UNIVERSITEIT GENT

FACULTEIT ECONOMIE
EN BEDRIJFSKUNDE

HOVENIERSBERG 24
B-9000 GENT
Tel. : 32 - (0)9 – 264.34.61
Fax. : 32 - (0)9 – 264.35.92

WORKING PAPER

Response Styles in Marketing Research : a Means And Covariance Structures Comparison of Modes of Data-Collection

Bert Weijters [1]

Niels Schillewaert [2]

Maggie Geuens [3]

December 2005

2005/349

[1] Vlerick Leuven Gent Management School, Reep 1, B-9000 Ghent, Belgium, E-mail: bert.weijters@vlerick.be
[2] Vlerick Leuven Gent Management School, Reep 1, B-9000 Ghent, Belgium, E-mail: niels.schillewaert@vlerick.be
[3] Ghent University and Vlerick Leuven Gent Management School, Hoveniersberg 24, B-9000 Ghent, Belgium, E-mail: maggie.geuens@ugent.be

Bert Weijters would like to thank ICM (Intercollegiate Centre for Management Science) for funding his research.

D/2005/7012/67

RESPONSE STYLES IN MARKETING RESEARCH : A MEANS AND COVARIANCE STRUCTURES COMPARISON OF MODES OF DATA-COLLECTION

ABSTRACT

Based on two data sets, we compare levels of response styles across three modes of data-collection: paper and pencil questionnaires, online questionnaires, and telephone interviews. Using Means And Covariance Structures (MACS), we find that data collected by different modes show differences in response styles levels. More specifically, telephone data show lower levels of midpoint responding. We propose a method to alleviate response style bias in cross-mode comparisons.

INTRODUCTION

In survey studies, researchers assume that the responses to items in a questionnaire reflect a respondent's true position with regard to the content of the question. This is, however, not always the case. The effect of random error has been generally accepted and is often accounted for by using multi-item scales (Churchill 1979). The effect of systematic error, on the other hand, poses more serious problems to the validity of survey research and has not been as widely recognized or investigated as would be warranted by its potential biasing effects (Baumgartner and Steenkamp 2001; Podsakoff, Mackenzie, Lee and Podsakoff 2003). Often, respondents seem to be prone to response styles, defined as "*[behavior patterns] where the individual tends to select disproportionately a particular response category regardless of item content*" (O'Neill 1967). While several methods of measuring response styles have been proposed (Baumgartner and Steenkamp 2001; Greenleaf 1992a), it is remarkable that response styles have not been operationalized as latent constructs, as pointed out by Podsakoff et al. (2003). Apart from being valid from a conceptual point of view, an important advantage of such an approach would be that random measurement error in the response style measures can be corrected for (Podsakoff et al. 2003). Additionally, the use of multi-group means and covariance structures (MACS), would enable researchers to verify whether response style factors are actually measuring the same construct across groups, a concept referred to as measurement invariance (Meredith 1993; Ployhart and Oswald 2004). Measurement invariance is a necessary condition for conducting substantive cross-group comparisons (Vandenberg and Lance 2000). The first objective of our research is to operationalize response styles as latent constructs in a MACS framework, such that we can (1) test for measurement invariance across groups, (2) make means and covariances comparisons across groups, and (3) implement response style corrections in MACS.

Our first objective at the same time is a means to an end, the end being a comparison of different modes of data collection in terms of response styles levels.This is our second research objective. Imagine a researcher wanting to compare levels of trust and loyalty of online and offline customers of a multi-channel retailer. Often, these respective groups would be most easily contacted online versus offline, due to contact data availability, cost considerations, etc. The comparison between the groups would

be rendered useless, however, if data collected online versus offline were contaminated by response styles to a different extent. Therefore, we believe it is crucial to investigate whether different modes of data collection bring along different levels of response styles. This is an important issue for both practical and academic marketing research with repercussions on the optimal choice of a data collection procedure. Especially with regard to the growing importance of the Internet and web surveys (Gunter et al. 2002; Johnson 2001; Griffis, Goldsby and Cooper 2003), such comparison would enrich our understanding of the comparability of various research methods. Although researchers have identified a wide range of possible (dis-)advantages of web surveys, the focus of previous research is mainly on response rate, response speed, costs, representativeness of samples, anonymity and confidentiality (Deutskens et al. 2004; Gunter et al. 2002; Truell 2003; also see Ployhart et al. 2003; Simsek and Veiga 2001; Thompson et al. 2003). To the best of our knowledge, however, no research is available that compares offline self-administered questionnaire data, telephone interview data and online self-administered data in terms of systematically measured response styles. This is done in the current paper. We test the null hypothesis that mode of data collection does not lead to different levels of response style bias. This hypothesis is relevant because it is the current working hypothesis of many marketing researchers. In order to address the above questions, we study two data sets, each consisting of three subsamples of respondents who answered the same questions via a different mode of data collection. Additionally, we test whether cross-mode comparisons of substantive[1] construct measures are biased by response styles and to what extent our proposed operationalization can correct for such bias if present.

The paper is organized as follows. In our theoretical framework, we respectively discuss response styles and modes of data collection. Next, we present two studies in which we use MACS to compare response styles levels in data from online, paper and pencil, and telephone respondents. Both studies have similar set-ups, but use different samples of respondents and items. In the second study, we additionally illustrate a correction method for response styles across modes of data collection. We end the

[1] We use the term 'substantive measures' to set them apart from 'methodological measures' like response styles.

paper with a conclusion based on our findings, including limitations of the current research and possibilities for future research.

THEORETICAL FRAMEWORK

RESPONSE STYLES

Measurement items are being designed to capture true variance. Unfortunately, this aim is not fully met due to the presence of error variance. Error variance has two components, a random and a systematic component. The systematic component can be split up into content related systematic error due to response sets, and non-content related systematic error due to response styles (Baumgartner and Steenkamp 2001). A response set means a respondent wants to create an impression of her/himself with regard to the item content. Social desirability is a well-known example of this (Mick 1996). A response style, on the contrary, is a tendency to answer items in a certain way regardless of content (O'Neill 1967). Response styles are not limited to specific domains, such as socially sensitive variables, or so-called 'dark side variables' (Mick 1996) but are omnipresent in marketing research (Baumgartner and Steenkamp 2001). This study exclusively deals with response styles. Following Baumgartner and Steenkamp (2001), we consider response styles as an interaction of personal dispositions and situational factors other than questionnaire content, e.g. mode of data collection.

Based on the impact they have on observed scores, we distinguish between two types of response styles: unidirectional and bidirectional. Unidirectional response styles refer to a respondent's preferred use of either positive, neutral or negative response options. The net result of these styles is a shift in the Intra-Subject Mean (ISM), i.e. the respondent's mean response over a set of unrelated items (Greenleaf 1992a; cfr. Net Acquiescence Response Style in Baumgartner and Steenkamp 2001). There are three such unidirectional response styles (Baumgartner and Steenkamp 2001): Acquiescence Response Style (ARS), i.e. the tendency to disproportionately use positive response categories; Disacquiescence Response Style (DARS), i.e. the tendency to disproportionately use negative response categories; and Midpoint Responding (MPR), i.e. the tendency to disproportionately use the midpoint of a scale. Bidirectional response styles, on the other hand, refer to a respondent's

tendency to use response categories in a narrow or broad range. Two response styles fall into this category (Baumgartner and Steenkamp 2001; Greenleaf 1992a, b): Extreme Response Style (ERS), the tendency to use the most extreme response options on both the left and the right hand side of the scale, and Response Range (RR), i.e. the tendency to use response options in a broad rather than a narrow range around the midpoint. The net result of these bidirectional styles is a change in the Intra-Subject Standard Deviation (ISSD), i.e. the respondent's standard deviation over a set of unrelated items (Greenleaf 1992a). Implicit in the above discussion is the distinction between raw response styles and net response styles. Raw response styles are the direct behavioral tendencies shown by the respondent (ARS, DARS, MPR, RR, ERS). Net response styles, ISM and ISSD, are summary measures of the impact of the raw behavioral tendencies on the observed scores of the respondent. More specifically, the unidirectional raw response styles ARS, DARS and MPR are the antecedents of the unidirectional net response style ISM, while ERS and RR affect the bidirectional response style ISSD. As demonstrated by Cheung and Rensvold (2000), response styles affect the observed scores and their relation to the latent variables they reflect. More specifically, in a measurement model where observed variable x is related to latent variable ξ such that $x = \tau + \lambda\xi + \delta$, higher/lower ISM leads to higher/lower measurement intercepts τ, and higher/lower ISSD leads to higher/lower factor loadings λ. Consequently, if groups have different levels of response styles, this will lead to between-group differences in measurement intercepts and loadings. However, to be able to compare groups in terms of latent means, scalar and metric invariance have to be satisfied (Little 1997; Vandenberg and Lance 2000). Scalar invariance refers to the condition in which the measurement intercepts τ are equal across groups, while metric invariance refers to the condition in which the measurement slopes λ are equal across groups (Steenkamp and Baumgartner 1998). Inter-group differences in response styles may threathen metric and scalar invariance, rendering inter-group comparisons impossible (Cheung and Rensvold 2002). Remarkably, however, the current measurement invariance literature is limited to diagnosing the net impact of response style differences, without (1) diagnosing the raw response style phenomena driving this impact, and (2) without trying to solve problems of measurement invariance if present. In this paper, we aim to address both gaps. Each of these two goals calls for a different approach though. On the one hand,

to be able to diagnose between-group differences in response styles, some conditions have to be met. First, we need an operationalization representing a complete profile of all unidirectional and bidirectional response styles. Second, we need an operationalization of response styles that has itself metric and scalar measurement invariance across groups. Not meeting the measurement invariance condition would invalidate multi-group comparisons of these measures themselves (Steenkamp and Baumgartner 1998). We therefore propose the use of multi-indicator measures of ARS, DARS, MPR, RR and ERS for the purpose of diagnosing between-group differences in response styles. This method will be applied in the response style mean comparisons of Study 1 and Study 2.

