HANS BAUMGARTNER and JAN-BENEDICT E.M. STEENKAMP*

Response styles are a source of contamination in questionnaire ratings, and therefore they threaten the validity of conclusions drawn from marketing research data. In this article, the authors examine five forms of stylistic responding (acquiescence and disacquiescence response styles, extreme response style/response range, midpoint responding, and noncontingent responding) and discuss their biasing effects on scale scores and correlations between scales. Using data from large, representative samples of consumers from 11 countries of the European Union, the authors find systematic effects of response styles on scale scores as a function of two scale characteristics (the proportion of reverse-scored items and the extent of deviation of the scale mean from the midpoint of the response scale) and show that correlations between scales can be biased upward or downward depending on the correlation between the response style components. In combination with the apparent lack of concern with response styles evidenced in a secondary analysis of commonly used marketing scales, these findings suggest that marketing researchers should pay greater attention to the phenomenon of stylistic responding when constructing and using measurement instruments.

Response Styles in Marketing Research: A Cross-National Investigation

It is generally assumed that when people fill out a questionnaire, their answers are based on the substantive meaning of the items to which they respond. However, it has been known for a long time that people's responses are also influenced by content-irrelevant factors, such as the rating scale associated with an item (Cronbach 1946; Lentz 1938). These non–content-based forms of responding are usually referred to as response styles, response sets, or response biases. Following Paulhus (1991, p. 17), response styles may be defined as tendencies to respond systematically to questionnaire items on some basis other than what the items were specifically designed to measure. Common examples of response styles are acquiescence, extreme responding, use of the middle response category on ratings scales, and socially desirable responding.[1]

Response styles are a source of concern in both domestic and international marketing research because they threaten the validity of empirical findings by contaminating respondents' answers to substantive questions (Craig and Douglas 2000; Greenleaf 1992a; Van de Vijver and Leung 1997). Despite their potentially biasing effects, response styles have not attracted much attention in the marketing literature (notable recent exceptions include Greenleaf [1992a] and Mick [1996]). This may be due to researchers' lack of familiarity with the large number of response styles that have

*Hans Baumgartner is Associate Professor of Marketing, The Mary Jean and Frank P. Smeal College of Business Administration, Pennsylvania State University (e-mail: jxb14@psu.edu). Jan-Benedict E.M. Steenkamp is CentER Research Professor of Marketing and GfK Professor of International Marketing Research, Tilburg University (e-mail: J.B.Steenkamp@kub.nl). The order of authorship is arbitrary; both authors contributed equally to this research. The authors thank the three anonymous *JMR* reviewers for their help during the review process.

[1]Some authors distinguish response styles from response sets (e.g., Rorer 1965; Watkins and Cheung 1995). A response style refers to a tendency to respond to questionnaire items independently of item content, whereas a response set indicates people's desire to give a particular picture of themselves by the way they respond to questionnaire items (which implies that the responses given depend on the item content). The five response biases studied in this article are examples of response styles, whereas social desirability is the most common response set. In this article, we use the terms "response style" and "response set" interchangeably, because the distinction is not widely accepted and the terms are used in different senses (Paulhus 1991). Moreover, the effect of response styles and response sets is the same, namely, a method bias in questionnaire ratings.

been identified (which often go by different names and are assessed in different ways), uncertainty about how to deal with stylistic responding, or the belief that response styles generally do not have deleterious effects on the validity of conclusions drawn from marketing research data.

The purpose of this article is to increase marketing researchers' concerns about the phenomenon of stylistic responding. After briefly reviewing seven common response styles, we present data, based on a secondary analysis of a large number of commonly used measurement scales, that show that many instruments have apparently been developed without adequate regard for the problems caused by stylistic responding. We then discuss two key adverse effects of stylistic responding (see Bagozzi 1994). First, response styles can contaminate observed responses by either inflating or deflating respondents' scores on measurement instruments. We develop a model that considers the influence of the different response styles on scale scores and suggest two scale characteristics that moderate the biasing effect of stylistic responding. Second, response styles can affect conclusions about the relationship between scales by either inflating or deflating the correlation between respondents' scores on measurement instruments. These correlations constitute the basis for most multivariate techniques. We develop a second model that investigates the correspondence between two kinds of correlations—correlations between observed scores and correlations between scores that have been purified of response style contaminations—and discuss under what conditions observed correlations will be inflated or deflated compared with the true correlation between substantive scales. We test the two models in a cross-national context using a unique sample of more than 10,000 nationally representative respondents from 11 countries of the European Union, which enables us to arrive at empirical generalizations (Bass and Wind 1995). Our basic conclusion is that ratings on substantive scales and relations between scales can be strongly contaminated by stylistic responding and that this can lead to seriously biased conclusions. It is thus necessary that marketing researchers pay greater attention to the phenomenon of response styles when constructing and using measurement instruments.

TYPES OF RESPONSE STYLES AND MARKETING RESEARCHERS' CONCERN WITH RESPONSE BIASES

Seven important response styles will be considered: acquiescence response style (ARS), disacquiescence response style (DARS), net acquiescence response style (NARS), extreme response style (ERS), response range (RR), midpoint responding (MPR), and noncontingent responding (NCR). Although several other response styles have been identified (see Broen and Wirt 1958; Messick 1968), these seven, with the exception of social desirability, are the ones most commonly encountered in the literature. The response styles considered can all be computed and examined from consumers' responses on the substantive scales in question.

The seven response styles are summarized in Table 1. Although Table 1 is more or less self-explanatory, several comments are in order. First, net acquiescence is listed as a separate response style because some authors (e.g., Greenleaf 1992a; Hui and Triandis 1985) have used it in their work. However, net acquiescence is simply a summary measure based on acquiescence and disacquiescence, and because these response styles are considered separately in this article, we do not discuss it further. Moreover, although in principle ERS and RR need not be highly correlated (e.g., if someone gives only extreme positive responses, his or her RR would be very small), research has shown that people who tend to give extreme positive responses also tend to give extreme negative responses (see Bachman and O'Malley 1984; Hamilton 1968). Therefore, the two response styles are expected to be closely related (a hypothesis that is confirmed in the empirical section), and we combine them in the remainder of the article.

Second, the theoretical explanations of stylistic responding have been of either the dispositional or situational variety. Dispositional explanations attribute stylistic responding to characteristics of the individual (e.g., extremity as a personality trait associated with intolerance of ambiguity and dogmatism); situational explanations attribute it to characteristics of the situation (e.g., acquiescence as a consequence of the ambiguity of items). The position adopted here is that stylistic responding is best understood as an interaction of personal dispositions and situational factors (see Snyder and Ickes 1985). That is, people differ in their inherent tendency to engage in stylistic responding, but this tendency may be encouraged or discouraged by situational determinants such as the issue about which respondents are queried, the way the questions are asked, the format of the response scale, or more general characteristics of the context in which the data are collected (e.g., time pressure).

Third, the major problem in measuring response styles is not to confound stylistic variance with substantive variance. The dominant approach to dealing with this issue has been to assess stylistic responding across many different items that are heterogeneous in content. Heterogeneity in content means that the entire set of items on which the response style measure is based does not refer to a substantively meaningful psychological construct and is psychologically diffuse. Operationally, heterogeneous items are selected by using items from a diverse set of scales that have little in common (see Couch and Keniston 1960). The assumption is that a respondent may agree or disagree with, respond extremely to, or endorse the midpoint on some items, but it is unlikely that across many items varying in content, the respondent's true position is characterized by consistent agreement or disagreement, extremity, or MPR. The major weakness of this approach is that if the items are substantively correlated, stylistic and substantive variance will be confounded. The second approach, applicable to acquiescent and disacquiescent responding, is to word some items in a scale positively (i.e., agreement with the question indicates a higher score on the underlying construct) and other items negatively. If an instrument has an equal number of positively and negatively keyed items, the scale is called balanced. The advantage of balanced scales is that they have a built-in control for stylistic responding because a high (low) score cannot be obtained simply because of yea-saying (nay-saying).

