs and nonresponse rates. Evidence from previous research has verified that

The authors thank Carolyn Massiah for her valuable comments on this article. The Netherlands Organization for Scientific Research (NWO) contributed to the realization of this project with a grant to the first author. Correspondence concerning this article should be addressed to Elisabeth Deutskens, University of Maastricht, Faculty of Economics and Business Administration, Department of Marketing, P.O. Box 616, 6200 MD Maastricht, the Netherlands; phone: +31 (43) 3883851; fax: +31 (43) 3884918; e-mail:ec.deutskens@mw.unimaas.nl.

Journal of Service Research, Volume 8, No. 4, May 2006 346-355
DOI: 10.1177/1094670506286323

online and telephone surveys may exhibit different underlying factor structures (Roster et al. 2004). Other studies indicate that online surveys produce more socially desirable answers, more extreme responses, higher item completion, higher item variability, and higher measurement errors (Klassen and Jacobs 2001; Shermis and Lombard 1999; Stanton 1998). In contrast, still other studies have found no significant differences between the online and mail methods (Epstein et al. 2001; Knapp and Kirk 2003).

Despite these findings, existing evidence regarding the comparison of Web-based and paper-and-pencil surveys remains both scarce and inconclusive. Assuming their equivalence still must be considered risky, because most of these studies have methodological limitations and conducted only limited statistical comparisons.

This article contributes to the extant service research literature in several important ways. Because of the inevitable time and resource constraints involved in service quality survey research, online surveys represent attractive alternatives or supplements to traditional mail surveys. However, mixed-mode surveys can serve as a trustworthy basis for decision making only if convergent validity between the two survey methods can be established. Therefore, we empirically examine the response quality and measurement invariance of Internet-based and traditional mail surveys in a service quality study. In addition, we attempt to align the diverse literature on the comparison between online and offline surveys by indicating how different aspects of response completeness and accuracy might be analyzed in the future. Furthermore, because the existing literature on measuring the response quality of surveys is sparse, we use a service quality survey from a large office equipment manufacturer to show how the measurement invariance (Jöreskog 1971; Vandenberg and Lance 2000) and response quality (Goetz, Tyler, and Lomax Cook 1984; Hansen 1980; McDaniel and Rao 1980) of online and mail surveys can be analyzed and compared. Also, we provide managerial guidelines on the combined use of both online and mail surveys that will enable practitioners to optimize their response rates, response times, and survey costs while maintaining high response quality. Finally, this study addresses the critical issue of obtaining high-quality, representative online samples in service research.

The remainder of the article is structured as follows: We begin with a review of the existing empirical evidence on the equivalence of online and mail surveys. Next, we assess the equivalence of on- and offline surveys in a service quality context. Finally, we conclude with a discussion of the results and their theoretical and practical implications.

LITERATURE REVIEW

Mixed-mode surveys combine two survey modes, such as online and mail surveys, to compensate for the weaknesses of each. For example, various persuasive arguments recommend online surveys over traditional methods because of their lower costs, faster response, and wider geographic reach (e.g., Green, Johnson, and Neal 2003; Illieva, Baron, and Healey 2002; Schuldt and Totten 1994). Furthermore, proponents of online surveys argue that the Internet provides uncomplicated directions (e.g., automatic routing), as well as richer and more interesting question formats (Klassen and Jacobs 2001; Simsek and Veiga 2001). However, in terms of coverage, the potential for invalid addresses, and representativeness, mail surveys are still preferred. Table 1 provides an overview of the strengths and weaknesses of both methods.

A mixed-mode design thus can reduce overall costs while maximizing response rates and minimizing nonresponse. But if online and mail surveys produce different results for the same study, their data cannot be aggregated. In Table 2, we provide an overview of existing empirical evidence regarding the difference between, or equivalence of, online and mail surveys, which demonstrates that this evidence is very inconclusive. Whereas several studies find differences—including factor structures, socially desirable answers, more extreme responses, higher item completion, higher item variability, and higher measurement error in online surveys (Klassen and Jacobs 2001; Shermis and Lombard 1999; Stanton 1998)—others indicate that online and mail surveys are equivalent (Epstein et al. 2001; Knapp and Kirk 2003). These conflicting results may be due to, for example, differences in the response processes for online and mail surveys. Because online respondents, for the most part, cannot scan, preview, review, skip, or change items, they may experience a different level of self-generated validity, which refers to the phenomenon in which responses to previous survey questions can affect answers to subsequent questions in the survey (Feldman and Lynch 1988). Moreover, computer anxiety might affect participants' responses (Buchanan and Smith 1999), or biases could occur in the way people perceive questions on screen versus on paper. Also, different screen formats and otherwise inconsistent computer administration, as well as technical or interface problems, can elicit different responses in an online survey. Finally, the print quality of paper-and-pencil surveys is still better than that of onscreen surveys, which could result in lower comprehension, less patience, and decreased reading speed on the computer (Nielsen 2000).