On the other hand, to be able to correct substantive models for measurement invariance due to response styles, we need a response style operationalization that meets the following conditions. First, parsimony rather than completeness is an advantage if the aim is to correct a substantive model, which in and of itself might already be complex. Second, ease of implementation would clearly increase the chance that in the future researchers will correct for response styles. In an SEM context, ease of implementation includes considerations of model identifiability and avoidance of multicollinearity. Third, measurement invariance of the response styles measures themselves has to be tested for. Finally, there has to be a clear theoretical link with measurement invariance of the substantive constructs to be corrected. More specifically, both scalar and metric invariance of the substantive model should be accounted for. We therefore propose the use of multi-indicator measures of ISM and ISSD to correct substantive models for response style bias. This method will be applied in part 2 of Study 2, "Impact of Response Styles on a Substantive Model".

For both the purpose of diagnosis and correction we suggest the use of multiple rather than single indicator measures for each response style. Each indicator will be based on a random subset of items taken from a broader set of heterogeneous items. The purpose of this approach is to (1) decrease the probability that systematic variance other than response style variance is included in the response style measure, and (2) assess measurement invariance of the response style measures. An example will clarify this. Imagine that respondents in one group systematically respond more positively to an item. When using three random subsets of items, this effect will be limited to one indicator. More specifically, the analyses will show the intercept of the

indicator containing the item is higher, whereas in the one-indicator case, this effect would be passed on to the structural mean. Figure 1 depicts how the use of multiple indicators allows the researchers to discern between an effect of the grouping variable on the indicator (dotted arrow, indicating scalar non-invariance), and an effect of the grouping variable on the factor (plain arrow between mode and response style).

< Insert Figure 1 about here >

MODES OF DATA COLLECTION

Notwithstanding the availability of several modes of data collection and the growing success of the Internet in this regard (Johnson 2001), little research is available that addresses the impact of mode of data collection on response styles. Jordan, Marcus and Reeder (1980) compare telephone and household interviews, and find more acquiescence and extremeness in the telephone interviews. Kiesler and Sproul (1986) compared electronic and paper mail self-administered surveys in terms of the contents of responses to a specifically health related questionnaire. They found that in the electronic surveys, people tended to show less inhibitions in their responses, and concluded that their results "*show considerable similarity of response between the paper and electronic survey but not so much that the two may be considered interchangeable without further research"*. The measures used by Jordan et al. (1980) and Kiesler and Sproul (1986), however, are constructed ad hoc and related to the specific content of the questionnaire (health issues in both cases), rendering assessment of response styles as non-content related patterns of responses tentative. The conclusions therefore may not be generalizable to consumer surveys on less privacy sensitive topics as are often encountered in marketing. Moreover, the stricter operationalizations suggested by Greenleaf (1992a) and Baumgartner and Steenkamp (2001) were not yet available at the time of their research, and the response styles were not operationalized as latent constructs, missing the opportunities this would offer, as discussed above. Additionally, by now, the Internet has become more commonly adopted, possibly leading to convergence of communication patterns. Consequently, the question remains open whether and to what extent mode of data collection systematically affects (non-content related) response styles. The topic of

mode comparability is becoming especially important since substantive questions need to be answered concerning the generalizability of marketing models from an offline to an online context (see for example Szymanski and Hise 2000; Venkatesh, Smith, and Rangaswamy 2003). Often, cost and data availability considerations lead to the situation in which respondents in the offline and online settings are easier to reach respectively by means of mailed paper surveys and e-mails linking to online questionnaires. However, to be able to assess generalizability of a substantive model across modes, a first prerequisite is generalizability of measurement. It is crucial to evaluate the hypothesis that bias due to response styles is not different across modes. Therefore, in this paper, we compare the levels of bias due to response styles in three types of questionnaire data: self-administered paper and pencil questionnaires, telephone interviews, and self-administered online questionnaires.

STUDY 1: DIAGNOSIS OF CROSS-MODE DIFFERENCES IN RESPONSE STYLES

In the first study, we specify and test a multi-group cross-mode MACS that allows us to assess response style measurement invariance across modes of data collection, and to compare levels of response style bias across modes of data collection.

METHODOLOGY

Respondent sampling

We collected data among three samples. The first data collection mode used was the online opt-in web panel of an Internet market research company. 170 panel members were randomly selected and sent an e-mail including a link to a web survey. This recruitment procedure generated 84 responses or a 49.4% response rate. Next to this, a total of 450 postal mail surveys were sent out, including a pre-paid envelope directly addressed to the researchers. The mailing generated 141 completed questionnaires or a 31.3% response rate. Finally, 122 responses were generated by means of a telephone interview using random digit dialing. In total, 350 people were called which implies a usable response rate of 34.9%. The three samples were comparable in social demographic composition. The respective samples (paper and pencil, telephone, online) had average ages of 40.2 (sd 15.6), 40.1 (sd 16.3), and 39.9 (sd 13.2), years of

formal education of 12.9 (sd 3.1), 13.2 (sd 2.5), and 13.1 (sd 2.3), and proportions of females of 58%, 49% and 55%. Based on analyses of variance and a chi square test, we found the samples to be similar in terms of educational level, age and gender (respective p-values for educational level, age and gender are .862, .200 and .720).

Questionnaire and item sampling

The questionnaire consisted of social demographic questions and 21 unrelated seven point Likert items measuring attitudes, interests and preferences concerning diverse topics such as leisure activities, fast moving consumer goods, fashion and others (e.g. "I like driving a nice car"). The average inter-item correlation was .12.

Response style indicator calculation

We randomly select three subsets of 7 items each[2]. Each subset is then used to compute response style indicators. This allows us to compute three indicators for each of the following response styles: ARS, DARS, ERS, MPR, RR. We apply the following formulas based on Baumgartner and Steenkamp (2001) and Greenleaf (1992a) and rescaled to a 0-100 range to make scores comparable across studies[3]. For each set of k items:

(1) $$ARS = 100 * [f(5)*1 + f(6)*2 + f(7)*3] / 3k$$

(2) $$DARS = 100 * [f(1)*3 + f(2)*2 + f(3)*1] / 3k$$

(3) $$ERS = 100 * [f(1) + f(7)] / k$$

(4) $$MPR = 100 * f(4) / k$$

(5) $$ISM = [f(1)*1 + f(2)*2 + f(3)*3 + f(4)*4 + f(5)*5 + f(6)*6 + f(7)*7]/k$$

(6) $$ISSD = [[(f(1)*(1-ISM)^2 + f(2)*(2-ISM)^2 + f(3)*(3-ISM)^2 + f(4)*(4-ISM)^2 + f(5)*(5-ISM)^2 + f(6)*(6-ISM)^2 + f(7)*(7-ISM)^2]/(k-1)]^{1/2}$$

(7) $$RR = 100* ISSD * [(k*3)^2 / (k-1)]^{-1/2}$$

---

[2] For a discussion of the advantage of three indicators (or 'parcels'), we refer to Little et al. (2002). The number of items needed as a basis for response style indicators is discussed in Appendix A.

[3] In line with our definition of response styles, we base our response styles on sets of unrelated items and use operationalizations that are not content related (more specifically, for ARS and DARS, we use 'ARS1' and 'DARS1' in Baumgartner and Steenkamp 2001).

where f(o) refers to the frequency of response option o, and ISM and ISSD (which were not rescaled to a 0-100 range) serve as temporary variables in computing RR[4] and are not used in the subsequent analyses.

MACS Model and data analysis

We compare response styles across three modes (paper and pencil, telephone and online) by specifying a multi-group MACS in which ARS, DARS, ERS, RR and MPR are freely covarying latent constructs. Each factor has three indicators. Across response styles, the indicators that are based on the same sets of items have correlated error terms to take into account the shared variance due to basing measures of response styles on the same items (see Figure 2)[5].

