Journal of the Association for Information Systems
JAIS

Research Article

Toward a Better Use of the Semantic Differential in IS Research: An Integrative Framework of Suggested Action

Tibert Verhagen
VU University Amsterdam
t.verhagen@vu.nl

Bart van den Hooff
VU University Amsterdam
b.j.vanden.hooff@vu.nl

Selmar Meents
Amsterdam University of Applied Sciences—International Business School
s.meents@hva.nl

Abstract

The semantic differential is a widely applied measurement technique in the information systems field. As we demonstrate in this study, however, there is evidence that many of the applications of the semantic differential seem to be subject to common shortcomings. In this study, we address these shortcomings by creating awareness of the requirements underlying semantic differentiation. We discuss the requirements of semantic differentiation and use them as a foundation to introduce a framework to assist researchers in applying the semantic differential more adequately. The framework puts renewed emphasis on bipolar scale selection and dimensionality testing, introduces semantic bipolarity as new criterion, and proposes distinct stages for the testing of wording and contextual contamination. We exemplify the framework using an illustration exercise, which centers on the assessment of the meaning of the concept "electronic marketplace quality". Using a mixture of qualitative and quantitative methods, the illustration exercise clarifies the prerequisites for semantic differentiation and provides suggestions for researchers. The paper concludes with a discussion of several methodological implications.

Keywords: *Semantic Differential, Linguistic Contrast, Psychological Bipolarity, Concept-scale Pairing, Contextual Contamination, Electronic Marketplace Quality, Measurement Validation.*

* Gary Klein was the accepting senior editor. This article was submitted on 27th June 2013 and went through three revisions.

Volume 16, Issue 2, pp. 108-143, February 2015

Toward a Better Use of the Semantic Differential in IS Research: An Integrative Framework of Suggested Action

1. Introduction

The semantic differential (SD) is frequently used in information systems (IS) research. Consisting of a set of bipolar scales that match the concept to be measured, the SD is an established technique of observing and measuring the meaning of concepts such as information system satisfaction (Xue, Liang, & Wu, 2011), attitude toward information technology (Bhattacherjee & Premkumar, 2004) information systems planning success (Doherty, Marples, & Suhaimi, 1999), perceived enjoyment (Luo, Chea, & Chen, 2011), and website performance (Huang, 2005).

Theory presents the SD as one of the most appropriate techniques to assess the intensity and direction of the meaning of concepts, especially complex and multidimensional concepts (Mindak, 1961). This is an important advantage of the SD to the IS field, in which diverse theoretical perspectives (e.g., technology acceptance, innovation diffusion, task-technology fit) have been used to decompose perceptions, beliefs, or attitudes regarding intangible and even invisible system characteristics into separate concepts that can be included and subsequently tested in various intricate nomological structures. The relevance of the SD is also mirrored in the increased attention it has received in recent IS publications. For example, Chin, Johnson, and Schwarts (2008) claim that an advantage of the SD is that it can be used as a short-form scale format to reduce survey completion time. The SD's methodological advantages are also backed by empirical results. For instance, the SD has been demonstrated to outperform Likert-based scaling or stapel scaling on robustness (Hawkins, Albaum, & Best, 1974), reliability (Wirtz & Lee, 2003), and validity (Van Auken & Barry, 1995).

Although IS scholars recognize the SD's value and frequently use the measurement technique, our review of the IS literature indicates that the application of SDs in IS research is subject to several common shortcomings. Accordingly, we contribute to IS research by proposing and illustrating a set of procedural guidelines for SD development and usage. What distinguishes semantic differentiation from other measurement techniques is its reliance on linguistics in assessing the meaning of concepts (Osgood, Suci, & Tannenbaum, 1957). Therefore, developing and applying SDs demands not only proper measurement validation, but also meeting more-particular requirements related to semantics (e.g., Heise, 2010). General scale validation guidelines, however (e.g., MacKenzie, Podsakoff, & Podsakoff, 2011; Straub, Boudreau, & Gefen, 2004), do not address specific semantic requirements. In addition, where the SD literature does provide insight into these semantic requirements, scholars have treated them in a rather isolated and incremental manner over the years, without specifying how the requirements relate to prevalent and recent measurement validation procedures (Straub et al., 2004). We fill this methodological gap by synthesizing established scale validation and semantics requirements in a framework of suggested action for SD development and usage.