TABLE 1
Comparison of Mail and Online Surveys

	Mail	*Online*
Coverage	High	Low
Speed	Low	High
Control of data collection environment	Low	Low
Response rates	Low	Low
Flexibility of data collection	Low	Moderate to high
Wrong addresses	Low	High
Labor needed	High	Low
Expertise to construct	Low	Medium to high
Costs		
Invitations	Letterhead; envelopes; postage; personnel to generate, fold, stuff, and mail packets	Personnel to generate e-mails
Data entry of results	Hand data entry or scanning form Quality control checks and reentry required Error rate in hand data entry	Data entered by respondents Zero data entry errors
Data handling	Manual processing of mailed surveys response forms Results not available until final data entry and analysis is performed	Data resides on server. Researcher can track results as responses are acquired.
Reminders	Letterhead; envelopes; postage; personnel to generate, fold, stuff, and mail packets	Personnel to generate e-mails

SOURCE: Adapted from Cobanoglu, Warde, and Moreo (2001); Malhotra (1999); and Online Survey Services (http://www.onlinesurveyservices.com/olss/cost.htm).

In addition to resulting in these inconclusive results, previous studies that compare online and mail surveys have methodological limitations. For example, in many cases, the results are based on small samples, target populations with strong technology familiarity (e.g., student or academic samples), or self-selected convenience samples. Knapp and Kirk (2003) used a different recruitment method; instead of sending an e-mail with the hyperlink to the Web questionnaire, they handed out an envelope containing the URL to respondents. This method requires more effort from the respondent and is more time-consuming and complicated. Their survey also suffered from technical problems; the questionnaire was offline for 27 hours (Knapp and Kirk 2003). In addition, the choice of environment often acts as a limitation in previous studies on this topic. Epstein et al. (2001), for example, chose to conduct their study in a highly controlled environment. Furthermore, most studies are conducted in research areas other than marketing, such as psychology or public opinion research (e.g., Buchanan and Smith 1999; Stanton 1998). Finally, some studies provide contradictory results and are limited in both their quantity and methodological quality. Many articles compare the means of the online and offline groups, but because these means are just approximations of the underlying true population means, it is impossible to say which is better and how much importance should be attached to small but significant differences.

Despite the widespread evidence of nonequivalence, we hypothesize that data collected through online and mail surveys are equivalent. First, more recent studies have found equivalence between the two methods (Epstein et al. 2001; Knapp and Kirk 2003), which may indicate that factors such as computer anxiety or privacy concerns have been reduced as people become more familiar with the Internet. Second, over time, respondents gain increased experience with online surveys, which makes it less likely that the response process or the way people perceive questions on a screen versus on paper will cause differences. Third, although online communication in general has been found to be more open, there should be no difference in perceived anonymity between online and mail questionnaires because both are filled out in the absence of an interviewer. Fourth, the number of online panel members or e-mail addresses available in databases continues to increase steadily, ameliorating the lower coverage problems. Fifth, recommendations included in the growing body of literature on the best design of online surveys (e.g., Couper, Traugott, and Lamias 2001; Schaefer and Dillman

TABLE 2
Overview of Existing Literature That Compares Mail and Online Surveys

Author(s)	Sample	Survey Topic	Survey Modes	Response Rate: Sample Size	Response Rate: n	Response Rate: %	Method Testing Equivalence	Findings
Stanton (1998)	Professional employees	Employees' perceptions of fairness in their day-to-day interactions with their supervisors	Mail Web	NA NA	181 50	~50 —	ANOVA, *t* test, SEM	Differences in missing values, reported level of fairness, measurement errors. Comparable in terms of item variability
Buchanan and Smith (1999)	Mainly students	Self-monitoring scale	Paper-and-pencil Web	NA NA	224 963	— —	CFA	Comparable model fits, similar psychometric properties
Shermis and Lombard (1999)	Organization members	Assessment of electronic media	Mail E-mail	585 585	211 176	36 30	Chi-square, multiple Mann-Whitney U test, *t* test, ANOVA	Differences in screening questions, item response rate, open-ended questions
Cobanoglu, Warde, and Moreo (2001)	Hospitality professors	Hospitality education	Mail Fax Web	99 100 95	26 17 42	26 17 44	ANOVA	Equivalence
Epstein et al. (2001)	Students	Physical and sexual attractiveness of men and women	Paper-and-pencil Web	NA NA	116 97	55 46	Equivalence criterion: 20% deviation of mean score of Web from paper-and-pencil	Equivalence in mean ratings of physical and sexual attractiveness Difference in rating of same-gender targets
Klassen and Jacobs (2001)	Senior production or marketing professionals from a broad sample of firms	Forecasting practices	Mail Fax Disk-by-mail Web	163 161 160 110	38 32 33 15	23 20 21 14	Regression	Differences in item completion rate. Limited evidence of systematic bias in use of forecasting methods. No difference in reporting of firm-level characteristics
Truell, Bartlett, and Alexander (2002)	Business education professionals	A dummy instrument called "Educator's survey"	Mail Web	153 153	81 78	51 49	*t* test	Differences in response completeness
Grandcolas, Rettie, and Marusenko (2003)	Students	Internet and market research	Paper-and-pencil Web	— 4,000	123 157	— 4	Chi-square, *t* test, regression	For some items differences in means, skewness, and kurtosis, variance, and extremity of responses
Griffis, Goldsby, and Cooper (2003)	Middle- and senior-level managers from manufacturers and merchandisers	Logistics performance measurement	Mail Web	585 1,776	59 254	10 14	*t* test	Slight differences in strength with which measurement items gained support, but they were attributed to the statistical power. Consistency in conclusions.
Knapp and Kirk (2003)	Students	Mischievousness, general honesty, academic honesty, interpersonal relationship, prejudice, illegal behavior, alcohol, substance use, violence, health, sex, miscellaneous	Mail Touch-tone Web	NA NA NA	174 121 57	— — —	Chi-square	Great variability in responses, but no significant difference with regard to major questions
McDonald and Adam (2003)	Members of Australian football club	Satisfaction of members	Mail Web	1,024 3,200	471 826	46 26	Frequencies, chi-square	Difference in missing values for demographic but not for content questions. Differences in variation of responses as well as mean responses. Differences in demographics.
Roster et al. (2004)	Adults in metropolitan area in southwest	Corporate/company reputation	Telephone Web	2,173 974	251 272	11.5 326	Chi-square, *t* test	Differences in factor structures, mean agreement scores, nonresponse, reliability
Smither, Walker, and Yap (2004)	Managers and supervisors of a financial service company	Upward feedback ratings	Opscan Web (Intranet)	2,739 2,518	— —	—	Principal axis factor analysis, *t* test, regression	Differences in rater and ratee characteristics. No differences in number of missing responses, factor structures, or means