Casual reading of scale development articles and the measurement section of empirical articles suggests that response styles are frequently not seen as a major threat to validity by marketing researchers, because no mention is made that stylistic responding was checked or controlled for. In an effort to get more detailed insights into this issue, we

Table 1
SUMMARY OF COMMON RESPONSE STYLES

Response Style	*Definition and Synonyms*	*Theoretical Explanations*	*Measurement*
ARS[12, 19]	The tendency to agree with items regardless of content. Also called agreement tendency, yea-saying, or positivity.	•Characteristic of stimulation-seeking extroverts who have a tendency to accept statements impulsively.[3, 16] •Due to uncritical endorsement of statements by respondents who are low in cognitive abilities or have low status.[9, 16, 20] •More common for items that are ambiguous, vague, or neutral in desirability or for issues about which respondents are uncertain.[14, 18, 19] •Most likely when respondents lack adequate cognitive resources because of distraction, time pressure, and so forth.[13]	Two general approaches:[12] •Extent of agreement with many items that are heterogeneous in content (which requires that multiple scales that have little in common are available).[1, 10, 24] •Extent of agreement with both positively and negatively worded items within the same scale (before negatively worded items have been reverse-scored).[24] A special case is balanced-worded, logically inconsistent statements (i.e., pairs of items that are identical in substantive content, with one item worded positively and the other worded negatively).[26]
DARS[3]	The tendency to disagree with items regardless of content. Also called disagreement tendency, nay-saying, or negativity.	•Characteristic of controlled and reflective introverts who try to avoid external stimulation.[3]	Same as ARS, except that disagreement is assessed instead of agreement.
NARS[4, 7]	The tendency to show greater acquiescence than disacquiescence. Also called directional bias.	[See explanations for ARS and DARS.]	In general, acquiescence minus disacquiescence. Most commonly measured as the mean response across many heterogeneous items.[4, 7]
ERS[5]	The tendency to endorse the most extreme response categories regardless of content.	•Reflection of rigidity, intolerance of ambiguity, and dogmatism.[6] •Associated with higher levels of anxiety and possibly deviant behavior.[6] •Characteristic of respondents with less differentiated cognitive structures and poorly developed schemas.[21] •Greater for meaningful stimuli (i.e., stimuli that are important or involving to respondents).[17]	Number or proportion of heterogeneous items on which the respondent endorses the most extreme (positive or negative) scale categories.[1, 2, 7, 10, 22] Greenleaf (1992b) suggests that the items should be uncorrelated and have equal extreme response proportions. In addition, the mean response to an item should be close to the midpoint of the scale.
RR[4, 7, 27]	The tendency to use a narrow or wide range of response categories around the mean response.	[Presumably similar to the explanations for ERS.]	Standard deviation of a person's responses across many heterogeneous items.[4, 7, 27]
MPR[15, 20]	The tendency to use the middle scale category regardless of content.	•Due to evasiveness (desire not to reveal one's true opinion), indecision (uncertainty about one's position), or indifference (lack of interest in an issue).[15, 20]	Number or proportion of heterogeneous items on which the respondent endorses the middle scale category.[2, 22]
NCR[11, 23]	The tendency to respond to items carelessly, randomly, or nonpurposefully.	•Due to lack of motivation to read the instructions and interpret items appropriately.[8]	Sum of absolute differences between responses to pairs of items, where the items in each pair are maximally correlated, have similar means across respondents, and are keyed in the same direction.[11, 23]

Notes: [1]Bachman and O'Malley (1984); [2]Chen, Lee, and Stevenson (1995); [3]Couch and Keniston (1960); [4]Greenleaf (1992a); [5]Greenleaf (1992b); [6]Hamilton (1968); [7]Hui and Triandis (1985); [8]Jackson (1967); [9]Knowles and Nathan (1997); [10]Marín, Gamba, and Marín (1992); [11]Marsh (1987); [12]Martin (1964); [13]McGee (1967); [14]Messick (1967); [15]Messick (1968); [16]Messick (1991); [17]O'Donovan (1965); [18]Paulhus (1991); [19]Ray (1983); [20]Schuman and Presser (1981); [21]Shulman (1973); [22]Stening and Everett (1984); [23]Watkins and Cheung (1995); [24]Watson (1992); [25]Wells (1963); [26]Winkler, Kanouse, and Ware (1982); [27]Wyer (1971).

performed a secondary analysis of the measurement instruments described in Bearden and Netemeyer's (1999) influential *Handbook of Marketing Scales*. This handbook offers a comprehensive collection of measurement scales from many different areas of marketing, and the authors carefully describe the scoring procedures necessary for administering these scales and summarize evidence of scale validity, including checks for response styles. The findings confirmed our suspicion that many scales commonly used in marketing apparently fail to control adequately for response styles.

Across all scales analyzed,[2] on average 18% of items are reverse-keyed (the median is 11%). Of the scales, 43% do

[2]In total, we analyzed 200 scales. This number differs slightly from the number of scales listed by Bearden and Netemeyer (1999) for two reasons. First, a few scales could not be considered in the present analysis because they are not reproduced in the book and the information provided in the description of the scale was insufficient to include them. Second, some multidimensional scales are listed under the same heading in the book, but they measure different constructs (e.g., role conflict, role ambiguity), so they are treated as separate scales in our analysis.

not contain a single reverse-keyed item, and only 9% of scales are balanced. If the analysis is performed at the factor rather than the scale level (which is probably more relevant, because multidimensional scales frequently are not combined into an overall composite), on average 15% of items are reverse-scored (the median being 5%), 48% of scales contain no reverse-keyed items, and 7% of scales are balanced. In 12% of cases, the scale was correlated with a measure of social desirability, and in 2% another check of stylistic responding was reported. If a scale development article reported a check for social desirability bias, the scale also tended to have a higher proportion of reverse-keyed items ($r = .32$, $p < .001$). This suggests that authors who are concerned with one kind of response bias are also concerned with other types of biases.[3]

Overall, these findings suggest that marketing researchers may not be particularly concerned about response biases, particularly acquiescence and disacquiescence (because reverse-keying items is supposed to control for these response styles) or social desirability (because only in a small fraction of scale development efforts was a check for social desirability bias reported; see also Mick 1996). The question is whether this lack of concern is warranted—because stylistic responding generally does not have damaging effects on the validity of research findings—or whether it leads to misleading conclusions. In the next section, we address this question theoretically, and the empirical study provides evidence from a large-scale multinational survey.

CONTAMINATING EFFECTS OF RESPONSE STYLES

Stylistic responding is a source of measurement error, leading to deviations between a respondent's true score and the score actually observed in the administration of an instrument designed to measure the underlying construct. In general, the error is nonrandom, so that response styles not only contribute to observed score variance (which is the major adverse consequence of random measurement error) but also may have systematic effects on scale scores and produce correlated errors of measurement. Thus, two types of contaminating effects of response styles must be considered (see Bagozzi 1994): Response styles may bias the assessment of true scores by inflating or deflating observed scale scores, and they may bias the investigation of relationships between constructs by inflating or deflating a scale's correlation with other scales. We discuss each of these consequences of stylistic responding in turn.

Influence of Response Styles on Scale Scores

In general, a person's observed score on a scale is a function of three components: a true score, systematic measurement error, and random measurement error. Here, we are interested in stylistic responding as a source of nonrandom measurement error or, more specifically, in the systematic effects of particular response styles on scale scores. We propose a model that considers the influence of ARS and DARS, ERS (including RR), MPR, and NCR on observed scores, and we suggest moderating effects of two scale characteristics—the extent to which the scale is balanced and the extent to which the scale mean deviates from the midpoint—on the relationship between scale scores and response styles.

The model is specified for the case in which respondents from several countries respond to multiple scales. The effects of stylistic responding on observed scores are allowed to vary by scale and country (i.e., heterogeneity of response style effects is explicitly taken into account), which makes it possible to investigate moderating influences of scale characteristics on stylistic responding. The determinants of cross-country differences in response style effects are not modeled explicitly, but the model enables an assessment of the proportion of stylistic variance in observed scores attributable to scales and countries. The model is general in nature, because it includes the domestic market research context as a special case, is applicable to any situation in which response style effects differ among groups (e.g., it can be used to model subcultural differences in stylistic responding; see Marín, Gamba, and Marín 1992), and can be easily extended to incorporate determinants of across-group differences.

Let y_{ijk} be the observed score of individual i from country k on scale j. Assuming that scale scores are a linear function of the different response styles, the influence of response styles on scale scores can be expressed as follows:

$$(1)\quad y_{ijk} = \pi_{0jk} + \pi_{1jk}ARS_{ijk} + \pi_{2jk}DARS_{ijk} + \pi_{3jk}ERS_{ijk} + \pi_{4jk}MPR_{ijk} + \pi_{5jk}NCR_{ijk} + \varepsilon_{ijk},$$

where π_{0jk} is the intercept of the equation, π_{pjk} ($p = 1, \ldots, 5$) is the effect of the five response styles on y_{ijk}, and ε_{ijk} is a random effect representing the deviation of respondent i's scale score from the predicted score based on the response styles. Note that the intercept of the equation and the effects of the response styles on scale scores are allowed to vary by both scale and country. The systematic effect of each response style on scale scores is shown by the sign of π_{pjk}. If the coefficient is positive, stylistic responding inflates observed scores; if it is negative, stylistic responding deflates observed scores.

What are the likely effects of the different response styles on scale scores? For ARS and DARS, an important determinant of the direction and magnitude of the effect is the proportion of items in the scale that are worded positively (i.e., the items are keyed in such a way that agreement results in a higher scale score) or negatively (i.e., agreement with an item results in a lower score). In the case of acquiescence, observed scores should be inflated when most items in the scale are worded positively and deflated when most items are worded negatively. The higher the proportion of positively or negatively worded items, the stronger the effect of acquiescent responding on scale scores should be. If the scale is balanced (i.e., there is an equal number of positively and negatively worded items), adverse effects of acquiescence should be minimized, because in that case the opposing effects of acquiescent responding for positively and negatively worded items cancel each other out (Paulhus 1991).