This paper is organized as follows. In Section 2, we briefly discuss the fundamentals of semantic differentiation, highlight prerequisites for applying the measurement technique, and distinguish five apparent common weaknesses in the application of the SD in published IS research. In Section 3, drawing on extensive literature study, we propose a framework of suggested action for SD development and usage. In Section 4 we subsequently describe an illustration exercise to exemplify how researchers could put the framework to practice. Using relevant theory, linguistic tests, expert interviews, pilot tests and data collected in three electronic marketplaces (EMs) in the Netherlands, this illustration exercise centers on developing a SD to assess the meaning of electronic marketplace quality (EMQ). Finally, in Section 5 we discuss our work, and conclude with recommendations for future research.

2. The Semantic Differential

2.1. Technique of Measurement

The SD is a technique to measure the psychological meaning of concepts, or a person's subjective perception of and affective reactions to the properties of concepts (Friedmann & Zimmer, 1988) through the use of bipolar scales or bipolar items[1]. Each of the bipolar scales that make up a SD consists of a pair of antonyms, which are usually two adjectives (e.g., difficult–easy; constrained–free). For more-complex concepts, researchers may develop more-elaborate bipolar scales by formulating contrasting phrases, in which the antonyms still remain the only two words that are opposite in meaning (e.g., difficult to use website–easy to use website) (cf. Dickson & Albaum, 1977; Hawkins et al., 1974). The opposites in each scale are linked in most cases by a continuum of seven or nine points[2] that respondents mark to show how they see the concept (Devellis, 2012). This form of measurement, in which the direction and the intention of meaning is controlled and allocated with bipolar scales, is what is known as semantic differentiation (Osgood et al., 1957).

The specific way in which semantic differentiation is conducted depends on whether the studied concept is single-dimensional or multidimensional in nature. A single-dimensional concept is measured via one set of bipolar scales that correlate well with one another (Devellis, 2012, p. 34). Following its focus on measuring the psychological meaning of concepts, this set of bipolar scales usually is operationalized with a reflective measurement approach (also see Hardin, Chang, & Fuller, 2008; Howell, Breivik, & Wilcox, 2007). For example, the well-known concept of user satisfaction with IS has been operationalized (e.g., Cenfetelli & Schwarz, 2011; Hong, Thong, & Tam, 2006) with one set of four closely related bipolar scales (see Figure 1). Based on the respondent's rating on the underlying bipolar scales, an overall score is computed, which will fall somewhere in the continuum of very low to very high. In SD terminology, this continuum is known *as semantic continuum (*Osgood et al., 1957).

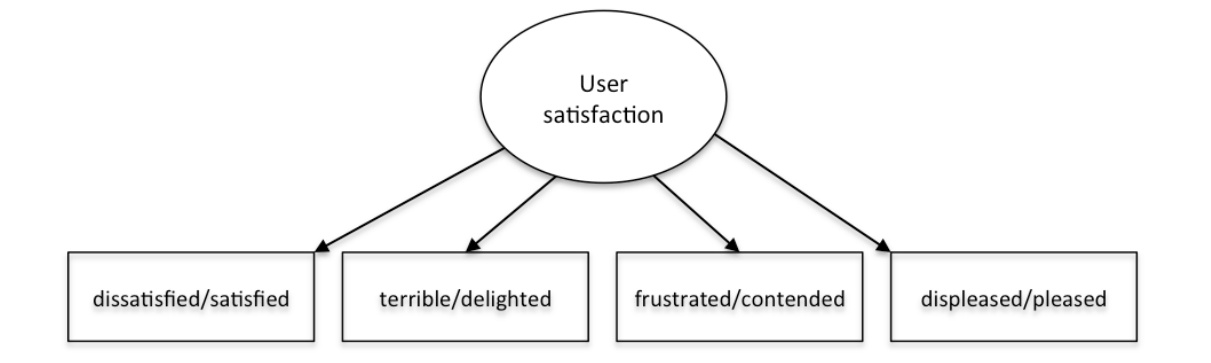

Figure 1. User Satisfaction Measured with the Semantic Differential

A multidimensional concept is formed by two or more related though still distinct dimensions that are each measured by a set of bipolar scales (cf. Netemeyer, Bearden, & Sharma, 2003). An example of such a multidimensional SD is online store image as conceptualized by Verhagen and Van Dolen (2009) (see Figure 2). Again, overall scores are calculated as described above, although this time for each of the dimensions (here: service, merchandise, atmosphere, navigation). This implies that a single semantic continuum no longer exists; rather, four semantic continua are used to score the overall meaning of the concept. In SD terminology, these semantic continua are also referred to as axes of a multidimensional space or semantic space.