NOTE: ANOVA = analysis of variance; SEM = structural equation modeling; CFA = confirmatory factor analysis.

1998) also has reduced problems because of different screen formats or other technical or interface problems. Therefore, we hypothesize that online and mail surveys are equivalent.

AN EMPIRICAL STUDY

To test the equivalence between mail and online surveys, we conducted a service quality survey with a major multinational office equipment manufacturer. The data were collected in the United States, which has one of the highest Internet acceptance rates worldwide (NUA Internet Surveys 2002). Customers in the traditional paper-and-pencil group received a mailing that contained an introduction letter, the questionnaire, and a prepaid return envelope. Customers in the online group received an e-mail invitation, including a short introduction to the study with a request to participate and the hyperlink to the Web questionnaire. We avoided double entries by using a unique, eight-digit identification code for each respondent.

The items used to assess equivalence between the survey modes measured service call quality, service visit quality, and the intention to use the services of this provider again (see Table 3). The items were strongly driven by the SERVQUAL dimensions developed by Parasuraman, Zeithaml, and Berry (1988) and have been used in previous studies on service contact modes (Van Birgelen et al. 2002). All items were measured on a 9-point scale that ranged from 1 (*much worse than expected*) to 9 (*much better than expected*) for service call quality and service visit quality and from 1 (*very unlikely*) to 9 (*very likely*) for intentions. In addition, the questionnaire contained six open-ended questions about the nature of the respondents' complaints, additional information the respondents would like to obtain, desired improvements, positive and negative experiences, and their intention to switch to a competitor.

Participants for both the online and offline groups were recruited from the manufacturer's customer database. A stratified sampling procedure (survey mode, business units, regions, product type) was used, in which we first divided the customer database into different business units (standard and customized specialty systems), then into different regions (central, east, southwest, west), product types (low-, medium-, high-volume machines), and finally survey modes (e-mail address is/is not available). Then, we drew a random sample from each group to ensure that we obtained a valid and representative sample of customers.

For the mail survey, we received 694 responses, which represents a response rate of 16.58%. In the online survey, 255 customers participated (response rate 28.47%). The smaller sample for the online survey reflects the common problem that customer databases do not yet contain all customers' e-mail addresses. To rule out that potential differences in the response numbers or rates could be due to different sample characteristics, we compared the online and mail sample on several important background variables. We did not find significant differences with respect to region ($p = .405$), position in the company ($p = .133$), decision power ($p = .126$), preferred mode of communication ($p = .182$), or whether they recently had bought new products ($p = .224$).

ANALYSES AND RESULTS

To assess equivalence, we considered two specific components of data quality, namely, the completeness and accuracy of respondents' answers (Goetz, Tyler, and Lomax Cook 1984; Hansen 1980; McDaniel and Rao 1980). Completeness was assessed by comparing the number of respondents who provided answers to open-ended questions and considering the length of answers. To assess accuracy and bias differences between the online and mail surveys, we followed Hansen's (1980) suggestion to compare the distribution, or summary, of responses (means and variances) from one subgroup with the summary of another subgroup. Furthermore, we used the more rigorous, powerful, and versatile multigroup confirmatory factor analysis (CFA) approach to assess measurement invariance, which basically determines whether different survey settings produce different measures of the same attribute (Steenkamp and Baumgartner 1998).

The majority of structural equation modeling (SEM) applications in the behavioral sciences employ the maximum likelihood (ML) estimation procedure to provide parameter estimates for the hypothesized models (e.g., Bollen 1989). However, the ML estimator will exhibit desirable statistical properties (unbiased, consistent, asymptotically efficient, and approximating a χ^2 distribution) only if several important assumptions are met (Bollen 1989; West, Finch, and Curran 1995). Chief among these is the assumption that the manifest variables follow a multivariate normal distribution. However, as in most customer satisfaction and service quality research (e.g., Brown, Churchill, and Peter 1993), this assumption frequently does not hold in behavioral research.