[3]We also conducted the foregoing analyses for Likert-type scales only because some response styles (e.g., acquiescence) may be more prevalent for this kind of scale and researchers may be more likely to take safeguards against such biases in this case. Of the scales, 55% used a Likert-type response scale. The results were more encouraging but far from satisfactory. Overall, on average 20% of items were reverse-keyed, 34% of scales contained no reverse-keyed items, and 8% of scales were balanced. The corresponding figures for the analysis conducted at the factor level are 17%, 38%, and 6%, respectively. In 16% of cases, the scale was correlated with a measure of social desirability, and in another 3% another check for response bias was reported.

Thus, although balanced scales do not eliminate the occurrence of acquiescence per se, they contain a built-in control for contamination of observed scores by yea-saying, because the bias is upward for half of the items and downward for the other half.

For DARS, the predictions are the opposite of those for acquiescence. That is, scores should be deflated for positively worded items and inflated for negatively worded items. Formally, these hypotheses can be stated as follows:

$$\pi_{1jk} = \beta_{10k} + \beta_{11k}PROP_{jk} + r_{1jk} \quad (2)$$

and

$$\pi_{2jk} = \beta_{20k} + \beta_{21k}PROP_{jk} + r_{2jk}, \quad (3)$$

where $PROP_{jk}$ is the proportion of positively or negatively worded items in scale j for country k, β_{10k} and β_{20k} are intercept terms, β_{11k} and β_{21k} are the effects of PROP on π_{1jk} and π_{2jk}, and r_{1jk} and r_{2jk} are random effects. Note that the β_{pqk} are allowed to vary across countries. Specifically, we assume that the country-specific coefficients vary randomly around some mean γ_{pq}. This leads to the following specification for the β_{pqk}:

$$\beta_{10k} = \gamma_{10} + u_{10k}, \quad (4)$$

$$\beta_{11k} = \gamma_{11} + u_{11k}, \quad (5)$$

$$\beta_{20k} = \gamma_{20} + u_{20k}, \text{ and} \quad (6)$$

$$\beta_{21k} = \gamma_{21} + u_{21k}, \quad (7)$$

where the u_{pqk} are random effects denoting the deviation of country j's coefficient from the overall mean. If PROP is coded such that PROP = 0 for a balanced scale, PROP = +1 for a scale in which all items are worded positively, and PROP = −1 for a scale in which all items are worded negatively, with intermediate values for different proportions of positively or negatively worded items, our hypotheses are that $\gamma_{11} > 0$ and $\gamma_{21} < 0$. Note that these predictions imply an interactive effect of ARS/DARS and PROP on scale scores (this can be seen by substituting Equations 4–7 into 2–3 and Equations 2–3 into 1), such that the effect of ARS (DARS) becomes more positive (more negative) as the proportion of positively worded items increases and more negative (more positive) as the proportion of negatively worded items increases.

For ERS and MPR, the direction and magnitude of the effect of stylistic responding on observed scores is hypothesized to depend on the positive or negative deviation of the scale mean (i.e., the average score across respondents on the instrument in question) from the midpoint of the response scale (i.e., 3 on a 1–5 scale). Consider the case of MPR first. If the scale mean were exactly at the midpoint, MPR would not systematically influence scale scores, because for respondents with true scores below the midpoint the bias would be positive, and for those with true scores above the midpoint the bias would be negative and the two effects would cancel each other out. In contrast, if the scale mean were greater than the midpoint, most true scores would be above the midpoint and stylistic endorsement of the middle scale category on average should decrease scores, whereas if the scale mean were smaller than the midpoint, scores generally should be inflated. The greater the deviation of the scale mean from the midpoint of the response scale, the greater the effect of MPR on scale scores should be.

The opposite is expected for ERS. Although, in principle, endorsement of the most extreme scale categories could either inflate or deflate scale scores—extreme positive responding would inflate (deflate) scale scores for positively (negatively) worded items, and extreme negative responding would deflate (inflate) scale scores for positively (negatively) worded items—it is unlikely that respondents will completely ignore scale content. The finding that people who engage in extreme positive responding also engage in extreme negative responding (Bachman and O'Malley 1984; Hamilton 1968) is consistent with this assumption. If responses are at least partly based on content, ERS should bias scores in the direction of the deviation of the scale mean from the midpoint of the response scale. In other words, if the scale mean is above the midpoint, stylistic endorsement of the most extreme scale categories should make scores even more positive, because for most people the bias is upward. However, if the scale mean is below the midpoint, stylistic responding will make scores even more negative, because for most people the bias is downward. If the scale mean is at the midpoint of the response scale, the bias should not be consistent, because for half the sample it is upward and for the other half it is downward. Formally, these hypotheses can be stated as follows:

$$\pi_{3jk} = \beta_{30k} + \beta_{31k}DMP_{jk} + r_{3jk} \quad (8)$$

and

$$\pi_{4jk} = \beta_{40k} + \beta_{41k}DMP_{jk} + r_{4jk}, \quad (9)$$

where DMP_{jk} is the (positive or negative) deviation of country k's mean on scale j from the midpoint of the scale, β_{30k} and β_{40k} are intercept terms, β_{31k} and β_{41k} are the effects of DMP on π_{3jk} and π_{4jk}, and r_{3jk} and r_{4jk} are random effects. As before, the β_{pqk} are allowed to vary randomly across countries:

$$\beta_{30k} = \gamma_{30} + u_{30k}, \quad (10)$$

$$\beta_{31k} = \gamma_{31} + u_{31k}, \quad (11)$$

$$\beta_{40k} = \gamma_{40} + u_{40k}, \text{ and} \quad (12)$$

$$\beta_{41k} = \gamma_{41} + u_{41k}, \quad (13)$$

where the u_{pqk} are random effects. Our hypotheses are that $\gamma_{31} > 0$ and $\gamma_{41} < 0$. Again, these predictions imply an interactive effect of ERS/MPR and DMP on scale scores (which can be seen by substituting Equations 10–13 into 8–9 and Equations 8–9 into 1), such that the effect of ERS (MPR) becomes more positive (more negative) as the deviation of true scores from the midpoint of the scale becomes more positive and more negative (more positive) as the deviation of true scores from the midpoint becomes more negative.

For NCR, no a priori hypotheses can be specified about the likely effect of this response style on scale scores. Unlike the other response styles, NCR may be less of a source of systematic measurement error and may instead predominantly contribute to random measurement error, as when a respondent fills out a questionnaire randomly. If this is the case, NCR should not systematically increase or decrease observed scores.

To complete the specification of the model, π_{0jk} (the intercept of Equation 1) and π_{5jk} (the effect of NCR on scale scores) are assumed to vary randomly by scale and country around some overall mean:

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \quad (14)$$

and

$$\pi_{5jk} = \beta_{50k} + r_{5jk}, \quad (15)$$

with

$$\beta_{00k} = \gamma_{00} + u_{00k} \quad (16)$$

and

$$\beta_{50k} = \gamma_{50} + u_{50k}. \quad (17)$$

The net effect of stylistic responding on scale scores (in terms of inflation or deflation) depends on the balance of the effects of each of the response styles. For example, if the total response style variance in a scale is dominated by acquiescent responding, scores should be inflated if most items in the scale are worded positively. A summary measure of the overall influence of stylistic responding on scale scores across respondents, scales, and countries is given by the coefficient of determination for Equation 1, which indicates the percentage of variance in scale scores that is accounted for by stylistic responding. In addition, it is possible to consider scale- and country-specific indices of variance accounted for. This more detailed analysis is useful for getting insights into the degree of contamination of observed scores by scale and country and for identifying scales and countries that seem particularly susceptible to response biases. These issues are addressed in more detail in the empirical part of the article.

Influence of Response Styles on Relationships Between Scales

It is well known that random measurement error attenuates the correlation between scales. Less well known is that systematic measurement error can either inflate or deflate correlations and can lead to correlations that have incorrect signs (Green and Citrin 1994; Green, Goldman, and Salovey 1993). To understand this, consider two sets of observed scores y_j and $y_{j'}$ for scales j and j′ (where we have omitted the subscript for individual for simplicity). Assume that these observed scores are a linear function of true scores τ_j and $\tau_{j'}$, response style components s_j and $s_{j'}$ (more generally, a component due to systematic error), and random errors δ_j and $\delta_{j'}$. By definition, random measurement error is uncorrelated with all other variables. If we also assume that true scores and response style components are uncorrelated, the correlation between y_j and $y_{j'}$, $c_{yj,yj'}$, is given by

$$c_{yj,yj'} = \frac{\text{Cov}(\tau_j, \tau_{j'}) + \text{Cov}(s_j, s_{j'})}{\sqrt{\text{Var}(y_j)\text{Var}(y_{j'})}}. \quad (18)$$

Inflation or deflation of the correlation between scales j and j′ depends on whether (1) the true correlation between the two substantive scales is positive or negative and (2) the response style components affecting the scales are positively or negatively correlated.