< Insert Figure 2 about here >

We specify nested models to test for measurement and structural invariance. In Appendix B - 1 we indicate the parameters that are fixed in each subsequent model. Based on a review of the measurement invariance literature, we formulate the following procedure to assess whether the subsequent null hypotheses of invariance are rejected (Cheung and Rensvold 2002; Jöreskog 1971; Vandenberg and Lance 2000; Little 1997; Meredith 1993; Ployhart and Oswald 2004; Steenkamp and Baumgartner 1998). First, the chi square difference test is evaluated (Jöreskog 1971). If it is insignificant, the invariance hypothesis is accepted. If it is significant, we

[4] RR and ISSD are operationally identical except for a scaling factor. Conceptually, however, ISSD is a resultant of RR and ERS, which are behavioral tendencies. The weighting implied in the RR measure (taken from Baumgartner and Steenkamp 2001), which makes it so similar to ISSD, is a convention rather than a necessity. We therefore maintain the distinction between RR and ISSD. Also note that extreme responses reflecting ERS are also a component of the RR and ISSD formulas. Hence, ERS affects ISSD, and this is reflected in the operationalization.

[5] Such model corresponds to a covariance matrix of the indicators in which not only the main diagonal (containing the variances) is systematically higher than the other values, but also the diagonals of each submatrix corresponding to indicators of different styles based on the same sets of items.

evaluate the change in CFI: a decrease in CFI equal to or higher than .01 leads to rejection of the null hypothesis of invariance (Cheung and Rensvold 2002). Additionally, in cases where the chi square difference test is significant, we evaluate to what extent indicators of local misfit, modification indices (M.I.'s) and standardized residuals (s.r.'s), show consistent patterns of significant values (Steenkamp and Baumgartner 1998; Little 1997). If the decrease in CFI is smaller than .01 and the local misfit indices do not show consistent patterns, the hypothesis of invariance is accepted.

FINDINGS STUDY 1

When testing the model for the three groups simultaneously, resulting model fit is good (see model A in Table 1). The standardized factor loadings and composite reliabilities are presented in Table 2.

< Insert Table 1 and Table 2 about here >

Next, we test for measurement and structural invariance. Metric invariance is accepted due to the insignificant chi square difference test. The chi square test for scalar invariance yields a significant result. The decrease in CFI is much smaller than .01 (Cheung and Rensvold 2002), however. Moreover, the indices of local misfit indicate that the misspecifications are small and randomly dispersed (Steenkamp and Baumgartner 1998; Little 1997): all modification indices for the intercepts are small (all M.I.'s < 6.63, which corresponds to a p-value of .01), as are the standardized residuals (all s.r.'s < 2.56, which corresponds to a p-value of .01). Therefore, following the procedure outlined above, we accept scalar invariance. In a subsequent test, structural means invariance is rejected, due to the highly significant chi square difference test and the decrease in CFI which is larger than .01 (Cheung and Rensvold 2002). We try to locate the reasons behind this misfit by evaluating the standardized residuals in the restricted model D in combination with the latent means in the unrestricted model C (Steenkamp and Baumgartner 1998; Little 1997). From this, it seems that the telephone group has a lower MPR mean, with standardized mean residuals ranging from -5.2 through -5.38, while having consistently higher means on the other response styles, especially RR and DARS. Based on these observations, we

test a model in which we free the latent means of the telephone group, while constraining the other two groups to equality. The resulting model compares very well to reference model C (scalar invariance): the chi square difference test is not significant on the .05-level. We therefore reject overall invariance of structural means, but accept invariance of means of the paper and pencil and online groups. The means as estimated in the partial structural mean invariance model E are reported in Table 3.

<Insert Table 3 about here.>

STUDY 2 : DIAGNOSIS OF CROSS-MODE DIFFERENCES IN RESPONSE STYLES

To test whether the above findings hold across different sets of items and respondents, we conduct a second study with a new data set. Additionally, we include a randomly selected model in the questionnaire with the aim of illustrating a proposed correction procedure for response styles. The latter topic is discussed later, under the heading "Study 2 : Impact of response styles on a Substantive Model". First we validate the findings of Study 1 on the new samples.

METHODOLOGY

Respondent sampling

We collect data among three samples of respondents, using identical questionnaires across three modes of data collection: (1) Self-administered paper and pencil questionnaire: N=655, recruited by means of a random walk procedure[6] (response rate 58.0%); (2) Telephone interview: N = 496 (response rate 32.0%); (3) Self-administered online survey among an online panel recruited by means of a personalized e-mail: N=1445 (response rate 48.2%).

[6] For each day of data collection, each data collector received one randomly generated address, covering city, suburb and countryside. From this start address, they followed a predefined procedure explaining how to select the next address. Questionnaires were collected two days later.

In order to obtain comparable samples, we resample three equally large samples from the above groups, matching for age, education level and sex. Since the telephone sample is the smallest group, it is used as the reference group in computing sampling probability weights. As intended, the resulting samples show no significant differences on the three demographic variables in chi square and anova tests (respective p-values for age, education and sex are .993, .856 and .434). The respective matched samples[7] (paper and pencil, telephone, and online) have average ages of 46.3 (sd 13.9), 46.3 (sd 13.0), and 46.2 (sd 13.4); average years of formal education of 12.5 (sd 2.7), 12.6 (sd 2.6), and 12.6 (sd 2.6); percentages of females of 64.9%, 65.7%, and 62.1%. Sample sizes are 501, 496, and 535 respectively.

Questionnaire and item sampling

From the marketing scales handbook by Bruner, James and Hensel (2001), we selected a total of 60 items, consisting of 52 unrelated items and 8 items measuring two constructs. All items were adapted to a seven point Likert scale. We subsequently discuss each group of items. First, we randomly selected 52 items from different scales. To be able to assess the impact of response styles on substantive measures (see below), we also included multi-item measures of two related constructs, together constituting what we will label the loyalty diad: trust in frontline employees (TRUST) and loyalty (LOYAL) in a clothing retail context. Both constructs are measured by means of four items each, taken from Sirdeshmukh, Singh, and Sabol (2002). For this measurement, respondents are asked to think back of their latest such encounter. The TRUST and LOYAL items are grouped in one block under this heading.

Response style indicator calculation

The 52 randomly selected items have an average inter-item correlation of .07. We randomly split them into three sets, each of which is used to calculate an indicator for each response style using equations 1 through 7. Each set now consists of 17 or 18 items.

[7] We tested whether our results were robust against fluctuations in sampling. This proved to be the case.

Model

We again compare mean response style scores across modes by means of the MACS model specified in Study 1 and Appendix B -1.

FINDINGS CROSS-MODE RESPONSE STYLE COMPARISON STUDY 2

The MACS model is fitted to the new data (see Table 1). Although the chi square value for the unconstrained model (model A) is significant, the alternative indices have acceptable values and there are no indications of particular misspecifications. Factor loadings and composite reliabilities are listed in Table 2. We gradually constrain the model further by imposing subsequent levels of invariance. To evaluate invariance, we use the procedure outlined in Study 1. Note, however, that sample sizes are larger than in Study 1, such that chance of rejecting the model based on chi square values can be expected to be higher (Marsh, Balla and McDonald 1988). Imposing metric invariance (model B), results in a chi square difference test which is significant on the .01-level but not the .001-level. The CFI value hardly decreases and the indices of local misfit do not show any meaningful patterns. Based on these observations, and in line with our procedure, we accept metric invariance. Scalar invariance (model C) results in a chi square difference test which is significant on the .001-level. The decrease in CFI is still quite small, however, and again there is no significant and consistent pattern to be found in the MI's (all below 4) and standardized residuals (only 3 unrelated intercepts' standardized residuals exceed 2.56). In line with our procedure, we therefore accept scalar invariance, while admitting that acceptance of scalar invariance is less obvious here than it was in Study 1. Imposing structural mean invariance (model D) leads to a highly significant chi square difference test and a small drop in CFI. In the telephone group, standardized mean residuals for MPR range from -3.19 through -3.64, while those of other response styles are positive (with one exception). This observation indicates that mean invariance is improbable. Therefore, in line with the results of Study 1, and based on an evaluation of the means in the unrestricted model and the standardized residuals in the restricted model, we test a model in which the latent means of the telephone group are freely estimated, while those of the paper and pencil and online group are constrained to equality. As in Study 1, this model's fit is comparable to that of its reference model (model C). The chi square difference test is significant on the .01 but

not the .001-level, and the decrease in CFI is very small. Indices of local misfit also do no longer show consistently significant patterns. Based on the above thorough evaluation, we reject full structural mean invariance, but we accept partial mean invariance of the paper and pencil and online groups. The latent means as estimated in the partial mean invariance model are presented in Table 3.

DISCUSSION RESPONSE STYLE COMPARISON STUDY 1 AND 2

The results reported above show that response styles can be usefully operationalized as latent constructs in means and covariance structures (MACS), thus meeting our first research objective. In both Study 1 and Study 2, we found measurement (metric and scalar) invariance and structural (mean) non-invariance. Measurement invariance indicates that the variance in the indicators is related in similar ways to the response styles they measure across the modes of data collection. Structural mean non-invariance indicates that response styles levels are not the same across modes. We discuss these differences in more detail.