Similarly, in our study there were significant deviations from multivariate normality in both the online and mail survey data sets, according to Mardia's (1970) and Srivastava's (1984) tests of multivariate skewness and kurtosis. When a data set deviates significantly from multivariate normality, the chi-square statistic is inappro-

TABLE 3
Measurement Items and Scale Reliabilities

Variable		*Offline*	*Online*	*Item*		*Offline*	*Online*
Service call quality	CR:	.95	.96	Competence of the [. . .] telephone support team.[a]	*M*	5.69	5.65
	AVE:	.87	.88		*SD*	2.09	2.18
				Feedback on when your software-related problem is being resolved.[a]	*M*	5.45	5.49
					SD	2.19	2.30
				Understanding of your needs by the [. . .] call-handling agent.	*M*	6.38	6.40
					SD	1.77	1.94
				Speed of response by the [. . .] call-handling agent.	*M*	6.38	6.43
					SD	1.79	1.87
				Competence of the [. . .] call-handling agent.	*M*	6.43	6.48
					SD	1.78	1.84
Service visit	CR:	.93	.93	Ability of the [. . .] service technician to solve your problem in one visit.[a]	*M*	6.59	6.57
	AVE:	.82	.82		*SD*	2.07	2.09
				The degree to which the [. . .] service technician provides feedback on the progress of the service visit.	*M*	6.71	6.68
					SD	1.94	1.98
				Competence of the [. . .] service technician.[a]	*M*	7.04	7.04
					SD	1.86	1.86
				Understanding of your needs by the [. . .] service technician.	*M*	6.97	7.01
					SD	1.85	1.85
				Amount of time it takes the [. . .] service technician to repair your equipment.	*M*	6.56	6.57
					SD	1.92	1.96
Intentions	CR:	.95	.96	I recommend [. . .] to someone who seeks my advice about [. . .].	*M*	6.49	6.50
	AVE:	.84	.84		*SD*	2.01	2.16
				I encourage associates, friends, and relatives to do business with [. . .].	*M*	6.19	6.31
					SD	2.09	2.23
				I intend to do more business with [. . .] in the next few years.	*M*	6.10	6.19
					SD	2.22	2.20
				I consider [. . .] to be my first choice for [. . .].	*M*	6.03	6.09
					SD	2.21	2.28

NOTE: CR = composite reliability; AVE = average variance extracted.
a. This item was eliminated from the analysis.

priate for assessing the fit of the CFA models, because it leads to an underestimation of the standard errors of the estimates (West, Finch, and Curran 1995). Several alternatives to ML estimation have been proposed for data sets that are not multivariate normal (cf. West, Finch, and Curran 1995), including the asymptotically distribution-free (ADF) estimator (Browne 1984) and the Satorra-Bentler scaled chi-square statistic (χ^2_{SB}) with robust standard errors (Satorra and Bentler 1994). Monte Carlo simulation studies have found that the statistic with robust standard errors outperforms the ADF estimator, especially when sample sizes are small (e.g., Chou, Bentler, and Satorra 1991). Therefore, to check for equivalence of the data sets between the online and mail surveys in our study, we employed a difference test for the statistic, as recommended by Satorra and Bentler(2001).[1]

We used EQS 6.1 to analyze the data. We first specified a baseline CFA model for both the online and offline samples that contained 14 items. The fit indices were modest for both the offline data, $\chi_{SB}(74) = 283.700$, Normed Fit Index (NFI) = .917, Confirmatory Fit Index (CFI) = .937, root mean square error of approximation (RMSEA) = .099, and the online data, $\chi_{SB}(74) = 180.448$, NFI = .881, CFI = .925, RMSEA = .122. On the basis of the pattern of the standardized residuals and the modification indexes (Langrangian multiplier tests), we decided to remove four items from the analyses (two items measuring service call quality and two measuring service visit). This reduction resulted in a significantly better fit for both groups: offline $\chi^2_{SB}(32) = 77.637$, NFI = .984, CFI = .991, RMSEA = .045, and online $\chi^2_{SB}(32) = 84.921$, NFI = .959, CFI = .974, RMSEA = .081.

We evaluated the reliability of our constructs through composite scale reliability and average variance extracted (e.g., Fornell and Larcker 1981) (see Table 3). The composite scale reliability ranged between .96 and .98; thus, all values exceeded the cutoff point of .7 suggested by Nunnally and Bernstein (1994). The average variance extracted ranged from .89 to .93, so all constructs exceeded the .5 cutoff value proposed by Fornell and Larcker

(1981). In addition, we assessed discriminant validity with a Satorra-Bentler difference test (Satorra and Bentler 2001) and found discriminant validity for all constructs at $p < .001$.

Recall that to assess completeness, we compared the number of respondents who provided answers to open-ended questions and the length of the answers. For four of the six open-ended questions, there were no significant differences between the online and offline samples in terms of how many people responded ($p = .786, .864, .319, .562$). However, for the remaining two questions, we found that the online sample provided significantly more improvement suggestions (34% versus 26.7%, $p = .008$) and indicated whether and to which competitor the respondents wanted to switch (8% versus 4.7%, $p = .019$). With respect to the number of words, there again were no significant differences for four open-ended questions ($p = .525, .298, .233$, and $.640$). However, online respondents provided significantly lengthier improvement suggestions (mean number of words 8.16 versus 6.93, $p = .006$) and examples of positive experiences with the company (mean number of words 11.06 versus 8.21, $p = .000$).