Assume that the true correlation between two substantive scales is positive. If the response style components of the two scales are also positively correlated, the observed correlation will be inflated (i.e., more positive than it should be). Conversely, if the response style components are negatively correlated, this will lead to a countervailing influence on the substantive correlation, and the observed correlation will be deflated (i.e., less positive or even negative). In this case, the response styles act as suppressors.

Assume that the true correlation between two substantive scales is negative. If the response style components of the two scales are also negatively correlated, the observed correlation will again be inflated (i.e., more negative than it should be). Conversely, if the response style components are positively correlated, the observed correlation generally will be deflated (i.e., less negative or even positive, because the response styles act as suppressors).

Formally, these hypotheses can be expressed as follows: The context of interest is again a situation in which respondents from several countries respond to multiple scales. Let $c_{jj',k}$ be the correlation between observed scores on scales j and j′ in country k, where the correlation is computed across individuals. Similarly, let $cc_{jj',k}$ be the correlation between corrected scores on scales j and j′ in country k, where "corrected" means that the response style component has been removed from observed scores. Operationally, this is done by partialling out ARS, DARS, ERS, MPR, and NCR from scale scores in Equation 1. Finally, let $rc_{jj',k}$ be the correlation between the response style components of scales j and j′ in country k. To test our hypotheses, consider the following model:

$$c_{jj',k} = \beta_{0k} + \beta_{1k}cc_{jj',k} + \beta_{2k}rc_{jj',k} + r_{jj',k}, \quad (19)$$

where β_{0k} is an intercept term, and β_{1k} and β_{2k} are the effects of $cc_{jj',k}$ and $rc_{jj',k}$ on $c_{jj',k}$. Note that as before, the β_{pk} are not specified to be homogeneous across countries. Specifically, we allow the β_{pk} to vary randomly across countries around some mean γ_p:

$$\beta_{0k} = \gamma_0 + u_{0k}, \quad (20)$$

$$\beta_{1k} = \gamma_1 + u_{1k}, \text{ and} \quad (21)$$

$$\beta_{2k} = \gamma_2 + u_{2k}. \quad (22)$$

If the difference between observed and corrected correlations is negligible, γ_0 and γ_2 should be close to zero and γ_1 should be close to one. However, if the correlation between the response style components of scales j and j′ has a significant impact on the observed correlation, γ_2 will be nonzero. Specifically, controlling for the true correlation ($cc_{jj',k}$), the average effect of $rc_{jj',k}$ on $c_{jj',k}$ will be positive (i.e., $\gamma_2 > 0$). In other words, for a given true correlation, a positive correlation between the response style components will make the observed correlation more positive (less negative), and a negative correlation between the response style components will make the observed correlation more negative (less positive). The total contribution of correlated response style components to observed correlations can be assessed by the increment in the coefficient of determination when $rc_{jj',k}$ is added to Equation 2. Furthermore, because β_{1k} and β_{2k} are allowed to vary across countries, differences in the correspondence between observed and corrected correlations and between observed correlations and correlations of the response style components across countries can be investigated, and the contribution of correlated response style components to observed correlations can be determined by country. These issues are investigated in more detail in the empirical section of the article.

METHOD

Subjects

A large pan-European market research agency collected nationwide data from representative samples of respondents in 11 countries of the European Union (i.e., Belgium, Denmark, France, Germany, Great Britain, Greece, Ireland, Italy, the Netherlands, Portugal, and Spain). The agency collected data by mail survey, using the script panel of the market research agency in each country. The mean response rate was 72%. After deleting cases with missing data on the variables of interest, we had an effective sample size of 10,477 people. The number of respondents that provided complete data in each country varied between 869 for Italy and 1057 for France.

Questionnaire

The questionnaire included questions about various consumer behavior issues and respondents' media behavior and certain demographics, among other things not relevant for the purposes of this study. Of primary interest are 60 attitudinal statements measured on five-point Likert scales, with scale steps of 1 = "strongly disagree," 2 = "disagree," 3 = "neither agree nor disagree," 4 = "agree," and 5 = "strongly agree." The 60 items were designed to measure the following constructs: attitude toward advertising in general (Gaski and Etzel 1986), attitude toward the past (Holbrook and Schindler 1994), change seeking (Steenkamp and Baumgartner 1995), consumer ethnocentrism toward the home country and the European Union (Shimp and Sharma 1987), deal proneness (Lichtenstein, Netemeyer, and Burton 1995), environmental consciousness (Grunert and Juhl 1995; Maloney, Ward, and Braucht 1976), exploratory acquisition of products (Baumgartner and Steenkamp 1996), exploratory information seeking (Baumgartner and Steenkamp 1996), health consciousness (Oude Ophuis 1989), price consciousness (Lichtenstein, Bloch, and Black 1988), product involvement (Mittal and Lee 1989), and quality consciousness (Steenkamp 1989). Although all items were taken from validated scales, only a subset of the complete set of items was used in most cases because of time constraints. With a few exceptions, the items were listed randomly in the questionnaire.

Calculation of the Response Style Indices and Preliminary Analyses

Some of the response style indices are calculated at the scale level, so it is necessary to check the factor structure of the items. A principal components analysis across all the respondents showed that 14 factors had eigenvalues greater than one, and the 14-factor solution was readily interpretable. For the most part, the factors were consistent with a priori expectations. However, the items indicating consumer ethnocentrism with respect to the home country and the European Union formed a single factor, whereas the items measuring attitude toward the past and exploratory information seeking split into two factors each (consisting of two and three items, respectively). Therefore, we treat the 60 items as measures of 14 different constructs, instead of the 13 factors initially hypothesized. With four exceptions, all loadings on the target factor were greater than .5 (the average loading was .73, and the smallest was .29), and there were only five nontarget loadings that were greater than .3 (the largest was .42). The average absolute interfactor correlation was .13, and the highest correlation between two factors was .40. In Table 2, we list the 14 constructs, the number of items per construct, the number of negatively worded items per construct, and the coefficient alpha (the interitem correlation for two-item constructs) of each construct. In total, we computed eight different response style indices. We briefly describe how we calculated these indices and how they are related to one another.

ARS and DARS. Two (dis)acquiescence indices were computed, corresponding to the two methods by which this response style has been measured in the past (Table 1). The first method of assessing (dis)acquiescence is based on the extent to which respondents agree with items that are heterogeneous in content (Bachman and O'Malley 1984; Marín, Gamba, and Marín 1992). The assumption of heterogeneity of content was met in the present case, because the average absolute intercorrelation of all 60 items was only .12 (with a range from −.32 to .79). If respondents strongly agreed with an item, they received a score of 2, and if they merely agreed with an item, a score of 1 was assigned. These scores were then averaged across all 60 items to calculate ARS1. We assessed DARS1 in the same way, except that it is based on disagreement with items.

The second method assesses (dis)acquiescence at the scale level and requires a mix of positively and negatively worded items within the same scale (Watson 1992). This means that only 6 of the 14 scales could be used in the calculation of ARS2 and DARS2 (see Table 1). For each of the six scales, subjects' responses to a negatively worded item were compared with their responses to the positively worded items as follows: If respondents agreed with both a positively and a negatively worded item within the same scale, a score of 1 was assigned; if they strongly agreed with a positively worded item and agreed with a negatively worded item (or vice versa), they received a score of 2; and if they strongly agreed with both a positively and negatively worded item, the score was 3. This procedure ensures that any form of simultaneous agreement with both positively and negatively worded items (not only strong agreement) contributes to acquiescence and at the same time takes into account the magnitude of acquiescent responding (i.e., acquiescence is stronger when people strongly agree with a statement and its opposite). The possible comparisons between a negatively worded item and the positively worded items within the same scale were averaged, and ARS2 was then computed as the average score across scales. Disacquiescence was assessed in the same way, except that it is based on simultaneous disagreement with both positively and negatively worded items.

The two acquiescence indices were substantially correlated (.56), as were the two disacquiescence indices (.50). This supports the convergent validity of the two methods for assessing (dis)acquiescence, and it increases our confidence that the measures of stylistic responding are not confounded with substantive content. The standardized indices based on the two methods were averaged to form overall measures of acquiescence and disacquiescence with reliabilities of .83 and .73, respectively.[4]

[4]These reliabilities are based on the formula for computing the reliability of linear combinations of measures (see Nunnally 1978). The reliabilities of the individual indices were computed on the basis of the internal consistency of responses to all 60 items or the scale-specific items across respondents.