In both Study 1 and Study 2, the telephone group's latent means markedly diverged from the other two modes (Table 3). The most consistent phenomenon that emerges from both studies, is that levels of acquiescence response style (ARS), disacquiescence response style (DARS), extreme response style (ERS) and response range (RR) are higher in the telephone data, while the midpoint responding (MPR) scores are substantially lower in the telephone group. In the telephone mode, the probability of respondents choosing the neutral point of a scale is smaller than in the other modes, leading to a shift to all other response options. This, in turn, may result in higher ARS, DARS, ERS and/or RR scores. While the current data do not allow us to conclusively interpret the mechanism underlying this phenomenon, a plausible interpretation is that respondents interviewed by telephone lack the visual representation that allows one to easily determine the neutral point in a scale. We note that the differences in response styles between modes are smaller in Study 2 than in Study 1. Probably, the quality of the items in this study was better. The items in Study 2 were all taken from throroughly validated marketing scales, while some items in Study 1 were ad hoc measures. So, while validated items may be less sensitive to response styles, they do definitely not rule out their effect.

Another important finding is that the online mode of data collection is similar to the paper and pencil mode of data collection in terms of response styles. This means it is possible to collect data offline and online and still obtain comparable measurements.

STUDY 2 : IMPACT OF RESPONSE STYLES ON A SUBSTANTIVE MODEL

In the above studies, our focus was on the diagnosis of response styles. As pointed out before, this objective called for completeness of the response style operationalization. In what follows, we demonstrate a proposed method to correct for cross-mode response style bias. For this aim, we make use of the net response styles ISM and ISSD (see theoretical framework).

MODEL AND METHODOLOGY

To correct a substantive model, we regress indicators of substantive measures on response styles (Baumgartner and Steenkamp 2001). As a substantive model, we use the loyalty diad. The model is represented in Figure 3.

< Insert Figure 3 about here >

FINDINGS

Calibration

From the online respondents not in the matched sample, we take a calibration sample which can be used to test the loyalty diad measurement model before validating it across the three groups. The sample size of this group is 500, and its demographic composition is comparable to the other online sample selected as the matched sample (see above). In the calibration sample, the two-factor structure with four indicators per factor shows mediocre fit: chi square(19)=81.398, p<.001; CFI=.973; TLI=.961; RMSEA=.092. After inspection of the modification indices, two items are deleted from the model, one for each factor . The resulting factor structure shows good fit (chi square(8)=10.763, p=.215; CFI=.999; TLI=.997; RMSEA=.030). Standardized factor loadings range from .76 through .94.

Cross-mode analysis

Using the three matched samples discussed above, we assess measurement and structural invariance of this nomological network across the three modes of data-collection by gradually imposing more stringent constraints on the model. In Appendix B - 2, we list the restricted parameters for each model.

In a first stage, only the uncorrected loyalty diad (depicted in the right hand pane of Figure 3) is subjected to invariance tests. In a second stage, we include ISM and ISSD in the model (the 'correction model' in the left pane of Figure 3) and again subject the loyalty diad to a sequence of invariance tests. Metric and scalar invariance for the ISM and ISSD indicators are imposed in all models, based on our earlier findings and a preliminary analysis of the ISM/ISSD measurement model in line with the procedure described above applied in the previous analyses. The ISM and ISSD factors are modeled to impact the indicators of both loyalty diad constructs. Table 4 presents the fit indices for the nested models, both the uncorrected and the corrected models. In Table 5 and 6, we compare the parameter estimates for the uncorrected and corrected models.

< Insert Table 4, Table 5 and Table 6 about here >

DISCUSSION STUDY 2: IMPACT OF RESPONSE STYLES ON A SUBSTANTIVE MODEL

Without scrutinizing all results in detail, we make some general observations concerning the differences between the model fit and parameter estimates in the uncorrected versus the corrected model. We discuss model fit, measurement model and structural model with and without correction for response styles.

As for model fit, we note that the correction model adds six observed variables, but succeeds in showing reasonably good fit as compared to the uncorrected model. While the model including the correction model shows statistical misfit in its chi square, the alternative fit indices have acceptable to good values (Hu and Bentler 1999). More importantly, this good fit is maintained throughout the further imposition of additional restrictions of invariance, while this is not the case for the uncorrected model (see Table 4).

As for the measurement model with and without response style correction, we especially note the inflation of factor loadings in the absence of response style correction. On average, standardized factor loadings in the uncorrected model are respectively 5%, 13% and 7% higher than in the corrected model for the paper and pencil, telephone and online groups (see Table 5). This illustrates how factor loadings can be substantially inflated when no response style correction is implemented, especially so in telephone interviews. All items are significantly and substantially influenced by ISM, with standardized regression coefficients ranging from .15 through .42. Just over half of the items are significantly influenced by ISSD. Here, standardized coefficients are lower, ranging from insignificant through .21 (see Table 5).

As for the structural model, the apparent mean difference across different modes of data collection is corrected for by including the ISM and ISSD model. First, the chi square difference test testing for mean/intercept invariance becomes insignificant (see Table 4). Second, the direct mean comparison shows that in the corrected model the initial significance of TRUST in the telephone group disappears (see Table 6a). As for the regression, a comparison of the uncorrected and corrected model indicates that the uncorrected standardized model estimates are inflated by 5 to 9% (see Table 6c).

In these data it seems that the ISM/ISSD correction has the effect of correcting for (1) measurement non-invariance, more particularly mode specifically inflated loadings (linking items to factors) and mode induced scalar non-invariance, and (2) structural non-invariance, more particularly mode specifically inflated regressions (linking factors to factors), and mode induced non-invariance of structural means.

CONCLUSION

In this paper, we compare levels of five response styles in three modes of data collection in a means and covariance structure (MACS): acquiescence response style (ARS), disacquiescence response style (DARS), extreme response style (ERS), midpoint responding (MPR), and response range (RR). The main advantages of the MACS approach are that it accounts for measurement error in the response style measures (Podsakoff et al. 2003), and allows to assess measurement invariance across groups of respondents (Little 1997). We apply this model to two data sets, each

consisting of respondents in three modes of data collection: (1) telephone interviews, (2) self-administered paper and pencil questionnaires, (3) self-administered online questionnaires. The operationalization shows measurement invariance across the three groups, which makes it an appealing method for use in similar settings. The findings of our mean comparison are important, in that they show the comparability of online and offline data when using self-administered questionnaires. Telephone interview data should be handled with caution, however, in that they show systematic bias as compared to other data. This conclusion is in line with findings by Jordan, Marcus and Reeder (1980) in a different context and using a more limited set of measures: these authors did not distinguish between ERS and MPR, and captured both response styles under the label extremeness. It is apparent from our data that telephone interviews result in lower MPR, and slightly but systematically higher levels of ARS, DARS, ERS and RR. Telephone survey participants seem to use a wider range of rating scale options away from the midpoint. This may be due to the stronger impact of primacy and recency effects for the recall of response options. During a telephone interview, the respondent cannot see all response options simultaneously, but has them presented consecutively by the interviewer.

In addition to demonstrating systematic cross-mode response style bias, based on data from the second study, we propose and apply a method to correct for such bias.

Using two response styles, intra-subject mean (ISM) and intra-subject standard deviation (ISSD), we correct for the effect of cross-mode differences in response styles in a basic marketing model with two latent variables. From this, we find that our proposed method largely corrects for cross-mode measurement non-invariance. Moreover, in all groups the factor loadings and regression weights turn out to be inflated in absence of our proposed correction, while some means are overestimated as well.

Our findings need to be taken into account in future research that aims to compare theoretical models in an online to an offline context. For such comparisons, we recommend the use of self-administered paper and pencil questionniares and self-administered online questionnaires, and not telephone interviews. Moreover, it is advisable to test for response style differences between modes of data collection before proceeding to the actual comparisons between online and offline measurement and structural models. Based on our research, we recommend the following procedure for cross-mode marketing research. (1) Include a set of unrelated items in your

questionnaire, or try to distill these from parts of the questionnaire that you do not need for your research question at hand. The latter is often possible when several research topics are pooled in one questionnaire. A minimum of 21 items is recommended, as discussed in Appendix A. (2) Diagnose response styles by means of the 5-style typology ARS, DARS, ERS, MPR, RR (as illustrated in Study 1 and in part 1 of Study 2, as well as Figure 2). (3) If significant differences in response style levels are apparent from the previous step, include ISM and ISSD in your model (as illustrated in part 2 of Study 2 and Figure 3).

LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

While this paper makes a clear contribution to the literature on response styles, several limitations provide opportunities for future research. First, like Greenleaf (1992a) and Baumgartner and Steenkamp (2001), we limit our scope to one type of measurement scale. All items used and discussed in this paper are seven-point Likert items. It might be interesting to study how scale format is related to response styles.