The second aspect of response quality pertains to bias or inaccuracy (Goetz, Tyler, and Lomax Cook 1984). Our results show that there are no significant differences in the means between online and offline groups, $\chi^2_{SB}(10) = 10.40$, $p = .406$, RMSEA < .001. Testing for the differences in the variance-covariance matrices, $\chi^2_{SB}(55) = 52.858$, $p = .557$, RMSEA < .001, CFI = 1.000; Tucker-Lewis Index (TLI) = 1.001, and simultaneously for the means and the variance-covariance matrices, $\chi^2_{SB}(65) = 62.453$, $p = .567$, RMSEA < .001, CFI = 1.000, TLI = 1.000, also resulted in a good model fit. Because the covariances and means were invariant across the survey modes, we could pool the data from the online and offline surveys, which means that additional analyses to test for configural, metric, scalar, and factor covariance, factor variance, and error variance invariance were unnecessary (Steenkamp and Baumgartner 1998; Vandenberg 2002; Vandenberg and Lance 2000).[2]

Therefore, we concluded that we could not identify any differences in the means and variance-covariance matrices with respect to the evaluations of service call quality, service visit quality, or intentions in the online and offline surveys. In four of the six open-ended questions, there also were no differences in the number of respondents who provided answers or in the length of their answers. Only for improvement suggestions and switching intentions did the online surveys provide slightly more responses. Online respondents also provided significantly lengthier improvement suggestions and examples of positive experiences with the company.

DISCUSSION

ESOMAR, the World Association of Research Professionals, estimates that in 2004, 35% of market research in the United States was conducted through online surveys (ESOMAR 2004). More and more companies now use online surveys, with their significantly lower costs, to assess service quality continuously. In the long term, however, companies only profit from the cost savings of online surveys if their response quality is similar to that of mail surveys, so that outcomes from mixed-mode surveys provide a reliable basis for managers' decision making.

The selection of a data collection technique generally is based on four criteria: response rate, response bias, costs, and completion time (e.g., Malhotra 1999). In terms of the response rate, the online survey has a higher response rate than the mail survey (28.47% versus 16.58%). Even though the online sample initially was smaller, the higher response rate suggests that online surveys are preferable for companies attempting to contact busy, hard-to-reach professionals.

With respect to response characteristics, the online and mail samples produce virtually identical results. Both the composite reliability and the average variance extracted show consistently high levels for both groups. The number and length of responses to open-ended questions also demonstrates similar results for the online and mail surveys. In terms of accuracy, we find that the means and the variance-covariance matrices are equal across modes. Thus, the results from online and offline surveys are comparable and produce equally usable data. The only minor differences are that online respondents provide more improvement suggestions, indicate more often to which competitor they want to switch, and provide lengthier answers about their positive experiences with the company and improvement suggestions. A possible explanation for this outspokenness by online respondents may be the reduced social context information on the Internet. Persons often are sensitive to the variables that can influence the content of a conversation and inhibit or facilitate what is said, how, and by whom. Empirical evidence from e-mail communications (Sproull and Kiesler 1986) suggests that social context information in online surveys is weak, which increases the respondents' perceived anonymity and produces relatively self-centered and unregulated behavior. Respondent behavior thus is likely to be more extreme, more impulsive, and less socially differentiated, because these people are relatively unconcerned about making a good appearance (Sproull and Kiesler 1986).

Unfortunately, we did not record information about data collection costs and response times, the third and fourth criteria. However, several studies and meta-

analyses suggest that online surveys are faster and cheaper (e.g., Illieva, Baron, and Healey 2002). Our findings of equivalence between online and mail surveys support the use of online surveys, whose response quality is comparable to that of mail surveys, whereas their response time is faster and their costs are lower. Equivalence between online and mail surveys is extremely compelling for companies that rely on continuous measurements of service quality. For example, decision-support tools use continuous input from customers to predict future behavior and revenues. Because customer input enables predictions about crucial factors such as future revenues, it is absolutely necessary that surveys produce timely and reliable results at a low cost, as can be achieved through online surveys.

However, even though our study supports the equivalence of online and mail surveys, we recognize that the adequacy of online surveys depends mainly on the strength of the online sample. In this study, our online and mail samples were comparable; thus, we could focus on examining whether there were any systematic differences in response quality or measurement invariance. We did not find any such differences, which is a promising result. However, differences between online and mail surveys might occur because of incomparable samples. For online surveys, we recommend that users carefully examine the sample frame before the data are collected. Because no post hoc comparisons are possible, the reliability and validity of a pure online survey stand and fall with the representativeness of the sample. In a business-to-business (B2B) context, many companies maintain a database from which they can draw a random sample, but most business-to-consumer (B2C) companies must rely on e-mail addresses obtained from online panels. Although most online panels are very large and employ sophisticated weighting techniques, they may not be able to replicate results from more traditional research methods (Sparrow and Curtice 2004). Furthermore, empirical evidence suggests that no simple weighting factor or adjustment strategy can make on- and offline samples comparable (e.g., Vehovar, Lozar Manfreda, and Batagelj 1999).