Table 2
SCALE CHARACTERISTICS FOR THE 14 CONSTRUCTS USED IN THIS STUDY

	Number of Items	*Negatively Worded Items*[a]	*Alpha (Correlation)*[b]
Attitude toward advertising in general	5	4	.74 (.62, .78)
Attitude toward the past (romanticism)	3	0	.69 (.58, .71)
Attitude toward the past (technology)	2	2	.42 (.28, .51)
Change seeker index	7	2	.75 (.60, .81)
Consumer ethnocentrism	8	0	.93 (.89, .95)
Deal proneness	4	1	.76 (.64, .82)
Environmental consciousness	5	1	.70 (.58, .72)
Exploratory acquisition of products	5	5	.80 (.71, .84)
Exploratory information seeking (advertising)	3	0	.64 (.55, .72)
Exploratory information seeking (shopping)	2	1	.42 (.26, .73)
Health consciousness	5	0	.85 (.79, .88)
Price consciousness	3	0	.66 (.60, .75)
Product involvement	3	1	.82 (.73, .86)
Quality consciousness	5	0	.79 (.67, .83)

[a]A negatively worded item is defined as an item for which higher scores indicate the opposite of what the name of the construct implies. For example, the scale measuring attitude toward advertising in general contains four items for which higher scores indicate a lower attitude toward advertising.

[b]The Pearson correlation is given for two-item scales, and coefficient alpha is given for scales with more than two items. Both are computed across respondents and countries. The numbers in parentheses show the lowest and highest reliabilities/correlations from the country-specific analyses.

ERS and RR. We calculated the index of ERS (ERS1) as the frequency (across all 60 items) with which a respondent strongly agreed or strongly disagreed with questionnaire statements (Chen, Lee, and Stevenson 1995; Hui and Triandis 1985). Similarly, we calculated the index of RR (RR1) as the standard deviation of a person's responses across all 60 items (Greenleaf 1992a). Because the average absolute intercorrelation of the items was low (.12), we did not deem it necessary to select items that are as uncorrelated as possible. The reliability of ERS1 was .93, and the correlation between ERS1 and RR1 was .92.[5] Because of the high correlation, we averaged the two indices to form an overall measure of extreme responding.

MPR. This response style was assessed as the frequency with which respondents endorsed the middle scale category on the 60 items (Chen, Lee, and Stevenson 1995; Stening and Everett 1984). The reliability of this measure was .85.[6]

NCR. Following Marsh (1987; see also Watkins and Cheung 1995), we constructed an index of NCR that was based on the notion that people should respond consistently to correlated item pairs. Operationally, this index is defined as the sum of absolute differences between people's responses to pairs of items, where the items in each pair are selected to be as highly correlated as possible and to have similar item means across respondents. For this measure not to be confounded with positivity and negativity, the two items in each pair must be scored in the same direction (i.e., both items must be worded either positively or negatively). We selected 24 pairs of items that fulfilled these criteria. Two examples of items pairs are, "I become incensed when I think about the harm being done to plant and animal life by

[5]The reliability of ERS1 was calculated as follows: First, we defined a variable of ERS for each of the 60 attitudinal statements. These variables were coded 1 for an extreme response (1 or 5 on the original five-point scale) and 0 otherwise. Second, we computed the internal consistency of people's scores on the 60 variables using coefficient alpha. A similar procedure was used for MPR and NCR.

[6]We also constructed indices of ARS, DARS, ERS, and MPR that were based on specially selected items that are minimally correlated. These indices were highly correlated with the response style measures used in this study. This provides further evidence that our measures of stylistic responding are based on heterogeneous items and do not confound stylistic with content variance.

pollution" and "When I think of the ways industries are polluting the environment, I get frustrated and angry" (from the environmental consciousness scale) and "I sacrifice a lot to eat as healthy as possible" and "I value my health so much that I sacrifice many things for it" (from the health consciousness scale). The average correlation between item pairs was .58 (range of .41 to .79), and the average difference in means was .03. The NCR measure had a reliability of .55, based on the internal consistency of scores for the 24 item pairs across respondents.

Relationships among the five categories of response styles. The correlations among the five response styles, computed as the average correlation across scales and countries, were as expected. First, ARS and DARS were negatively correlated (r = –.16), though the correlation was not high. Second, ERS was positively correlated with both ARS (r = .59) and DARS (r = .41). Third, MPR had a negative relationship with all other response styles (r = –.48 with ARS, r = –.40 with DARS, and r = –.55 with ERS). Finally, NCR was positively correlated with ARS (r = .18), DARS (r = .30), and ERS (r = .38) and negatively correlated with MPR (r = –.19).

Model Estimation

The model expressing scale scores as a function of response styles is technically a three-level hierarchical linear model. The Level 1 (or individual-level) model is given by Equation 1, the Level 2 (or scale-level) model consists of Equations 2, 3, 8, 9, 14, and 15, and the Level 3 (or country-level) model comprises Equations 4–7, 10–13, and 16–17. We estimated the model using individual respondents' scores on the 14 scales and the five response style indices discussed previously. To avoid confounding of substantive and stylistic variance, we computed scale-specific response style indices for each person by excluding the items belonging to the scale in question. For example, for the regression of consumer ethnocentrism on the various response styles, we did not use the ethnocentrism items in the calculation of any of the response style indices. If a particular substantive scale was regressed on a response style measure that included items from the same scale, the amount of shared variance would be exaggerated because of item overlap (see Bagozzi 1994). We centered the response style variables by scale and country so that the scale- and country-specific intercepts π_{0jk} correspond to the mean score of a country on a particular scale (Bryk and Raudenbush 1992). $PROP_{jk}$ is coded as described previously, and more positive (negative) values indicate a larger proportion of positively (negatively) worded items. Three scales had a predominance of negatively worded items, ten scales had mostly positively worded items, and one scale was balanced. Although most of the PROP values are positive, they span the continuum from –1 to +1. DMP_{jk} is defined as the positive or negative deviation of country k's average score on scale j from the midpoint of the scale.

The model investigating response style effects on correlations between scales is a two-level model.[7] The first level is given by Equation 19, and the second level by Equations 20–22. The three types of correlations represented in Equation 19 were obtained as follows: The uncorrected correlation, $c_{jj',k}$, is simply the correlation between observed scores for scales j and j′ in country k, where the correlation is computed across individuals. When observed scores are regressed on the five response style variables for each scale and country, observed scores can be partitioned into a response style component (predicted scores) and a residual component (scale scores purified of the response style component). The correlation between corrected scores, $cc_{jj',k}$, is obtained by correlating the residual scores for scales j and j′ in country k. Finally, the correlation between the response style components, $rc_{jj',k}$, is obtained by correlating the predicted scores for scales j and j′ in country k. The model is estimated across the 91 possible pairwise correlations of the 14 substantive scales in each of the 11 countries.

We estimated both models using the program HLM (Bryk, Raudenbush, and Congdon 1996), which enables us to estimate an integrated model in which the regression coefficients are allowed to vary across subsets of observations, variation in the coefficients can be related to hypothesized predictors, and error terms are handled appropriately. Under certain assumptions described by Bryk and Raudenbush (1992), HLM produces maximum-likelihood estimates of all regression coefficients and variance/covariance components and associated tests of statistical significance.

RESULTS

We present the findings in two sections. We first discuss the influence of the five response styles on scale scores, and then we investigate the effect of stylistic responding on correlations between scales.

Effect of Response Styles on Scale Scores

As a first step, we estimated a three-level model in which no predictor variables were specified at Levels 2 and 3.[8] This model serves as a baseline model and can be used to partition the total variation in the effect of each response style on scale scores into components that are due to scales (Level 2) and countries (Level 3). The results indicate that the influence of all five forms of stylistic responding on observed scores varied significantly by scale (all *p*-values were smaller than .0001) but that the variation across countries was negligible (all *p*-values were nonsignificant). For each response style, at least 98% of the variation in the π_{pjk} was attributable to differences across scales, and at most 2% of the variation was due to cross-country differences.

Next, PROP and DMP were introduced as Level 2 predictors. The parameter estimates for this model are shown in the left-hand columns of Table 3.[9] The first prediction was that as the proportion of positively (negatively) worded items in a scale increased, acquiescence would have a more positive (negative) effect on scale scores. This implies that $\gamma_{11} > 0$. The opposite was predicted for disacquiescence, which suggests that $\gamma_{21} < 0$. The results are consistent with these hypotheses. The estimate of γ_{11} is .146 ($p < .0001$), and the estimate of γ_{21} is –.093 ($p < .0001$). For a balanced

[7]The variables of interest in this model are correlations, which are computed across individuals, so the individual-level model vanishes.

[8]For this model to converge, we needed to constrain the error term for the intercept in the Level 3 model to zero.

[9]We needed to set the error terms of the response style effects in the Level 3 model (i.e., the u_{pjk}, p = 1, ..., 5) to zero in this analysis to achieve convergence. This is unproblematic because the associated variance components previously were shown to be negligible.