[1] Both terms are used interchangeably in the SD literature, and we use both interchangeably here.

[2] Recent study draws attention to the value of continuous scaling as an alternative SD scaling format. Continuous scaling may have certain benefits in terms of the detection of small differences (i.e., less information loss) and suitability for more-robust factor analytic approaches (e.g., see Funke & Reips, 2012; Treiblmaier & Filzmoser, 2011).

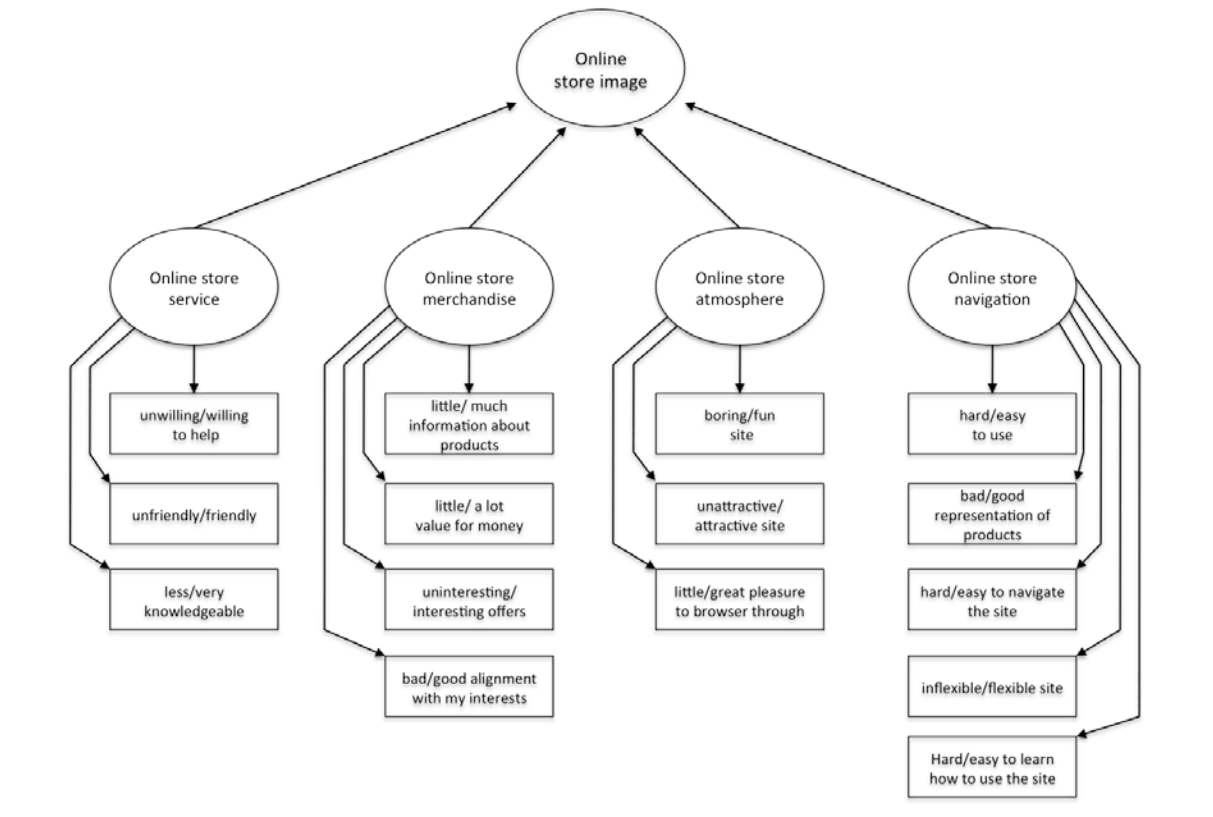
Online store image
Online store service
Online store merchandise
Online store atmosphere
Online store navigation
unwilling/willing to help
unfriendly/friendly
less/very knowledgeable
little/ much information about products
little/ a lot value for money
uninteresting/ interesting offers
bad/good alignment with my interests
boring/fun site
unattractive/ attractive site
little/great pleasure to browser through
hard/easy to use
bad/good representation of products
hard/easy to navigate the site
inflexible/flexible site
Hard/easy to learn how to use the site

Figure 2. Online Store Image Measured with the Semantic Differential

2.2. Requirements of Semantic Differentiation

The literature emphasizes that the SD is a measurement technique that requires careful consideration of the research context in terms of whether the selected bipolar scales fit the concept being judged (i.e. concept delineation) and the subject group being used (i.e., population specification) (Berthon, Pitt, Ewing, & Carr, 2002; Dickson & Albaum, 1977; Osgood et al., 1957).