If only a limited number of e-mail addresses is available, mixed-mode surveys might be a viable alternative to contact those members for whom an e-mail address is available online while surveying others with a regular mail survey (Dillman 2000). Using mixed-mode surveys can increase both response rates and the representativeness of the sample and maintain the online survey benefits of faster response times and lower costs. We demonstrate how the strength of the online sample can be assessed through a comparison, with the mail survey, of relevant sample characteristics. Thus, the response quality of, and equivalence between, online and mail surveys can be tested easily by comparing (a) their completeness and accuracy according to the number and length of open-ended questions, (b) the distribution of responses by considering the means and variances, and (c) measurement invariance.

LIMITATIONS AND FUTURE RESEARCH GUIDELINES

Our study was conducted within only one country, the United States. Additional studies should examine whether our findings hold true for other countries, especially those with lower rates of Internet adoption. Moreover, we compared online and mail surveys because the type of research for which mail surveys are used is most likely to be supplemented by online surveys. However, it might be interesting to examine how alternative modes, such as telephone surveys, compare to online surveys. Further studies could also examine the impact of any potential moderating factors on the results.

In addition, the quality of online samples must receive further attention, although we did not find systematic biases between online and mail surveys. Also, the customers in our study are not end consumers but rather business customers; the findings for end customers might differ from those presented herein. According to the company, the small number of e-mail addresses in the database is due to erratic updates by the sales force, not a systematic bias in, for example, the innovativeness of those customers for whom an e-mail address is available. However, a selection bias cannot be ruled out completely because the availability of e-mail addresses might contain an element of self-selection.

The quality of online studies depends on the quality of the online sample. In general, the best way to secure representative results is by maximizing response rates (Hansen 1980) to reduce nonresponse error. However, this conventional wisdom may not be true for online panels, in which many respondents are members of several panels or participate because of the money they can win (e.g., "The Free Get Paid to Take Surveys Online Guide;" http://www.surveys4money.com/). A large response rate also could indicate that more highly motivated, survey-prone, or professional respondents have participated. Even if the demographic profile is representative, it is reasonable to assume that the responses are not in such a case. Insights into these issues will advance knowledge about the quality of online surveys and thereby help empirically determine the potential of Internet-based research.

NOTES

1. The difference test for the Satorra-Bentler scaled statistic can be implemented as follows:

$$\Delta\chi^2_{SB} = \frac{\Delta\chi^2}{c_d},$$

where

$$\Delta\chi^2 = \chi^2_1 - \chi^2_2,$$

and

$$c_d = \frac{df_1 c_1 - df_2 c_2}{df_1 - df_2},$$

where

$$c_1 = \frac{\chi^2_1}{\chi^2_{SB,1}},$$

and

$$c_2 = \frac{\chi^2_2}{\chi^2_{SB,2}}.$$

2. To rule out that the underlying factor structures might be different, we assessed measurement invariance following the procedure recommended by Vandenberg and Lance (2000). To examine invariance—starting with the analysis of configural, then metric, scalar, and factor covariance, and finally factor variance invariance—we tested increasingly restrictive hypotheses. In line with Steenkamp and Baumgartner (1998), we also tested for error variance invariance. Our results indicate that the data fit well and that the Satorra-Bentler difference test (Satorra and Bentler 2001) was not significant ($p \geq .124$). For more information on testing measurement invariance, see Steenkamp and Baumgartner (1998) and Vandenberg and Lance (2000).

REFERENCES

Bollen, Kenneth A. (1989), *Structural Equations with Latent Variables*. New York: John Wiley.

Brown, Tom J., Gilbert A. Churchill Jr., and J. Paul Peter (1993), "Improving the Measurement of Service Quality," *Journal of Retailing*, 69 (1), 127-39.

Browne, Michael W. (1984), "Asymptotically Distribution-Free Methods for the Analysis of Covariance Structures," *British Journal of Mathematical and Statistical Psychology*, 37, 62-83.

Buchanan, Tom and John L. Smith (1999), "Using the Internet for Psychological Research: Personality Testing on the World Wide Web," *British Journal of Psychology*, 90 (1), 125-44.

Chou, Chih-Ping, Peter M. Bentler, and Albert Satorra (1991), "Scaled Test Statistics and Robust Standard Errors for Nonnormal Data in Covariance Structure Analysis: A Monte Carlo Study," *British Journal of Mathematical and Statistical Psychology*, 44, 347-57.

Cobanoglu, Cihan, Bill Warde, and Patrick J. Moreo (2001), "A Comparison of Mail, Fax and Web-based Survey Methods," *International Journal of Market Research*, 43 (4), 441-52.

Couper, Mick P., Michael W. Traugott, and Mark J. Lamias (2001), "Web Survey Design and Administration," *Public Opinion Quarterly*, 65 (2), 230-53.

Dillman, Don A. (2000), *Mail and Internet Surveys, the Tailored Design Method*, 2nd ed. New York: John Wiley.