Table 3
EFFECTS OF RESPONSE STYLES ON SCALE SCORES AND CORRELATIONS BETWEEN SCALES

Model 1: Effect of Response Styles on Scale Scores				*Model 2: Effect of Response Styles on Correlations Between Scales*			
Parameter	*Coefficient (Variance Component)*	*t-Ratio (χ^2-Value)*	*p-Value*	*Parameter*	*Coefficient (Variance Component)*	*t-Ratio (χ^2-Value)*	*p-Value*
γ_{00}	3.241	89.48	.000	γ_0	.034	16.67	.000
γ_{10}	−.030	−3.36	.001	γ_1	.887	47.90	.000
γ_{11}	.146	18.78	.000	γ_2	.133	13.37	.000
γ_{20}	−.027	−5.10	.000	$Var(u_{0k})$	.000	17.99	.055
γ_{21}	−.093	−17.94	.000	$Var(u_{1k})$	.002	25.35	.005
γ_{30}	.013	1.42	.155	$Var(u_{2k})$	.001	170.75	.000
γ_{31}	.106	8.26	.000				
γ_{40}	−.017	−3.09	.002				
γ_{41}	−.061	−6.91	.000				
γ_{50}	−.003	−.97	.331				
$Var(r_{0jk})$	.196	30609.21	.000				
$Var(r_{1jk})$	.009	872.17	.000				
$Var(r_{2jk})$	.002	413.47	.000				
$Var(r_{3jk})$	.009	930.98	.000				
$Var(r_{4jk})$	.003	480.08	.000				
$Var(r_{5jk})$	.001	420.71	.000				
$Var(u_{00k})$	.0004	6.42	>.50				

scale, PROP = 0, so the intercepts γ_{10} and γ_{20} should be nonsignificant, because with a balanced scale acquiescence and disacquiescence are presumably controlled for. In the present case, the estimated intercepts are nonzero but small in absolute magnitude ($\hat{\gamma}_{10} = -.030$, $p < .001$; $\hat{\gamma}_{20} = -.027$, $p < .0001$). Thus, balancing scales largely, but not completely, eliminates variance due to (dis)acquiescence. This finding is consistent with Jackson's (1967) and Messick's (1991) arguments that balancing scales does not necessarily remove all acquiescence bias. Overall, PROP accounted for a high 60% and 62% of the variation in the effect of acquiescence and disacquiescence on scale scores, respectively. This suggests that the extent to which a scale is balanced is an important determinant of the contaminating influence of acquiescence and disacquiescence on scale scores.

Our predictions for ERS and MPR were that if the scale mean was greater than the midpoint, extreme responding (stylistic endorsement of the middle scale category) should increase (decrease) scores, whereas if the scale mean was smaller than the midpoint, scores should be deflated (inflated). This implies that $\gamma_{31} > 0$ and $\gamma_{41} < 0$. As seen in Table 3, both hypotheses are supported. The estimate of γ_{31} is .106 ($p < .0001$), and the estimate of γ_{41} is −.061 ($p < .0001$). When the deviation from the midpoint was 0, the estimated effect of ERS (i.e., the intercept) was .013 (nonsignificant), and the estimated effect of MPR was a small −.017 ($p < .01$). Overall, DMP accounted for a substantial 35% and 33% of the variation in the impact of ERS and MPR on scale scores, respectively.

Although the variation in the effect of NCR on scale scores was highly significant, on average NCR did not systematically contaminate scale scores ($\hat{\gamma}_{50} = -.003$, nonsignificant).

It is instructive to examine the proportion of variance in scale scores accounted for by stylistic responding (see Table 4). Across all scales and countries, these figures range from 0% to 29%. On average, 8% of the variance in observed scores can be attributed to stylistic responding. Although some scales are virtually free of stylistic variance (e.g., on average, only 2% of the variance in the change seeker index is stylistic variance; the highest value is 4%), several scales are fairly strongly contaminated by response biases. For example, on average 22% of the variance in health con-

Table 4
PROPORTION OF VARIANCE IN SCALE SCORES ACCOUNTED FOR BY STYLISTIC RESPONDING

	Mean Contamination Across 11 Countries	*Median Contamination Across 11 Countries*	*Range of Contamination Across 11 Countries*
Attitude toward advertising in general	.04	.02	.01–.13
Attitude toward the past (romanticism)	.10	.11	.05–.16
Attitude toward the past (technology)	.07	.06	.04–.15
Change seeker index	.02	.01	.01–.04
Consumer ethnocentrism	.16	.15	.11–.23
Deal proneness	.03	.02	.00–.06
Environmental consciousness	.12	.10	.08–.26
Exploratory acquisition of products	.10	.10	.06–.15
Exploratory information seeking (advertising)	.05	.05	.02–.11
Exploratory information seeking (shopping)	.01	.01	.00–.04
Health consciousness	.22	.23	.12–.29
Price consciousness	.04	.02	.01–.15
Product involvement	.03	.02	.01–.06
Quality consciousness	.14	.12	.07–.28

sciousness (range of 12% to 29%), 16% of the variance in consumer ethnocentrism (range of 11% to 23%), 14% of the variance in quality consciousness (range of 7% to 28%), and 12% of the variance in environmental consciousness (range of 8% to 26%) is due to the operation of response biases. These are also the scales that contained no or very few reverse-scored items (only environmental consciousness has one reverse-scored item), and for three of the four scales the scale mean deviated the most from the midpoint of the response scale of all scales (the deviation was .85, .74, and .60 for quality consciousness, environmental consciousness, and health consciousness, respectively). Overall, stylistic responding appears to be a significant source of extraneous variation in at least some of the substantive scales investigated in this article.

Effect of Response Styles on Relationships Between Scales

The parameter estimates for this model are shown in the right-hand columns of Table 3. There is a strong, positive relationship between corrected and uncorrected correlations ($\hat{\gamma}_1$ = .887, $p < .0001$), but the relationship is smaller than one. The estimate of γ_2 is .133 ($p < .0001$), which means that the correlation between the response style components of a scale has a significant impact on observed correlations. Corrected correlations ($cc_{jj'.k}$) account for 64% of the variation in uncorrected correlations, and when the correlation between the response style components ($cc_{jj',k}$) is added as an additional predictor, 91% of the variance in observed correlations is accounted for. This indicates that, overall, a substantial portion of the observed correlation between scales is due to correlations between the systematic errors in the scales involved.

It is instructive to examine the inflation or deflation of corrected correlations when the response styles are either positively or negatively correlated. As expected, if the response style components are positively (negatively) correlated and the correlation between corrected substantive scales is also positive (negative), the correlation between uncorrected scores is even more positive (negative). The average correlation increases from .13 to .20 in the positive/positive cell (an inflation of 54%) and from –.09 to –.15 in the negative/negative cell (an inflation of 67%). Conversely, in the two cells in which the substantive and response style components move in opposite directions, observed correlations are biased toward zero (from .11 to .05 and from –.07 to .01, respectively).

The foregoing results are based on all possible correlations between scales, including scales that are virtually free of stylistic variance. To illustrate the extent to which response styles can affect correlations between substantive scales when stylistic variance accounts for a substantial portion of the total variance, consider the correlations among the four scales that were found to be relatively strongly contaminated by response styles: health consciousness (HCO), consumer ethnocentrism (CET), quality consciousness (QCO), and environmental consciousness (ECO). The correlations between uncorrected scale scores are all substantial (HCO–QCO .40, HCO–ECO .33, QCO–ECO .31, HCO–CET .28, QCO–CET .19, and ECO–CET .15). Theoretically, a correlation can be expected among health consciousness, environmental consciousness, and quality consciousness, because many Europeans perceive health and environmental concerns as components of product quality (e.g., Grunert et al. 1996). However, it is unclear why consumer ethnocentrism should be correlated with any of the three other constructs. After response style variance is taken into account, all correlations are substantially reduced (HCO–QCO .20, HCO–ECO .15, QCO–ECO .13, HCO–CET .02, QCO–CET .00, and ECO–CET .01). Thus, the major reason these scales are correlated is that they share common response style variance. If the covariation due to stylistic responding is eliminated, these scales are uncorrelated or much more weakly correlated. Consistent with theoretical expectations, ethnocentrism is uncorrelated with health consciousness, quality consciousness, and environmental consciousness. Although the correlations among health consciousness, quality consciousness, and environmental consciousness remain significant (which is in line with expectations), their magnitude is reduced by approximately half.

As seen in Table 3, there was significant variation in both β_{1k} and β_{2k} across countries (i.e., the variances of u_{1k} and u_{2k} were significantly different from zero). However, this effect was almost entirely due to 2 of the 11 countries. Specifically, the correspondence between uncorrected and corrected scores was much lower and the contribution of shared response style variance to observed correlations was much higher for Greece and Portugal than for the other countries. When dummy variables contrasting these two countries with the remaining countries were added to the Level 2 model, 96% and 90% of the variation in the estimates of β_{1k} and β_{2k} were explained by the two dummies.

In conclusion, our results show that response styles, in general, are not a negligible source of variation in scale scores. The findings suggest that stylistic responding can have systematic and substantial effects on correlations between scales and that researchers should be concerned about the deleterious effects of stylistic responding on research results. There was also evidence that the problems created by response styles are more severe in some countries than others, but further research is needed to explain what the theoretical reasons for these differences are.