To exemplify this notion, we refer to the SD used to measure user satisfaction as displayed in Figure 1. A first inspection of the bipolar scales gives rise to some questions on the general applicability of a scale anchor such as “delighted”, which represents feelings of great pleasure and is a rather emotional term rooted in consumer behavior research (see Laros & Steenkamp, 2005). Whereas “delighted” may be applicable in the context of hedonic IS such as online shopping environments and online gaming applications, it seems less suitable when the IS concept under study is more utilitarian in nature (e.g., ERP, office applications). Furthermore, whereas “delighted” may be an appropriate term to measure IS satisfaction among consumers, it may be less applicable for chief information officers. Thus, both the concept being measured and the subject group being used influences the applicability of the bipolar scales, which emphasizes the relevance of using bipolar scales that fit the research context.

Consequently, a solid validation of the terms used in the bipolar scales seems a key necessity when applying the SD. Such validation should not only incorporate generic measurement validation principles, but also require specific attention to the typical linguistic properties of the SD (Xiong, Logan, & Franks, 2006). In order to create an integrated and structured overview of SD validation issues, we conducted a systematic literature study. Using three academic databases (ScienceDirect, ABI/INFORM, and Wiley), we searched for relevant literature using search terms such as “semantic differential” and “bipolar scale” (abstract, title, keywords). We then searched in the selected papers for requirements and recommendations underlying the use of the SD. We selected the most widely mentioned requirements and organized them in line with existing measurement validation principles as proposed in the IS field (e.g., Lewis, Templeton, & Byrd, 2005; MacKenzie et al., 2011). Table 1 overviews these requirements and their corresponding key references from the academic literature[3].

Table 1. Requirements of Semantic Differentiation

#	Requirement	Description	Key references
1	Collection of a set of bipolar scales which covers the whole domain.	The bipolar scales should be widely distributed in meaning and be relevant to the concept under study to adequately constitute the semantic continuum or semantic space that is used to measure the concept under study.	Bearden, Hardesty, & Rose (2001), Hardesty & Bearden (2004), Hawkins et al. (1974), Kelly & Stephenson (1967), Osgood et al. (1957)
2	Linguistic and psychological bipolarity.	Scale anchors have to be bipolar from both a linguistic point of view and in relation to the concept being measured. The pure linguistic antonym of “engaging”, for instance, would be “disengaging”. When used in relation to the design of a website, however, a better opposite would be “unattractive”, since “disengaging” would not be perceived as a logical antonym by a website user.	Dickson & Albaum (1977), Falthzik & Johnson (1974), Heise (2010), Mindak (1961), Osgood et al. (1957), Schriesheim & Klich (1991), Snider & Osgood (1969)
3	Linguistic clarity.	The wording of bipolar scales should be clear in relation to each other and in relation to the concept being measured. In the case of SDs consisting of contrasting phrases, the wording in each contrasting phrase should also be clear.	Blair & Presser (1992), Foddy (2004b), Kahn & Cannell (2004), Reynolds, Diamantopoulos, & Schlegelmilch (2004)
4	Unidimensionality and distinctiveness of dimension(s) of the concept.	The dimension(s) that forms the axis of the semantic continuum (single-dimensional concept) or the axes of the semantic space (multi-dimensional concept) must be unidimensional and distinctive.	Gerbing & Anderson (1988), Gerbing & Hamilton (1996), Netemeyer et al., (2003), Osgood et al. (1957), Straub (1989), Xiong, Franks, & Logan (2003), Xiong et al. (2006)
5	Absence of contextual contamination.	The scales in the SD should be structured in such a way that respondent’s responses to preceding scales within the SD are not used as frame of reference for responses to the remaining scales.	Bickart (1993), Handling (1994), Landon (1971), Osgood et al. (1957), Tversky & Kahneman (1974)

Based on this overview of the basic requirements of semantic differentiation, two interesting questions arise that we address in Section 2.3: a) what would be the possible consequences if one did *not* sufficiently adhere to the principles? and b) how prominent is this lack of adherence in the current IS field?