Epstein, Joel, W. Dean Klinkenberg, D. Wiley, and L. McKinley (2001), "Insuring Sample Equivalence across Internet and Paper-and-Pencil Assessments," *Computers in Human Behavior*, 17 (3), 339-46.

ESOMAR (2004), "Marketing and Opinion Research Industry in the USA Is the Largest Market Worldwide." Retrieved October 14, 2004, from http://www.esomar.org/esomar/show/id=107403.

Feldman, Jack M. and John G. Lynch Jr. (1988), "Self-Generated Validity and Other Effects of Measurement of Belief, Attitude, Intention, and Behavior," *Journal of Applied Psychology*, 73 (3), 421-35.

Fornell, Claes and David F. Larcker (1981), "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research*, 18 (3), 39-50.

Goetz, Edward G., Tom R. Tyler, and Fay Lomax Cook (1984), "Promised Incentives in Media Research: A Look at Data Quality, Sample Representativeness, and Response Rate," *Journal of Marketing Research*, 21 (2), 148-54.

Grandcolas, Ursula, Ruth Rettie, and Kira Marusenko (2003), "Web Survey Bias: Sample or Mode Effect?" *Journal of Marketing Management*, 19 (5/6), 541-61.

Green, Paul E., Richard M. Johnson, and William D. Neal (2003), "The *Journal of Marketing Research*: Its Initiation, Growth, and Knowledge Dissemination," *Journal of Marketing Research*, 40 (1), 1-9.

Griffis, Stanley E., Thomas J. Goldsby, and Martha Cooper (2003), "Web-Based and Mail Surveys: A Comparison of Response, Data, and Cost," *Journal of Business Logistics*, 24 (2), 237-58.

Hansen, Robert A. (1980), "A Self-Perception Interpretation of the Effect of Monetary and Nonmonetary Incentives on Mail Survey Respondent Behavior," *Journal of Marketing Research*, 17 (1), 77-83.

Illieva, Janet, Steve Baron, and Nigel M. Healey (2002), "Online Surveys in Marketing Research: Pros and Cons," *International Journal of Market Research*, 44 (3), 361-82.

Jöreskog, Karl G. (1971), "Simultaneous Factor Analysis in Several Populations," *Psychometrika*, 36 (4), 409-26.

Klassen, Robert D. and Jennifer Jacobs (2001), "Experimental Comparison of Web, Electronic and Mail Survey Technologies in Operations Management," *Journal of Operations Management*, 19 (6), 713-28.

Knapp, Herschel and Stuart A. Kirk (2003), "Using Pencil and Paper, Internet and Touch-Tone Phones for Self-Administered Surveys: Does Methodology Matter?" *Computers in Human Behavior*, 19 (1), 117-34.

Malhotra, Naresh K. (1999), *Marketing Research: An Applied Orientation*, 3rd ed. Englewood Cliffs, NJ: Prentice Hall.

Mardia, Kanti V. (1970), "Measures of Multivariate Skewness and Kurtosis with Applications," *Biometrika*, 57, 519-30.

McDaniel, Stephen W. and C. P. Rao (1980), "The Effect of Monetary Inducement on Mailed Questionnaire Response Quality," *Journal of Marketing Research*, 17 (2), 265-68.

McDonald, Heath and Stewart Adam (2003), "A Comparison of Online and Postal Data Collection Methods in Marketing Research," *Marketing Intelligence & Planning*, 21 (2), 85-95.

Nielsen, Jakob (2000), *Designing Web Usability*. Indianapolis, IN: New Riders Publishing.

NUA Internet Surveys (2002), "Ipsos-Reid: Internet Use Climbing in Most Markets." Retrieved January 29, 2004, from http://www.nua.com/surveys/index.cgi?f=VS&art_id=905358657&rel=true.

Nunnally, J. C. and I. H. Bernstein (1994), *Psychometric Theory*. New York: McGraw-Hill.

Parasuraman, A., Valarie A. Zeithaml, and Leonard L. Berry (1988), "SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality," *Journal of Retailing*, 64 (1), 12-40.

Roster, Catherine A., Robert D. Rogers, Gerald Albaum, and Darin Klein (2004), "A Comparison of Response Characteristic from Web and Telephone Surveys," *International Journal of Market Research*, 46 (3), 359-73.

Satorra, Albert and Peter M. Bentler (1994), "Corrections to Test Statistic and Standard Errors in Covariance Structure Analysis," in *Analysis of Latent Variables in Developmental Research*, A. von Eye and C. C. Clogg, eds. Thousand Oaks, CA: Sage, 399-419.

——— (2001), "A Scaled Difference Chi-Square Test Statistic for Moment Structure Analysis," *Psychometrika*, 66 (4), 507-14.

Schaefer, David R. and Don A. Dillman (1998), "Development of a Standard E-Mail Methodology," *Public Opinion Quarterly*, 62 (3), 378-97.

Schuldt, Barbara A. and Jeff W. Totten (1994), "Electronic Mail vs. Mail Survey Response Rates," *Marketing Research*, 6 (1), 36-39.

Shermis, Mark D. and Danielle Lombard (1999), "A Comparison of Survey Data Collected by Regular Mail and Electronic Mail Questionnaires," *Journal of Business & Psychology*, 14 (2), 341-54.