Also, it would be enlightening to further study what causes different levels of response styles in different data collection settings. Experimental work might be used to manipulate factors like mood, fatigue or cognitive load and see how these relate to response styles levels.

APPENDIX A

NUMBER OF ITEMS AND EXPLAINED INDICATOR VARIANCE

In Study 1, we compute response style indicators based on sets of 7 items, in Study 2, we use sets of 17 (to 18) items. In the former study, the average inter-item correlation is .12, while in the latter, it is .07. In this appendix, we shortly illustrate how these parameters might affect the proportion of indicator variance explained by the latent response style variable and the expected factor loading for such indicators. We use ISM as the most straightforward example.The total variance in a summed set of items can be expressed as follows:

$$\sigma_s^2 = \Sigma\, \sigma_i^2 + \Sigma\Sigma\, \sigma_{ij}, \quad (1)$$

where σ_s^2 is total variance in the summed set score, σ_i^2 is the variance in item i, and σ_{ij} denotes the covariance between item i and j, and the double summation is taken over all combinations of i and j where $i \neq j$. Formula (1) shows that the variance of the summed set score can be divided in a component due to item variances and a component due to item covariances, assumed to reflect response styles in this study, given the unrelatedness of the items in terms of content. Assumptions in this illustration are (1) the covariances between items that are not related in terms of content reflect common response style variance; and (2) for reasons of simplicity we set all variances equal to one, and all covariances equal to the average inter-item covariance. Assuming variances of one and covariances of respectively .12 and .07, we can then calculate the hypothetical proportions of variance due to the common underlying response style factor ISM, which will be reflected in the factor loading, and the component due to item specific variances, which will be reflected in the measurement residual. We summarize these results for a hypothetical ISM factor in Table A-1. In this table, a stands for $\Sigma\, \sigma_i^2$, b stands for $\Sigma\Sigma\, \sigma_{ij}$, b/(a+b) refers to the common variance divided by the total variance, and lambda stands for the estimated standardized loading assuming two similar indicators of one response style factor. The number of items is equal to component a due to the assumption that all variances are 1. From the data in Table A-1, and conditional on the above assumptions, we can

derive the following recommendations for two different situations, namely (1) cases like Study 1, where inter-item correlations are around .12; (2) cases where inter-item correlations are around .07. For (1), to obtain response style indicator factor loadings of about .60, sets of 6 or more items suffice (as in Study 1); to obtain response style indicator factor loadings of .70 or more, sets of 10 or more items are recommended. For (2), to obtain response style indicator factor loadings of about .60, sets of 9 items or more suffice; to obtain response style indicator factor loadings of .70 or more, sets of 15 items or more are recommended (as in Study 2, Part 1). Note however, that the best guarantuee for good response style measures lies in the sampling of items. These should be heterogeneous in content and unrelated both to one another and to substantive constructs in the model. Watson (1992) for example, constructs a response style measure by including a single indicator that is not related to the indicators of substantive measures in the model. This illustrates how it is not necessarily considered self-evident to apply common standards of explained variance or size of factor loadings to operationalizations of response style. We especially caution against the uncritical inclusion of more items to enhance such indicators of measurement quality, since this may lead to confounding response style based covariance with true content based covariance. Also, making a questionnaire extraordinarily long to better grasp response styles might end up as a self-fulfilling prophecy, as longer questionnaires might induce more respondent fatigue and an increase in response style bias.

APPENDIX B - 1

MACS FOR RESPONSE STYLE MEAN COMPARISON IN STUDY 1 AND STUDY 2

$x^{(g)} = \tau^{(g)} + \Lambda^{(g)} \xi^{(g)} + \delta^{(g)}$

where g refers to groups (1) paper and pencil, (2) telephone and (3) online; $x^{(g)}$ is a 15*1 vector with observed response style indicators; $\tau^{(g)}$ is a 15*1 vector with measurement intercepts of the response style indicators; $\Lambda^{(g)}$ is a 15*5 matrix with factor loadings of the response style indicators; $\xi^{(g)}$ is a 5*1 vector with latent response styles; $\delta^{(g)}$ is a 15*1 vector with residuals of the response style indicators. Moreover, means of $\xi^{(g)}$ are represented by $\kappa^{(g)}$, a 5∗1 vector. For reasons of identifiability, for each latent factor $\xi^{(g)}$, we fix one $\lambda^{(g)}$ to 1 and we fix one $\tau^{(g)}$ to 0 in all models. Consequently, in the unconstrained model, for each group 10 λ's out of 15 and 10 τ's out of 15 are freely estimated.

Model	Additionally constrained parameters
A. Base model	
B. Metric invariance:	$\Lambda_x^{(1)} = \Lambda_x^{(2)} = \Lambda_x^{(3)}$
C. Scalar invariance:	$\Lambda_x^{(1)} = \Lambda_x^{(2)} = \Lambda_x^{(3)}$
	$\tau_x^{(1)} = \tau_x^{(2)} = \tau_x^{(3)}$
D. Structural mean invariance:	$\Lambda_x^{(1)} = \Lambda_x^{(2)} = \Lambda_x^{(3)}$
	$\tau_x^{(1)} = \tau_x^{(2)} = \tau_x^{(3)}$
	$\kappa^{(1)} = \kappa^{(2)} = \kappa^{(3)}$
E. Partial structural means invariance:	$\Lambda_x^{(1)} = \Lambda_x^{(2)} = \Lambda_x^{(3)}$
	$\tau_x^{(1)} = \tau_x^{(2)} = \tau_x^{(3)}$
	$\kappa^{(1)} = \kappa^{(3)}$

APPENDIX B - 2

MODEL SPECIFACTION : IMPACT OF RESPONSE STYLES ON A SUBSTANTIVE MODEL

$$r^{(g)} = \tau_r^{(g)} + \Lambda_{r\rho}^{(g)} \rho^{(g)} + \delta_r^{(g)}$$

$$x^{(g)} = \tau_x^{(g)} + \Lambda_{x\xi}^{(g)} \xi^{(g)} + \Lambda_{x\rho}^{(g)} \rho^{(g)} + \delta_x^{(g)}$$

$$y^{(g)} = \tau_y^{(g)} + \Lambda_{y\eta}^{(g)} \eta^{(g)} + \Lambda_{y\rho}^{(g)} \rho^{(g)} + \varepsilon^{(g)}$$

$$\eta^{(g)} = \alpha^{(g)} + B^{(g)} \eta^{(g)} + \Gamma^{(g)} \xi^{(g)} + \zeta^{(g)}$$

where

g refers to groups (1) paper and pencil, (2) telephone and (3) online; $r^{(g)}$ is a 6*1 vector with response style indicators; $\tau_r^{(g)}$ is a 6*1 vector with measurement intercepts for the response style indicators; $\Lambda_{r\rho}^{(g)}$ is a 6*2 vector with factor loadings regressing the response style indicators on the latent response styles; $\rho^{(g)}$ is a 2*1 vector with latent response style scores; $\delta_r^{(g)}$ is a 6*1 vector with residuals for the response style indicators; $x^{(g)}$ is a 3*1 vector with observed independent scores; $\tau_x^{(g)}$ is a 3*1 vector with measurement intercepts; $\Lambda_{x\xi}^{(g)}$ is a 3*1 matrix with factor loadings; $\xi^{(g)}$ is a 1*1 vector with independent latent scores; $\Lambda_{x\rho}^{(g)}$ is a 3*1 matrix with factor loadings regressing the x-scores on the latent response styles; $\delta_x^{(g)}$ is a 3*1 vector with residuals; $y^{(g)}$ is a 3*1 vector with observed dependent scores; $\tau_y^{(g)}$ is a 3*1 vector with measurement intercepts; $\Lambda_{y\eta}^{(g)}$ is a 3*1 vector with factor loadings; $\eta^{(g)}$ is a 1*1 vector with latent dependent scores; $\Lambda_{y\rho}^{(g)}$ is a 3*1 matrix with factor loadings regressing the y-scores on the latent response styles; ε is a 3*1 vector with residuals; $\alpha^{(g)}$ is a 1*1 vector with structural intercepts; $B^{(g)}$ is a 1*1 vector with regression weight 0; $\Gamma^{(g)}$ is a 1*1 vector with regression weights; $\zeta^{(g)}$ is a 1*1 vector with residuals. Moreover, means of $\xi^{(g)}$ are represented by $\kappa^{(g)}$; $\phi^{(g)}$ is a 1*1 vector representing the variance of $\xi^{(g)}$; $\Psi^{(g)}$ is a 1*1 vector representing the variance of $\eta^{(g)}$. In all models and all groups, one loading per factor is fixed to one. Also, the intercepts of one ISM indicator and one ISSD indicator are fixed to zero in all models and all groups. Moreover, we fix the TRUST and LOYALTY factor means to zero in the paper and pencil group. This makes it easy to interpret the mean differences in the corrected and the uncorrected model in similar ways, by assessing the critical ratio's (estimate / standard error) of the other groups means.