2.3. Consequences and Prominence of Unawareness of SD Requirements

Neglecting the five mentioned requirements of semantic differentiation constitutes a shortcoming in the research design because it creates the risk of systematic measurement error, a consistent bias in measurement that negatively affects the psychometric quality of measurement and jeopardizes the adequate assessment of statistical relationships between concepts (Straub et al., 2004). Table 2 summarizes the most important consequences of such shortcomings.

[3] In line with the SD literature, these requirements directly relate to the process of semantic differentiation (see Section 2.1) and/or tap into the unique linguistic properties of the SD. We acknowledge that some (elements) of these requirements may apply to other measurement methods (e.g., Likert scaling). Following our research objectives, however, here we specifically focus on their role in semantic differentiation.

Table 2. Shortcomings in Semantic Differentiation and Measurement Consequences

#	Shortcoming	Potential implications in terms of systematic measurement error	Possible psychometric consequences**
1	Relevance of bipolar scales not tested.	Questionable content adequacy of the set of bipolar scales. The bipolar scales may not be sufficiently relevant and large to measure the concept's meaning (Netemeyer et al., 2003; Nunnally & Bernstein, 1994). This is likely to lead to neutral judgments when allocating the meaning of the concept on the bipolar scales (Osgood et al., 1957).	Limited content validity and face validity (Furr & Bacharach, 2008; Lewis et al., 2005; Raykov & Marcoulides, 2011; Straub, 1989). Artificially high reliability (Netemeyer et al., 2003; Nunnally & Bernstein, 1994). Unidimensionality brought about artificially (Netemeyer et al., 2003).
2	Missing linguistic and psychological bipolarity tests.	Disputable scale anchoring between both antonyms (linguistic bipolarity). Disputable scale anchoring between the antonyms in relation to the concept to be measured (psychological bipolarity) (Barrett & Russell, 1998; Green, Goldman, & Salovey, 1993). This is likely to lead to artificial and random skewness in the allocation of the meaning of the concept across the used set of bipolar scales (Cogliser & Schriesheim, 1994; Tourangeau, Rips, & Rasinski, 2000).	Limited content validity, in particular from a psychological bipolarity perspective (Osgood et al., 1957; Snider & Osgood, 1969). Lower reliability (Kerlinger, 1973; Netemeyer et al., 2003). Limited convergent and discriminant validity (Osgood et al., 1957; Schriesheim & Klich, 1991). Limited predictive and nomological validity (Cogliser & Schriesheim, 1994; Osgood et al., 1957; Schriesheim & Klich, 1991).
3	Within-scale linguistic combinations tests lacking.	Ambiguous and/or too complex wording of the bipolar scale items leading to comprehension problems (Groves et al., 2004). This is likely to trigger respondents to engage in neutral responses or to rely on their generic beliefs and opinions (automatic response tendencies) when allocating the meaning of the concept (Dillman, 2007; Podsakoff, Mackenzie, Lee, & Podsakoff, 2003).	Lower reliability of measurement (Fowler, 2009; Groves et al., 2004). In case of neutral responding: limited predictive and nomological validity (Coaley, 2010; Nunnally & Bernstein, 1994). In case of automatic responding: artificial improvement of predictive and nomological validity (Furr & Bacharach, 2008; Podsakoff et al., 2003).

Table 2. Shortcomings in Semantic Differentiation and Measurement Consequences (cont.)

#	Shortcoming	Potential implications in terms of systematic measurement error	Possible psychometric consequences**
4	Absence of dimensionality (pilot) test.	Uncertainty regarding the accuracy of the dimension(s) used to map the meaning of the concept in the semantic continuum/space (Furr & Bacharach, 2008; Nunnally & Bernstein, 1994; Snider & Osgood, 1969). It is likely that a lack of unidimensionality leads to wrong impressions of how respondents actually see the theorized concept (Bearden, Netemeyer, & Haws, 2011; Osgood et al., 1957).	Threat to unidimensionality (Lewis et al., 2005; Snider & Osgood, 1969). Lower reliability (Cortina, 1993; Lewis et al., 2005). Limited convergent validity, discriminant validity, predictive validity and nomological validity (Gerbing & Anderson, 1988; Netemeyer et al., 2003; Neuberg, West, Judice, & Thompson, 1997).
5	Contextual contamination not tested.	Uncertainty regarding the biasing effect of the presentation order of the bipolar scales (Krosnick, 1999; Lasorsa, 2003). It is likely that the meaning of the concept is measured inaccurately when respondents' ratings on bipolar scales systematically carry over to their ratings on later bipolar scales (Dillman, 2007; Krosnick, 1999; Schuman & Presser, 1996).	Threat to unidimensionality (Netemeyer et al., 2003). Limited discriminant validity (Schriesheim, 1981; Schriesheim, Solomon, & Kopelman, 1989). Artificial increase in convergent validity (Schriesheim, 1981; Schriesheim et al., 1989). Artificial increase in predictive validity and nomological validity (Doty & Glick, 1998; Furr & Bacharach, 2008).