Simsek, Zeki and John F. Veiga (2001), "A Primer on Internet Organizational Surveys," *Organizational Research Methods*, 4 (3), 218-35.

Smither, James W., Alan G. Walker, and Michael K. T. Yap (2004), "An Examination of the Equivalence of Web-Based Versus Paper-and-Pencil Upward Feedback Ratings: Rater- and Ratee-Level Analyses," *Educational and Psychological Measurement*, 64 (1), 40-61.

Sparrow, Nick and John Curtice (2004), "Measuring the Attitudes of the General Public via Internet Polls: An Evaluation," *International Journal of Market Research*, 46 (1), 23-44.

Sproull, Lee and Sara Kiesler (1986), "Reducing Social Context Cues: Electronic Mail in Organizational Communication," *Management Science*, 32 (11), 1492.

Srivastava, Muni S. (1984), "Measure of Skewness and Kurtosis and a Graphical Method for Assessing Multivariate Normality," *Statistical Probability Letters*, 2 (5), 263-367.

Stanton, Jeffrey M. (1998), "An Empirical Assessment of Data Collection Using the Internet," *Personnel Psychology*, 51 (3), 709-25.

Steenkamp, Jan-Benedict E. M. and Hans Baumgartner (1998), "Assessing Measurement Invariance in Cross-National Consumer Research," *Journal of Consumer Research*, 25 (1), 78-90.

Truell, Allen D., James E. Bartlett II, and Melody W. Alexander (2002), "Response Rate, Speed, and Completeness: A Comparison of Internet-Based and Mail Surveys," *Behavior Research Methods, Instruments, & Computers*, 34 (1), 46-49.

Van Birgelen, Marcel, Ko de Ruyter, Ad de Jong, and Martin G. M. Wetzels (2002), "Customer Evaluations of After-Sales Service Contact Models: An Empirical Analysis of National Culture's Consequences," *International Journal of Research in Marketing*, 19 (1), 43-64.

Vandenberg, Robert J. (2002), "Toward a Further Understanding of and Improvement in Measurement Invariance Methods and Procedures," *Organizational Research Methods*, 5 (2), 139-58.

——— and Charles E. Lance (2000), "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research," *Organizational Research Methods*, 3 (1), 4-70.

Vehovar, Vasja, Katja Lozar Manfreda, and Zenel Batagelj (1999), "Web Surveys: Can The Weighting Solve The Problem?" in *Proceedings of the Survey Research Methods Section*, Dick Kulka, ed., conference chair. St. Pete Beach, FL: American Statistical Association, 962-67.

West, Stephan G., John F. Finch, and Patrick J. Curran (1995), "Structural Equation Models with Nonnormal Variables," in *Structural Equation Modeling: Concepts, Issues and Applications*, R. H. Hoyle, ed. Thousand Oaks, CA: Sage, 56-75.

Zeithaml, Valarie A. and Mary Jo Bitner (2003), *Services Marketing: Integrating Customer Focus across the Firm*, 3rd ed. New York: McGraw-Hill.

Elisabeth Deutskens is an assistant professor in the Department of Marketing at the University of Maastricht, the Netherlands. She recently defended her dissertation titled "From Paper-and-Pencil to Screen-and-Keyboard: Studies on the Effectiveness of Internet-Based Marketing Research." During the fall semester 2004, she was a visiting PhD student at Arizona State University. Her research interests are in the area of (online) market research, services marketing, shared consumption experiences, and value creation by customers. She has presented her work at several international conferences, and she has published in *Marketing Letters.*

Ko de Ruyter is a professor of interactive marketing and a professor of international service research at the University of Maastricht, the Netherlands, and chair of the Department of Marketing. He has published six books and numerous scholarly articles in, among others, the *Journal of Marketing*, *Management Science*, *Decision Science*, *Marketing Letters*, the *Journal of Retailing*, the *Journal of Economic Psychology*, the *International Journal of Research in Marketing*, the *Journal of Service Research*, the *International Journal of Service Industry Management*, the *Journal of Business Research*, the *European Journal of Marketing*, *Information and Management*, and *Accounting, Organisation and Society*. He serves on the editorial boards of various international academic journals, among them the *Journal of Service Research* and the *International Journal of Service Industry Management*. His research interests concern international service management, e-commerce, and customer satisfaction and dissatisfaction.

Martin Wetzels is a professor of marketing and supply chain research in the Department of Marketing at the University of Maastricht, the Netherlands. His work has resulted in more than 40 scholarly articles in journals such as *Management Science*, *Marketing Letters*, the *Journal of Economic Psychology*, the *International Journal of Research in Marketing*, the *Journal of Service Research*, the *International Journal of Service Industry Management*, the *Journal of Interactive Marketing*, the *Journal of Management Studies*, the *Journal of Business Research*, *Industrial Marketing Management*, *International Business Review*, *European Journal of Marketing*, and *Accounting, Organisation and Society*. Moreover, he has contributed more than 70 papers to conference proceedings. His main research interests are customer satisfaction/dissatisfaction, customer value, quality management in service organizations, services marketing, online marketing research, technology infusion in marketing, and relationship marketing.