Model	Constrained parameters
A. Base model:	$\tau_r^{(1)} = \tau_r^{(2)} = \tau_r^{(3)}$; $\Lambda_{r\rho}^{(1)} = \Lambda_{r\rho}^{(2)} = \Lambda_{r\rho}^{(3)}$;
B. Metric invariance:	$\tau_r^{(1)} = \tau_r^{(2)} = \tau_r^{(3)}$; $\Lambda_{r\rho}^{(1)} = \Lambda_{r\rho}^{(2)} = \Lambda_{r\rho}^{(3)}$;
	$\Lambda_{x\xi}^{(1)} = \Lambda_{x\xi}^{(2)} = \Lambda_{x\xi}^{(3)}$; $\Lambda_{y\eta}^{(1)} = \Lambda_{y\eta}^{(2)} = \Lambda_{y\eta}^{(3)}$
C. Scalar invariance:	$\tau_r^{(1)} = \tau_r^{(2)} = \tau_r^{(3)}$; $\Lambda_{r\rho}^{(1)} = \Lambda_{r\rho}^{(2)} = \Lambda_{r\rho}^{(3)}$;
	$\tau_x^{(1)} = \tau_x^{(2)} = \tau_x^{(3)}$; $\tau_y^{(1)} = \tau_y^{(2)} = \tau_y^{(3)}$;
	$\Lambda_{x\xi}^{(1)} = \Lambda_{x\xi}^{(2)} = \Lambda_{x\xi}^{(3)}$; $\Lambda_{y\eta}^{(1)} = \Lambda_{y\eta}^{(2)} = \Lambda_{y\eta}^{(3)}$
D. Structural means/intercepts:	$\tau_r^{(1)} = \tau_r^{(2)} = \tau_r^{(3)}$; $\Lambda_{r\rho}^{(1)} = \Lambda_{r\rho}^{(2)} = \Lambda_{r\rho}^{(3)}$;
	$\tau_x^{(1)} = \tau_x^{(2)} = \tau_x^{(3)}$; $\tau_y^{(1)} = \tau_y^{(2)} = \tau_y^{(3)}$;
	$\Lambda_{x\xi}^{(1)} = \Lambda_{x\xi}^{(2)} = \Lambda_{x\xi}^{(3)}$; $\Lambda_{y\eta}^{(1)} = \Lambda_{y\eta}^{(2)} = \Lambda_{y\eta}^{(3)}$;
	$\kappa^{(1)} = \kappa^{(2)} = \kappa^{(3)}$; $\alpha^{(1)} = \alpha^{(2)} = \alpha^{(3)}$
F. Structural regression invariance:	$\tau_r^{(1)} = \tau_r^{(2)} = \tau_r^{(3)}$; $\Lambda_{r\rho}^{(1)} = \Lambda_{r\rho}^{(2)} = \Lambda_{r\rho}^{(3)}$;
	$\tau_r^{(1)} = \tau_r^{(2)} = \tau_r^{(3)}$; $\Lambda_{r\rho}^{(1)} = \Lambda_{r\rho}^{(2)} = \Lambda_{r\rho}^{(3)}$;
	$\tau_x^{(1)} = \tau_x^{(2)} = \tau_x^{(3)}$; $\tau_y^{(1)} = \tau_y^{(2)} = \tau_y^{(3)}$;
	$\Lambda_{x\xi}^{(1)} = \Lambda_{x\xi}^{(2)} = \Lambda_{x\xi}^{(3)}$; $\Lambda_{y\eta}^{(1)} = \Lambda_{y\eta}^{(2)} = \Lambda_{y\eta}^{(3)}$;
	$\kappa^{(1)} = \kappa^{(2)} = \kappa^{(3)}$; $\alpha^{(1)} = \alpha^{(2)} = \alpha^{(3)}$;
	$\phi^{(1)} = \phi^{(2)} = \phi^{(3)}$; $\Psi^{(1)} = \Psi^{(2)} = \Psi^{(3)}$;
	$\Gamma^{(1)} = \Gamma^{(2)} = \Gamma^{(3)}$

TABLES

TABLE 1

FIT INDICES FOR NESTED MODELS TESTING MEASUREMENT AND STRUCTURAL INVARIANCE (STUDY 1 AND STUDY 2)

	Model	χ^2	df	p	χ^2 diff	df diff	p	TLI	CFI	RMSEA	reference model
Study 1	A. Unconstrained	158.57	150	0.300				0.996	0.998	0.013	
	B. Metric invariance	182.85	170	0.237	24.28	20	0.231	0.995	0.997	0.015	A
	C. Scalar invariance	232.00	190	0.020	49.15	20	0.000	0.986	0.992	0.025	B
	D. Structural means invariance	319.84	200	0.000	87.85	10	0.000	0.962	0.976	0.042	C
	E. Partial structural means invariance	241.95	195	0.012	9.95	5	0.077	0.985	0.990	0.026	C
Study 2	A. Unconstrained	288.32	150	0.000				0.989	0.995	0.026	
	B. Metric invariance	326.87	170	0.000	38.54	20	0.008	0.989	0.994	0.026	A
	C. Scalar invariance	414.58	190	0.000	87.72	20	0.000	0.986	0.992	0.030	B
	D. Structural means invariance	524.68	200	0.000	110.10	10	0.000	0.981	0.988	0.035	C
	E. Partial structural means invariance	431.04	195	0.000	16.46	5	0.006	0.986	0.991	0.030	C

df=degrees of freedom; χ^2 diff = χ^2 difference test; df diff=degrees of freedom of the χ^2 difference test

TABLE 2

STANDARDIZED FACTOR LOADINGS AND COMPOSITE RELIABILITIES OF THE RESPONSE STYLE MEASURES (STUDY 1 AND STUDY 2)

	Study 1					Study 2				
	ARS	DARS	ERS	MPR	RR	ARS	DARS	ERS	MPR	RR
Standardized factor loadings	0.57	0.65	0.73	0.59	0.65	0.75	0.69	0.90	0.70	0.88
	0.66	0.68	0.75	0.63	0.62	0.79	0.75	0.91	0.73	0.83
	0.79	0.85	0.81	0.80	0.75	0.79	0.71	0.92	0.76	0.85
Composite reliability	0.72	0.77	0.81	0.72	0.72	0.82	0.76	0.94	0.78	0.89

ARS = acquiescence response style; DARS = disacquiescence response style; ERS = extreme response style; MPR = midpoint responding; RR = response range. Note: the estimates are based on the metric invariant model (model B in Table 2) in the paper and pencil group

TABLE 3

LATENT MEANS IN THE PARTIAL STRUCTURAL MEAN INVARIANCE MODEL

(STUDY 1 AND STUDY 2)

	Study 1				Study 2			
	paper & pencil and online*		Telephone		paper & pencil and online*		Telephone	
	mean	s.e.	mean	s.e.	mean	s.e.	mean	s.e.
ARS	27.90	1.84	**32.39**	2.18	32.33	0.36	**35.64**	0.53
DARS	15.66	1.36	**23.66**	1.76	21.65	0.30	**22.42**	0.39
ERS	24.37	2.01	**33.96**	2.67	27.90	0.65	**28.45**	1.01
MPR	**17.64**	1.63	7.84	1.18	**20.31**	0.43	15.16	0.50
RR	12.64	1.07	**15.36**	1.18	14.48	0.10	**14.85**	0.13

* Means of the paper & pencil and online groups were fixed to equality based on the invariance tests.

The highest value per study/response style is printed in boldface;

ARS = acquiescence response style; DARS = disacquiescence response style; ERS = extreme response style; MPR = midpoint responding; RR = response range;

s.e. = standard error of the mean estimate

TABLE 4

MODEL FIT INDICES FOR LOYALTY DIAD MODELS, CORRECTED / NOT CORRECTED FOR ISM/ISSD (STUDY 2)

	Model	χ^2	df	p	χ^2 diff	df diff	p diff	TLI	CFI	RMSEA
Uncorrected	Base model	48.6	24	0.002				0.990	0.995	0.028
	Metric invariance	64.4	32	0.001	15.8	8	0.045	0.991	0.993	0.028
	Scalar invariance	91.4	40	0.000	27.0	8	0.001	0.988	0.989	0.031
	Structural means	158.3	44	0.000	66.9	4	0.000	0.976	0.976	0.045
	Structural regression invariance	181.1	50	0.000	22.8	6	0.001	0.976	0.973	0.045
Corrected	Base model	236.3	127	0.000				0.979	0.986	0.026
	Metric invariance	250.2	135	0.000	13.9	8	0.083	0.979	0.985	0.026
	Scalar invariance	259.3	143	0.000	9.0	8	0.338	0.980	0.985	0.025
	Structural mean and intercept invariance	263.7	147	0.000	4.4	4	0.354	0.980	0.985	0.025
	Structural regression invariance	285.1	153	0.000	21.4	6	0.002	0.978	0.983	0.026

df=degrees of freedom; χ^2 diff = χ^2 difference test statistic; df diff=degrees of freedom of the χ^2 difference test