** In this table, we focus on those psychometrics most commonly reported in the IS research field: unidimensionality, content validity, face validity, convergent validity, discriminant validity, predictive validity, nomological validity, and reliability.

Overall, the table illustrates the seriousness of the possible consequences of not testing whether the requirements of semantic differentiation are met, and thus highlights the need for adequate SD testing procedures.

To obtain insight into the extent to which these shortcomings are present in the IS field, we studied the papers published by six prominent academic IS journals from 2001-2010. To put our findings into a broader perspective, we also studied two leading journals in the marketing field and two in the management field[4]. The study resulted in a total sample of 269 papers that documented the use of one or more multi-item SDs, which confirms the widespread acceptance of the SD as measurement technique. Next, we assessed the extent to which the sample of papers was subject to the five identified shortcomings by coding each paper. Two members of the research team, who independently observed whether each paper did or did not heed the requirements of semantic differentiation, conducted the coding. Both researchers discussed the results of the coding and, in the cases of interpretation difficulties that were caused by ambiguous wording in the papers under study, together determined the most appropriate interpretation. Then, we consulted a panel of five academic experts (cf. Torkzadeh & Dhillon, 2002), all experienced IS researchers with a background in scaling procedures. After a brief introduction to our study and the expert panel's purpose, every expert

[4] We examined papers from the following journals using both computer-based bibliographic databases and issue-by-issue browsing: *MIS Quarterly*, *Information Systems Research*, *Journal of Management Information Systems*, *Decision Support Systems*, *Information & Management, European Journal of Information Systems*, *Journal of Marketing*, *Journal of Marketing Research*, *Management Science*, and *Organization Science.*

independently evaluated whether each of the 269 studies employing the SD was indeed subject to the shortcomings observed by the authors. Afterwards, the experts collectively determined the degree of correspondence between their evaluations. Overall, these evaluations were found to be in agreement with each other and with our findings. Table 3 shows the findings of the literature study.

Table 3. Literature Study on Shortcomings in Semantic Differentiation (2001-2010)

		Shortcomings				
		1	2	3	4	5
Journal	**# studies applying the SD N**	**Relevance of bipolar scales not tested. % (n)**	**Missing linguistic and psychological bipolarity tests % (n)**	**Within-scale linguistic combinations tests lacking % (n)**	**Absence of dimensionality (pilot) test % (n)**	**Contextual contamination not tested. % (n)**
MISQ	24	83% (20)	100% (24)	79%(19)	75%(18)	100% (24)
ISR	10	70% (7)	100% (10)	70% (7)	80% (8)	100% (10)
JMIS	22	100% (22)	100% (22)	59% (13)	73% (16)	100% (22)
DSS	20	65% (13)	100% (20)	70% (14)	50% (10)	85% (17)
I&M	42	60% (25)	100% (42)	71% (30)	60% (25)	100% (42)
EJIS	15	93% (14)	100% (15)	73% (11)	87% (13)	100% (15)
JM	62	53% (33)	100% (62)	71% (44)	76% (47)	100% (62)
JMR	55	76% (42)	100% (55)	85% (47)	93% (51)	100% (55)
MS	7	71% (5)	100% (7)	57% (4)	57% (4)	86% (6)
OS	12	67% (8)	100% (12)	42% (5)	92% (11)	100% (12)
Total	**269**	**70.3% (189)**	**100% (269)**	**72.1% (194)**	**75.5% (203)**	**98.5% (265)**

While part of the results may be attributed to non-reporting practices (a weakness in itself), the high percentages do give the impression that a considerable body of research seems to be subject to the identified shortcomings, both in the IS field and in the two related fields. Interpreting the SD rather narrowly as an alternative scaling format, which can be applied directly and universally without aligning it to the research context and without further explicit semantic testing, seems common practice. All in all, increasing IS researchers' awareness of how to apply the principles of semantic differentiation seems warranted.