DISCUSSION

Response styles are a source of systematic measurement error and thus threaten the validity of conclusions drawn from marketing research data (Bearden and Netemeyer 1999). In this article, we examined five forms of stylistic responding: ARS, DARS, ERS/RR, MPR, and NCR. We developed two models, one dealing with the systematic effects of each response style on respondents' answers to substantive questions and the other investigating the biasing effect of stylistic responding on correlations between substantive scales. Using large, representative samples of consumers from 11 countries of the European Union, which allowed for a strong test of the empirical generalizability of our findings in an international context (Van de Vijver and Leung 1997), we showed that response styles can lead to nontrivial contamination of observed scores and substantial bias in observed correlations between scales that are contaminated by stylistic variance. Compared with the variation across scales, the variation across countries was small. This provides evidence for the robustness of response style effects across different countries, cultural settings, and languages.

Two moderating factors of the degree of contamination of observed scores by stylistic responding were considered. One was the proportion of positively and negatively keyed items. Reverse-scoring items has frequently been suggested as an effective means of controlling for acquiescence and disacquiescence biases (e.g., Paulhus 1991), but empirical demonstrations of the beneficial effects of balanced scales are rare. Our findings show that the extent to which a scale is unbalanced has an important influence on the degree of contamination of scale scores and that the bias can be positive or negative depending on whether most items in a scale are positively or negatively worded. Although these results confirm the conventional wisdom that balancing scales is effective in counteracting the adverse influence of yea-saying and nay-saying, they are in stark contrast with the finding from our secondary analysis of the measurement instruments contained in Bearden and Netemeyer's (1999) *Handbook of Marketing Scales*, which showed that close to half of the instruments failed to contain a single reverse-scored item and fewer than 10% of instruments were balanced.[10]

The other moderating factor investigated in this article was the extent to which the scale mean deviates from the midpoint of the response scale. Our findings show that this scale characteristic is an important determinant of whether ERS and MPR will contaminate scale scores. In both cases, the problem becomes more severe as the deviation of the scale mean from the midpoint of the scale increases, but the bias works in opposite directions. For ERS, a positive (negative) deviation inflates (deflates) scores, whereas for MPR, a positive (negative) deviation deflates (inflates) scores. This suggests that researchers should be particularly concerned about ERS and MPR when, on average, respondents tend to have relatively strong opinions (in either a positive or negative direction) about an issue.

The two moderating factors point to conditions under which certain response styles may be expected to have a biasing effect on observed scores, and they can thus serve as warning signs alerting researchers to possible problems with response styles in a given situation. However, they do not shed light on why some people are more prone to stylistic responding or why certain situations encourage or discourage stylistic response tendencies. Our study is also silent about how response styles exert their influence on responses to substantive questions. Most of the research on the antecedents of stylistic responding is old and has not taken advantage of the advances in personality psychology that have occurred in recent years (see John 1990). Substantial progress also has been made in investigating the cognitive mechanisms underlying people's responses to survey questions and the biases that may arise as a function of question wording and question order (see Schwarz, Groves, and Schuman 1998). This research should be extended to a consideration of response styles so that the knowledge of the consequences of stylistic responding and the conditions under which stylistic responding has biasing effects on scale scores and correlations between scales (which was the major concern of this article) is supplemented by greater understanding of its antecedents.

An important issue in research with response styles has always been to cleanly separate stylistic variance from substantive variance. If the two sources of variance are confounded, conclusions about the contamination of scale scores and correlations between scales will be exaggerated. We are confident that this is not the case in our study. First, the response style indices were calculated from many different scales that, on average, were not strongly correlated, so the items are truly heterogeneous as required. Also, multiple indicators based on different methods were used for several response styles to establish convergent validity. Second, scale-specific response style indices were computed to avoid item overlap and guard against spurious response style contamination. Third, the influence of stylistic responding on scale scores varied systematically with variables that theoretically should act as determinants of response style differences but should have no effect if something other than response styles is assessed. For example, it is unclear why the proportion of reverse-scored items should moderate anything but the effect of acquiescent responding on scale scores. All these points argue against the charge that stylistic variance may be confounded with substantive variance. However, it would be useful to extend the present research by using some of the scales that have been constructed specifically to assess particular response styles. For example, Greenleaf (1992b) has developed an instrument measuring ERS, and Wells (1963) has proposed a scale of yea-saying/nay-saying based on previous work by Couch and Keniston (1960).

As does any study, our research has several limitations that present opportunities for future work. The present investigation is based on five-point labeled Likert scales (with scale steps of "strongly disagree" to "strongly agree"). Further research could examine response tendencies for other scale formats. One important contribution of such work would be the identification of response formats that suffer the least from response styles, both domestically and cross-nationally. Research on response styles could also be extended to other countries. Most previous studies have been conducted in the United States and eastern Asia. The present study deals with the European Union. Although some research has been conducted among Hispanics who had recently migrated to the United States, research on response tendencies in other parts of the world, such as the emerging economies of Latin America and Eastern Europe, is scarce. Finally, our findings are restricted to the instruments for which we had data, and in most cases we used only a subset of the items from the full scales. Although they were all carefully validated measurement scales, they represent a limited cross-section of the instruments used in marketing, and further research is needed to show whether our conclusions hold for other scales as well. In particular, it would be interesting to find whether scales used in applied marketing research (e.g., lifestyle instruments) are more or less susceptible to response biases than the scales used in this research.

In summary, on the basis of our results, we recommend that marketing researchers pay greater attention to the phenomenon of response styles. Ideally, this would involve the

[10]It should be mentioned that the use of balanced scales is not a panacea for acquiescence bias. As Jackson (1967) and Messick (1991) point out, it is difficult to construct truly balanced scales, and balancing scales does not necessarily eliminate acquiescence bias completely. However, as shown in our study, the magnitude of acquiescence bias is reduced substantially when a balanced scale is used.

development of instruments that minimize opportunities for stylistic responding (e.g., by using balanced scales to guard against acquiescence biases, by eliminating items that are susceptible to social desirability biases). If this is not possible, stylistic responding should be controlled for post hoc by purifying scale scores of response styles. This requires that both positively and negatively worded items from the same scale are available (to compute an index of acquiescent responding) or that items that are heterogeneous in content are included in the questionnaire (to compute indices of all the response styles studied in this article). After suitable response style indices have been calculated (see Table 1), respondents' scores on the substantive scales can be regressed on the response style indices, and all substantive analyses can be carried out on the residualized scores (i.e., scores that have been purified of extraneous method variance). Corrected covariance or correlation matrices based on purified scale scores can serve as input to tests of structural equation models. Failure to apply these corrections may lead to spurious results when analyses are based on contaminated scale scores or correlations between scales.

REFERENCES

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

Bagozzi, Richard P. (1994), "Measurement in Marketing Research: Basic Principles of Questionnaire Design," in *Principles of Marketing Research*, Richard P. Bagozzi, ed. Cambridge, MA: Blackwell, 1–49.

Bass, Frank M. and Jerry Wind (1995), "Introduction to the Special Issue: Empirical Generalizations in Marketing," *Marketing Science*, 14 (3-2), G1–G6.

Baumgartner, Hans and Jan-Benedict E.M. Steenkamp (1996), "Exploratory Consumer Buying Behavior: Conceptualization and Measurement," *International Journal of Research in Marketing*, 13 (2), 121–37.

Bearden, William O. and Richard G. Netemeyer (1999), *Handbook of Marketing Scales: Multi-Item Measures for Marketing and Consumer Behavior Research*, 2d ed. Newbury Park, CA: Sage Publications.

Broen, William E., Jr., and Robert D. Wirt (1958), "Varieties of Response Sets," *Journal of Consulting Psychology*, 22 (June), 237–40.

Bryk, Anthony and Stephen Raudenbush (1992), *Hierarchical Linear Models*. Newbury Park, CA: Sage Publications.

———, ———, and Richard Congdon (1996), *HLM: Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L Programs*. Chicago: Scientific Software.

Chen, Chuansheng, Shin-ying Lee, and Harold W. Stevenson (1995), "Response Style and Cross-Cultural Comparisons of Rating Scales Among East Asian and North American Students," *Psychological Science*, 6 (May), 170–75.

Couch, Arthur and Kenneth Keniston (1960), "Yeasayers and Naysayers: Agreeing Response Set as a Personality Variable," *Journal of Abnormal and Social Psychology*, 60 (2), 151–72.

Craig, C. Samuel and Susan P. Douglas (2000), *International Marketing Research*. New York: John Wiley & Sons.

Cronbach, Lee J. (1946), "Response Set and Test Validity," *Educational and Psychological Measurement*, 6 (Winter), 475–94.

Gaski, John F. and Michael J. Etzel (1986), "The Index of Consumer Sentiment Toward Marketing," *Journal of Marketing*, 50 (July), 71–81.