TABLE 5

FACTOR LOADINGS OF THE LOYALTY DIAD MODELS CORRECTED / NOT CORRECTED FOR ISM/ISSD (STUDY 2)

	Observed variable	Latent variable	Uncorrected stdd focal loading	Corrected stdd focal loading	Bias	ISM stdd loading	ISSD stdd loading
P&P	TRUST1	TRUST	0.88	0.84	5%	0.15	0.21
	TRUST2		0.85	0.83	3%	0.15	0.15
	TRUST3		0.81	0.75	8%	0.28	0.15
	LOYAL1	LOYAL	0.85	0.82	3%	0.21	*0.05*
	LOYAL2		0.77	0.72	8%	0.32	*0.02*
	LOYAL3		0.85	0.82	3%	0.17	0.13
Tele	TRUST1	TRUST	0.88	0.79	12%	0.34	0.15
	TRUST2		0.82	0.74	11%	0.31	0.14
	TRUST3		0.75	0.65	15%	0.30	0.17
	LOYAL1	LOYAL	0.89	0.80	11%	0.33	0.15
	LOYAL2		0.71	0.60	18%	0.42	*0.00*
	LOYAL3		0.85	0.77	10%	0.28	0.21
Online	TRUST1	TRUST	0.93	0.91	3%	0.25	*-0.03*
	TRUST2		0.86	0.82	5%	0.26	*-0.01*
	TRUST3		0.84	0.79	6%	0.28	*-0.01*
	LOYAL1	LOYAL	0.89	0.83	7%	0.32	*0.03*
	LOYAL2		0.79	0.71	12%	0.37	*0.02*
	LOYAL3		0.90	0.85	6%	0.30	*0.00*

Loadings printed in italics are non-significant on the .05-level; P&P = Paper and pencil; Tele = telephone; Online = online panel; Uncorrected stdd focal loading = standardized factor loading on the focal latent variable (TRUST or LOYAL) in the uncorrected model; Corrected stdd focal loading = standardized factor loading on the focal latent variable (TRUST or LOYAL) in the corrected model; Bias = [(Uncorrected stdd focal loading - Corrected stdd focal loading)/ Corrected stdd focal loading]; ISM/ISSD stdd loading = standardized loading of the indicator on the latent ISM/ISSD variable;

Table 6a

STRUCTURAL MEANS / INTERCEPTS OF THE LOYALTY DIAD WITH AND WITHOUT CORRECTION FOR ISM/ISSD (STUDY 2)

		Uncorrected			Corrected			Bias
		Mean / intercept	s.e.	C.R.	Mean / intercept	s.e.	C.R.	C.R. points
P&P	TRUST	0.00			0.00			
	LOYAL	0.00			0.00			
Tele	TRUST	0.57	0.08	7.09	-2.05	1.66	-1.24	8.33
	LOYAL	-0.09	0.08	-1.18	-0.71	1.53	-0.47	-0.71
Online	TRUST	0.08	0.08	1.03	0.06	1.85	0.04	0.99
	LOYAL	-0.10	0.07	-1.42	-2.44	1.71	-1.42	0.00

TABLE 6B

STRUCTURAL VARIANCES OF THE LOYALTY DIAD WITH AND WITHOUT CORRECTION FOR ISM/ISSD (STUDY 2)

		Uncorrected			Corrected			Bias
		Var.	s.e.	C.R.	Var.	s.e.	C.R.	%
P&P	TRUST	1.20	0.10	12.16	1.11	0.10	11.69	8%
	LOYAL	0.95	0.09	10.58	0.94	0.09	10.37	1%
Tele	TRUST	1.28	0.10	12.62	1.05	0.09	11.63	23%
	LOYAL	0.56	0.06	9.14	0.55	0.06	9.24	0%
Online	TRUST	1.06	0.08	12.89	0.99	0.08	12.25	7%
	LOYAL	0.67	0.06	10.68	0.64	0.06	10.31	5%

TABLE 6C

STRUCTURAL REGRESSION WEIGHTS OF THE LOYALTY DIAD WITH AND WITHOUT CORRECTION FOR ISM/ISSD (STUDY 2)

		Uncorrected				Corrected				Bias	
		Unstdd	s.e.	C.R.	Stdd	Unstdd	s.e.	C.R.	Stdd	% Unstdd	% Stdd
P&p	LOYAL on TRUST	0.61	0.05	11.38	0.57	0.58	0.06	10.14	0.53	6%	7%
tele	LOYAL on TRUST	0.79	0.04	17.73	0.77	0.72	0.05	14.30	0.70	10%	9%
Online	LOYAL on TRUST	0.79	0.05	16.36	0.70	0.72	0.05	14.41	0.67	9%	5%

P&P = Paper and pencil; Tele = telephone; Panel = online panel; Unstdd = Unstandardized parameter estimate; Stddd = Standardized parameter estimate; s.e.= standard error of the estimate; C.R.= critical ratio (Unstandardized estimate/s.e.); Uncorrected / Corrected = estimated in the model not including / including the ISM and ISSD factors; Bias = [(Uncorrected parameter estimate – corrected parameter estimate)/Corrected parameter estimate]; All estimates are based on the scalar invariant model. The means/intercepts are estimated by fixing the P&P means/intercepts to zero while estimating all TRUST and LOYAL indicator intercepts as free but group invariant parameters;

TABLE A-1: NUMBER OF ITEMS USED IN RESPONSE STYLE INDICATORS (FOR ISM) AND THEIR SHARED VARIANCE COMPONENTS

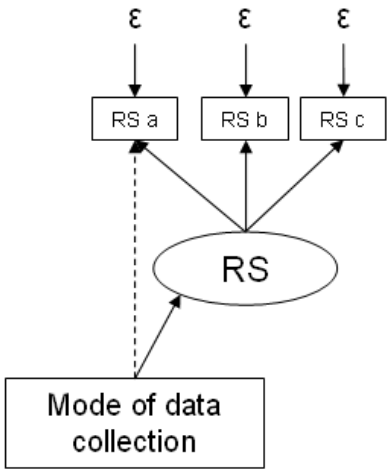
	average inter-item covariance .07			average inter-item covariance .12		
n items (a)	b	b/(a+b)	lambda	b	b/(a+b)	lambda
1	0.0	0.00	0.00	0.00	0.00	0.00
5	1.4	0.22	0.47	2.40	0.32	0.57
10	6.3	0.39	0.62	10.80	0.52	0.72
15	14.7	0.49	0.70	25.20	0.63	0.79
20	26.6	0.57	0.76	45.60	0.70	0.83
25	42.0	0.63	0.79	72.00	0.74	0.86
30	60.9	0.67	0.82	104.40	0.78	0.88
35	83.3	0.70	0.84	142.80	0.80	0.90
40	109.2	0.73	0.86	187.20	0.82	0.91
45	138.6	0.75	0.87	237.60	0.84	0.92
50	171.5	0.77	0.88	294.00	0.85	0.92

a stands for $\Sigma \sigma_i^2$, b stands for $\Sigma\Sigma \sigma_{ij}$, b/(a+b) refers to the common variance divided by the total variance, and lambda stands for the estimated standardized loading assuming identically loading indicators of one response style factor

FIGURES

Figure 1

Multiple indicators and scalar invariance



RS = response style; RS a through c = response style indicators. The dotted arrow indicates scalar non-invariance of RS a.