3. A Framework for Semantic Differential Development and Usage

To stimulate more-adequate usage of the SD technique by the IS research community in the future, we suggest a set of procedural guidelines for SD development and usage. We have structured these guidelines in a framework, which synthesizes several semantic-testing and measurement validation procedures. Table 4 displays the framework.

Table 4. Framework for Developing and Applying Semantic Differentials

	Stage	Summary	Suggested actions
1	Establishment of a sample of valid bipolar scales.	Collection and applicability assessment of bipolar scales. These should be widely distributed in meaning and sufficient in terms of number and relevance.	In case of re-using a validated measurement instrument: • Collect a sample of existing bipolar scales, or • Convert an existing sample of Likert scales into bipolar scales. In case of developing a measurement instrument: • Generate a preliminary set of bipolar scales by making use of literature study, observation, or expert interviews. In any circumstance: • Use existing SD works or thesauri to decide which antonyms to include as scale anchors. • Assess the content validity of the bipolar scales in relation to the concept being judged through a pretest with experts.
2	Linguistic test of semantic bipolarity.	Assessment and confirmation of linguistic contrast and psychological bipolarity of the scale anchors.	Pretest for linguistic contrast with a sample of native speakers. Test for psychological bipolarity with expert panel, judging the linguistic alignment of each bipolar scale in relation to the concept under study.
3	Linguistic test of SD wording.	Evaluation and determination of comprehensible combinations of introductions, adjectives, verbs, and nouns in each bipolar scale.	Construct the draft questionnaire and, if necessary, translation of this questionnaire into the language it is to be administered in. Translators should be bilingual and have an understanding of the concept to be measured. Pretest with expert panel to evaluate the clarity and understandability of the bipolar scales, their introductions, and the involved concept(s).
4	Pilot test of SD dimensionality.	Assessment and establishment of the unidimensionality of the concept.	Pilot survey study to test the dimensionality of the SD. It is recommended to use a sample that matches the subject group being used for the final data collection. In case of both developing new SDs and re-using a validated measurement instrument: • Apply exploratory factor analysis. • Apply confirmatory factor analysis. Use model fit indices to assess unidimensionality. • After establishing unidimensionality: initial analyses of reliability, convergent validity, and discriminant validity.
5	Test of contextual contamination.	Assessment of the sensitivity of the SD to anchoring effects between the dimensions/items.	Conduct a measurement invariance test to test for possible anchoring effects. When anchoring effects are found, the researcher should: • Consider randomization of the order of the bipolar scales in the final application of the SD. • Make use of statistical solutions in the final application of the SD that aim to filter out the anchoring effects.
6	Application of SD.	Apply the SD in final data collection. Confirmation of good psychometrics.	• Final data collection • Apply exploratory factor analysis. • Apply confirmatory factor analysis to confirm unidimensionality (fit indices). • Test of reliability, convergent validity and discriminant validity. • Test of predictive validity and nomological validity of the SD.

3.1. Setup of the Framework

The framework consists of six steps. The first five steps correspond, both in terms of subject and of sequence, with the five SD requirements that are displayed in Table 1. The last step does not address an individual requirement, but rather concerns the actual application of the SD in its final form. It is important to state that, while the sequence of these steps represents a structured approach, the proposed framework should be interpreted as guiding rather than normative. Furthermore, we acknowledge that the framework may not contain a complete overview of all possible techniques available. Rather, we focused on those actions and techniques that received much attention in the literature and that seem feasible for the researcher in terms of time and technical knowledge.

We used three main criteria to develop the framework. First, the selected stages had to be well grounded in the established SD literature, yet also be in line with more recent instrument development procedures in the IS field. Second, following our research goals, we set priorities for the typical SD procedures because these have not received much attention in the IS field. Third, the suggested actions had to offer recommendations in situations of re-using existing SDs and developing new SDs to make the framework of practical value to a substantial number of researchers.

The framework differs from existing general paradigms for measurement development and validation in two main ways. First, it advocates particular attention for collecting the set of relevant bipolar scales (stage 1) and establishing SD dimensionality (stage 4). The fact that most SD applications fail to address these rather basic requirements of semantic differentiation (see Table 3) highlights the need for this focus. Second, the framework emphasizes the adequate use and application of linguistics by adding a novel stage for linguistic testing of semantic bipolarity (stage 2) and by proposing distinct stages for testing of wording (stage 3) and contextual contamination (stage 5). As such, it complies with the seminal works of Osgood et al. and colleagues who state that linguistics forms the crux of the SD method (e.g., Osgood et al., 1957; Snider & Osgood, 1969).