Green, Donald P. and Jack Citrin (1994), "Measurement Error and the Structure of Attitudes: Are Positive and Negative Judgments Opposites?" *American Journal of Political Science*, 38 (February), 256–81.

———, Susan L. Goldman, and Peter Salovey (1993), "Measurement Error Masks Bipolarity in Affect Ratings," *Journal of Personality and Social Psychology*, 64 (June), 1029–1041.

Greenleaf, Eric A. (1992a), "Improving Rating Scale Measures by Detecting and Correcting Bias Components in Some Response Styles," *Journal of Marketing Research*, 29 (May), 176–88.

——— (1992b), "Measuring Extreme Response Style," *Public Opinion Quarterly*, 56 (Fall), 328–51.

Grunert, Klaus G., Allan Baadsgaard, Hanne H. Larsen, and Tage K. Madsen (1996), *Market Orientation in Food and Agriculture*. Boston: Kluwer Academic Publishers.

Grunert, Suzanne C. and Hans J. Juhl (1995), "Values, Environmental Attitudes, and Buying of Organic Foods," *Journal of Economic Psychology*, 16 (1), 39–62.

Hamilton, David L. (1968), "Personality Attributes Associated with Extreme Response Style," *Psychological Bulletin*, 69 (March), 192–203.

Holbrook, Morris B. and Robert M. Schindler (1994), "Age, Sex, and Attitude Toward the Past as Predictors of Consumers' Aesthetic Tastes for Cultural Products," *Journal of Marketing Research*, 31 (August), 412–22.

Hui, C. Harry and Harry C. Triandis (1985), "The Instability of Response Sets," *Public Opinion Quarterly*, 49 (Summer), 253–60.

——— and ——— (1989), "Effects of Culture and Response Format on Extreme Response Style," *Journal of Cross-Cultural Psychology*, 20 (September), 296–309.

Jackson, Douglas N. (1967), "Acquiescence Response Styles: Problems of Identification and Control," in *Response Set in Personality Assessment*, Irwin A. Berg, ed. Chicago: Aldine Publishing Company, 71–114.

John, Oliver P. (1990), "The 'Big Five' Factor Taxonomy: Dimensions of Personality in the Natural Language and in Questionnaires," in *Handbook of Personality: Theory and Research*, Lawrence A. Pervin, ed. New York: The Guilford Press, 66–100.

Knowles, Eric S. and Kobe T. Nathan (1997), "Acquiescent Responding in Self-Reports: Cognitive Style or Social Concern," *Journal of Research in Personality*, 31 (June), 293–301.

Lentz, T.F. (1938), "Acquiescence as a Factor in the Measurement of Personality," *Psychological Bulletin*, 35 (November), 659.

Lichtenstein, Donald R., Peter H. Bloch, and William C. Black (1988), "Correlates of Price Acceptability," *Journal of Consumer Research*, 15 (September), 243–52.

———, Richard G. Netemeyer, and Scot Burton (1995), "Assessing the Domain Specificity of Deal Proneness: A Field Study," *Journal of Consumer Research*, 22 (December), 314–26.

Maloney, Michael P., Michael P. Ward, and G. Nicholas Braucht (1976), "A Revised Scale for the Measurement of Ecological Attitudes and Knowledge," *American Psychologist*, 30 (July), 787–91.

Marín, Gerardo, Raymond J. Gamba, and Barbara V. Marín (1992), "Extreme Response Style and Acquiescence Among Hispanics," *Journal of Cross-Cultural Psychology*, 23 (December), 498–509.

Marsh, Herbert W. (1987), *The Self-Description Questionnaire I: Manual and Research Monograph*. San Antonio, TX: Psychological Corporation.

Martin, John (1964), "Acquiescence—Measurement and Theory," *British Journal of Social and Clinical Psychology*, 3 (October), 216–25.

McGee, Richard K. (1967), "Response Set in Relation to Personality: An Orientation," in *Response Set in Personality*

Assessment, Irwin A. Berg, ed. Chicago: Aldine Publishing Company, 1–31.

Messick, Samuel (1967), "The Psychology of Acquiescence: An Interpretation of Research Evidence," in *Response Set in Personality Assessment*, Irwin A. Berg, ed. Chicago: Aldine Publishing Company, 115–45.

——— (1968), "Response Sets," in *International Encyclopedia of the Social Sciences*, Vol. 13, David L. Sills, ed. New York: Macmillan, 492–96.

——— (1991), "Psychology and Methodology of Response Styles," in *Improving Inquiry in Social Science: A Volume in Honor of Lee J. Cronbach*, Richard E. Snow and David E. Wiley, eds. Hillsdale, NJ: Lawrence Erlbaum Associates, 161–200.

Mick, David Glen (1996), "Are Studies of Dark Side Variables Confounded by Socially Desirable Responding? The Case of Materialism," *Journal of Consumer Research*, 23 (September), 106–19.

Mittal, Banwari and Mysung-Soo Lee (1989), "A Causal Model of Consumer Involvement," *Journal of Economic Psychology*, 10 (November), 363–89.

Nunnally, Jum C. (1978), *Psychometric Theory*, 2d ed. New York: McGraw-Hill.

O'Donovan, Denis (1965), "Rating Extremity: Pathology or Meaningfulness?" *Psychological Review*, 72 (5), 358–72.

Oude Ophuis, Peter A.M. (1989), "Measuring Health Orientation and Health Consciousness as Determinants of Food Choice Behavior: Development and Implementation of Various Attitudinal Scales," in *Marketing Thought and Practice in the 1990s*, George A. Avlonitis, ed. Athens, Greece: European Marketing Academy, 1723–25.

Paulhus, Delroy L. (1991), "Measurement and Control of Response Bias," in *Measures of Personality and Social Psychological Attitudes*, John P. Robinson, Phillip R. Shaver, and Lawrence S. Wright, eds. San Diego, CA: Academic Press, 17–59.

Ray, John J. (1983), "Reviving the Problem of Acquiescent Response Bias," *Journal of Social Psychology*, 121 (October), 81–96.

Rorer, Leonard G. (1965), "The Great Response-Style Myth," *Psychological Bulletin*, 63 (March), 129–56.

Schuman, Howard and Stanley Presser (1981), *Questions and Answers in Attitude Surveys*. New York: Academic Press.

Schwarz, Norbert, Robert M. Groves, and Howard Schuman (1998), "Survey Methods," in *The Handbook of Social Psychology*, Vol. 1, 4th ed., Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey, eds., Boston: McGraw-Hill, 143–79.

Shimp, Terence A. and Subhash Sharma (1987), "Consumer Ethnocentrism: Construction and Validation of the CETSCALE," *Journal of Marketing Research*, 24 (August), 280–89.

Shulman, Art (1973), "A Comparison of Two Scales on Extremity Response Bias," *Public Opinion Quarterly*, 37 (Fall), 407–12.

Snyder, Mark and William Ickes (1985), "Personality and Social Behavior," in *Handbook of Social Psychology*, Vol. 2, 3d ed., Gardner Lindzey and Elliot Aronson, eds., New York: Random House, 883–948.

Steenkamp, Jan-Benedict E.M. (1989), *Product Quality: An Investigation into the Concept and How It Is Perceived by Consumers*. Assen: Van Gorcum.

——— and Hans Baumgartner (1995), "Development and Cross-Cultural Validation of a Short Form of CSI as a Measure of Optimum Stimulation Level," *International Journal of Research in Marketing*, 12 (2), 97–104.

Stening, B.W. and J.E. Everett (1984), "Response Styles in a Cross-Cultural Managerial Study," *Journal of Social Psychology*, 122 (April), 151–56.

Van de Vijver, Fons J.R. and Kwok Leung (1997), "Methods and Data Analysis of Comparative Research," in *Handbook of Cross-Cultural Psychology, Volume 1: Theory and Method*, John W. Berry, Ype H. Poortinga, and Janak Pandey, eds. Boston: Allyn and Bacon, 257–300.

Watkins, David and Steven Cheung (1995), "Culture, Gender, and Response Bias: An Analysis of Responses to the Self-Description Questionnaire," *Journal of Cross-Cultural Psychology*, 26 (September), 490–504.

Watson, Dorothy (1992), "Correcting for Acquiescent Response Bias in the Absence of a Balanced Scale," *Sociological Methods & Research*, 21 (August), 52–88.

Wells, William D. (1963), "How Chronic Overclaimers Distort Survey Findings," *Journal of Advertising Research*, 3 (2), 8–18.

Winkler, John D., David E. Kanouse, and John E. Ware Jr. (1982), "Controlling for Acquiescence Response Set in Scale Development," *Journal of Applied Psychology*, 67 (October), 555–61.

Wyer, Robert S., Jr. (1971), "The Effects of General Response Style on Measurement of Own Attitude and the Interpretation of Attitude-Relevant Messages," *British Journal of Social and Clinical Psychology*, 8 (2), 105–15.