Figure 2

MACS for mean comparison Study 1 and Study 2

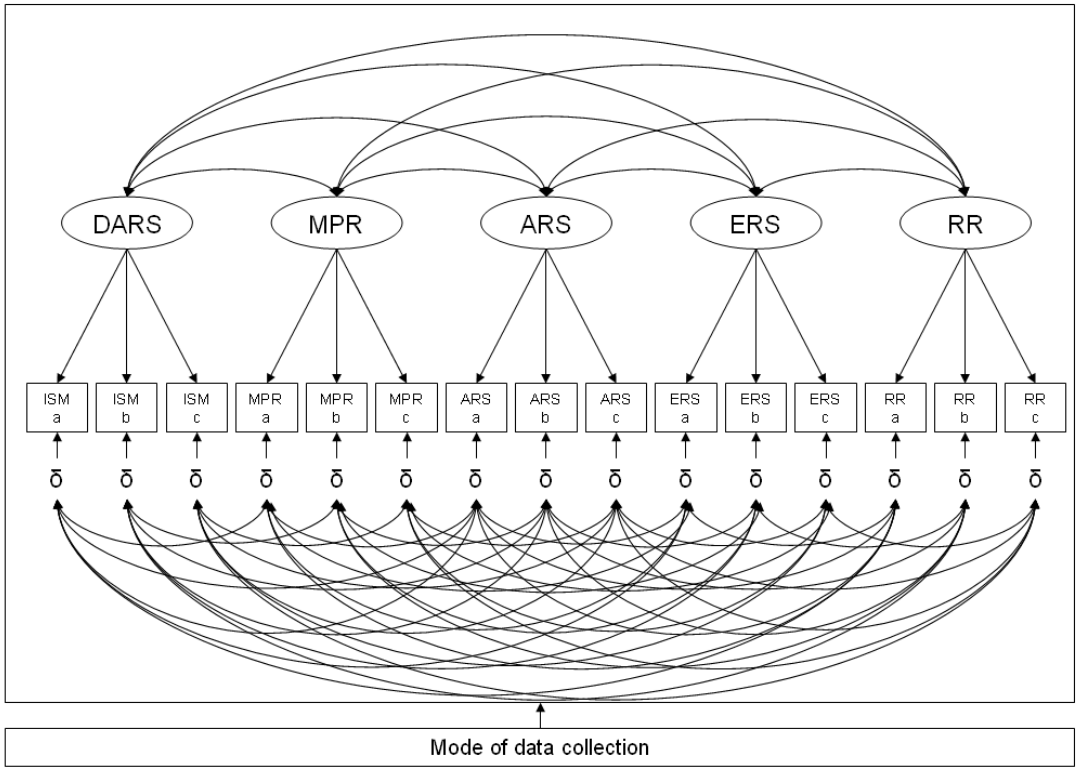
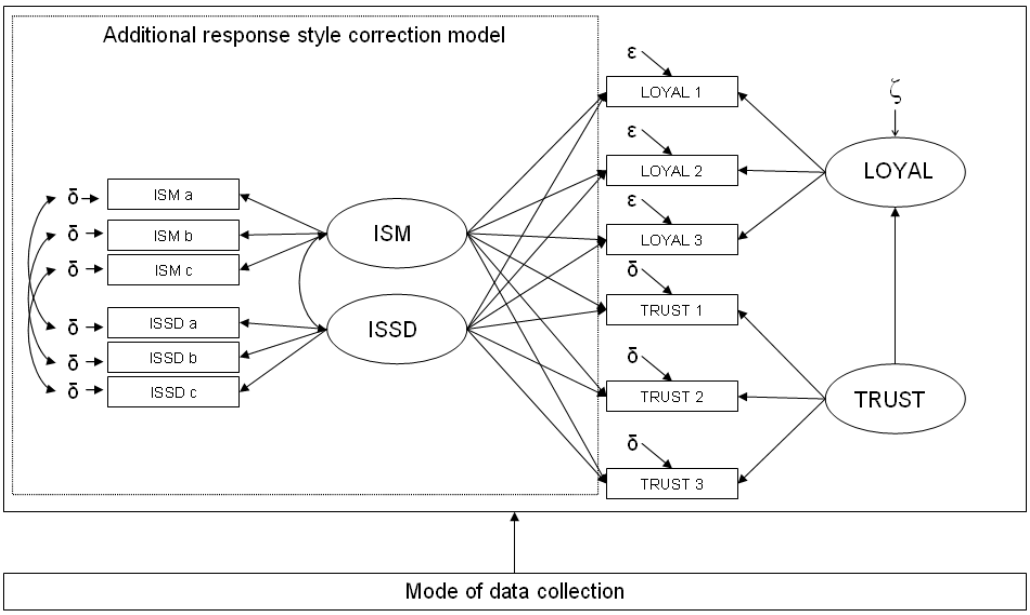


Figure 3

Loyalty diad corrected for ISM/ISSD



REFERENCES

Baumgartner, Hans and Jan-Benedict E.M. Steenkamp (2001), "Response Styles in Marketing Research: a Cross-national Investigation," *Journal of Marketing Research*, 38 (May), 143-156.

Bruner, Gordon C., Karen E. James, and Paul J. Hensel (2001), "*Marketing Scales Handbook, A Compilation of Multi-Item Measures*", Volume III. American Marketing Association, Chicago, Illinois USA.

Cheung, Gordon W., and Roger B. Rensvold (2000), "Assessing Extreme and Acquiescence Response Sets in Cross-Cultural Research Using Structural Equation Modeling," *Journal of Cross-cultural Psychology,* 31(2), 187-212.

Cheung, Gordon W., and Roger B. Rensvold (2002), "Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance," *Structural Equation Modeling*, 9(2), 233-255.

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

Deutskens, Elisabeth, Ko De Ruyter, Martin Wetzels, and Paul Oosterveld (2004), "Response Rate and Response Quality of Internet-Based Surveys: An Experimental Study," *Marketing Letters, 15(1), 21-36.*

Greenleaf, Eric A. (1992a), "Improving rating scale measures by detecting and correcting bias components in some response styles," *Journal of Marketing Research*, 29(May), 176-188.

Greenleaf, Eric A. (1992b), "Measuring Extreme Response Style," *Public Opinion Quarterly*, 56(3), 328-350.

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), 237-258.

Gunter, Barrie, David Nicholas, Paul Huntington and Peter Williams (2002), "Online versus Offline Research: Implications for Evaluating Digital Media", *Aslib Proceedings*, 54(4), 229-239.

Hu, Li-tze and Bentler, P.M. (1999), "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives," *Structural Equation Modeling*, 6, 1, 1-55.

Johnson, Eric J. (2001), "Digitizing Consumer Research," *Journal of Consumer Research*, 28(Sept), 331-336.

Jordan, Lawrence A., Alfred C. Marcus, and Leo G. Reeder (1980), "Response Styles in Telephone and Household Interviewing: A Field Experiment," *Public Opinion Quarterly*, 44(2), 210-222.

Jöreskog, Karl G. (1971), "*Simultaneous Factor Analysis in Several Populations,"* Psychometrika, 36 (Dec.), 409-426.

Kiesler, Sara, and Lee S. Sproul (1986), "Response Effects in the Electronic Survey", *Public Opinion Quarterly*, 50, 402-143.

Little, Todd D. (1997), "Mean and Covariance Structures (MACS) Analyses of Cross-Cultural Data: Practical and Theoretical Issues," *Multivariate Behavioral Research*, 32(1), 53-76.

Little, Todd D., William A. Cunningham, Golan Shahar, and Keith F. Widaman (2002), "To Parcel or Not to Parcel: Exploring the Question, Weighing the Merits," *Structural Equation Modeling*, 9(2), 151-173.

Marsh, Herbert W., John R. Balla, and Roderick McDonald (1988), "Goodness-of-Fit Indexes in Confirmatory Factor Analysis: The Effect of Sample Size," *Pscyhological Bulletin*, 103(3), 391-410.

Meredith, W. (1993), "Measurement invariance, factor analysis and factorial invariance," *Psychometrika*, 58, 525-543.

Mick, David Glen (1996), "Are studies of dark side variables confounded by socially desirable responding? The case of materialism," *Journal of Consumer Research*, 23(Sept), 106-119.

O'Neill, Harry W. (1967), "Response Style Influence in Public Opinion Surveys," *Public Opinion Quarterly,* 31(1), 137-157.

Ployhart, Robert E., Jeff A. Weekley, Brian C. Holtz, and Cary Kemp (2003), "Web-Based and Paper-and-Pencil Testing of Applicants in a Proctored Setting: are Personality, Biodata, and Situational Judgment Tests Comparable?" *Personnel Psychology*, 56(Autumn), 733-752.

Ployhart, Robert E., and Frederick L. Oswald (2004). "Applications of Means and Covariance Structure Analysis: Integrating Correlational and Experimental Approaches," *Organizational Research Methods*, 7(1), 27-65.

Podsakoff, Philip M., Scott B. MacKenzie, Jeong-Yeon Lee, and Nathan P. Podsakoff

(2003), "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies," *Journal of Applied Psychology,* 88(5), 879-903.

Simsek, Zeki, and John F. Veiga (2001), "A Primer on Internet Organizational Surveys," *Organizational Research Methods*, 4(July), 218-235.

Sirdeshmukh, Deepak, Jagdip Singh, and Barry Sabol (2002). "Consumer Trust, Value, and Loyalty in Relational Exchanges," *Journal of Marketing*, 66(1), 15-37.

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

Szymanski, David M., and Richard T. Hise (2000) "e-Satisfaction: An Initial Examination", *Journal of Retailing*, 76(3), 309-322.

Thompson, Lori Foster, Eric A. Surface, Don L. Martin, and Michael G. Sanders (2003), "From Paper to Pixels: Moving Personnel Surveys to the Web," *Personnel Psychology*, 56(Spring), 197-227.

Truell, Allen D. (2003), "Use of Internet Tools for Survey Research," *Information Technology, Learning, and Performance Journal*, 21(1), 31-37.

Vandenberg, Robert J., 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.

Venkatesh, Shankar, Amy K. Smith, and Arvind Rangaswamy (2003), "Customer Satisfaction and Loyalty in Online and Offline environments", *International Journal of Research in Marketing,* 20, 153- 175.

Watson, Dorothy (1992), "Correcting for Acquiescent Response bias in the Absence of a Balanced Scale: an Application to Class Consciousness", *Sociological Methods and Research*, 21(August), 52-88.