3.2. Description of the Stages

After delineating the research domain and defining the research concept, the first stage toward using a SD is establishing a sample of relevant bipolar scales. To establish such a sample, researchers can collect existing bipolar scales from one or more extant SDs, convert scales using other formats into bipolar scales (see e.g., Menezes & Elbert, 1979), and generate new bipolar scales based on literature study, observation, or expert interviews. Regardless of how the sample of bipolar scales is established, the SD literature prescribes that researchers need to test whether the scales, including the selected antonym pairs, are relevant for the focal concept (Dickson & Albaum, 1977; Sharpe & Anderson, 1972). Irrelevant (i.e., non-matching) concept-scale pairings are likely to jeopardize content validity or result in neutral responses (i.e., a check-mark in the middle of the scale) and, thus, reduce the amount of information gathered (Osgood et al., 1957, p. 78-79).

The next stage in the framework concerns ensuring the bipolarity of the selected scale anchors (Dickson & Albaum, 1977; Falthzik & Johnson, 1974), preferably by empirically pretesting for bipolarity (Cacioppo & Berntson, 1994). Such pretesting should concern both linguistic contrast and psychological bipolarity. Linguistic contrast implies that the scale anchors of each bipolar scale reflect a contrasting relationship from a purely linguistic point of view (Eggins, 2004); that is, the scale anchors function as grammatical antonyms (Yorke, 2001). Psychological bipolarity extends this view by assuming that the selected scale anchors are not only bipolar in isolation, but also in relation to the particular concept to be measured (functional antonyms; see Carroll, 1959). As such, psychological bipolarity demands linguistic matching of the polar terms to the concept under study to assure that these are also psychological opposites (Carroll, 1959; Yorke, 2001). An established procedure to test for linguistic contrast is a pretest with a sample of native speakers (Bachman, 1990). To assure psychological bipolarity, the researcher(s) should make use of an expert panel to test and establish the linguistic matching of each of the polar terms to the concept under study (Heise, 2010; Schriesheim & Klich, 1991).

The third step in the framework concerns a thorough testing of the SD's wording in the form (questionnaire) and language it is to be administered in. First, a draft questionnaire should be constructed that includes the concept(s), the bipolar scales, and introduction(s) to the bipolar scales. If necessary, this draft questionnaire needs to be translated into the native language of the respondents to be used in the final data collection (stage 6), preferably using bilingual translators who also have an understanding of the focal concept and its cultural meaning (cf. Sekaran, 1983). Next, the comprehensibility of the SD within-scale wording in the (translated) questionnaire should be established (Cliff, 1959; Osgood et al., 1957). An evaluation of the wording entails determining, preferably in a pretest, whether the bipolar scales and the text that introduces the scales are formulated in such a way that, combined in a questionnaire, they provide a comprehensible and unambiguous context for an individual to correctly interpret the studied concept (cf. Foddy, 2004a; Reynolds et al., 2004). For SDs consisting of descriptive phrases as opposites, the evaluation of wording should also encompass testing of the wording comprehensibility of the combination of adjectives, nouns, or verbs in each bipolar scale.

As a fourth step, in order to accurately allocate a concept’s meaning, it should be statistically determined that the bipolar scales assess a single concept (in the case of a semantic continuum) or only one underlying factor or dimension (in the case of a semantic space) (Landon, 1971; Osgood et al., 1957). Such unidimensionality or homogeneity (Clark & Watson, 1995) is a prerequisite for reliability and validity (Netemeyer et al., 2003; Ping, 2004). Accordingly, particular consideration has to be given to discovering and defining the SD’s dimensionality (Deese, 1964) using factor analytical procedures such as exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) (see e.g., Gerbing & Anderson, 1988; Netemeyer et al., 2003). EFA is useful to statistically identify both known and unknown factor structures underlying a SD (cf. Treiblmaier & Filzmoser, 2010), which are then to be validated using CFA (Gerbing & Anderson, 1988; Gerbing & Hamilton, 1996) via second-generation covariance-based structural equation modeling (SEM) software such as analysis of moment structures (Amos), linear structural relations (LISREL), and statistical analysis system (SAS). The data collected in the pilot study to establish the unidimensionality of the particular SD can also be used to make an initial assessment of whether the SD satisfies other conventional criteria for psychometric measurement (Dickson & Albaum, 1977; Osgood et al., 1957), such as reliability, convergent validity, and discriminant validity (Netemeyer et al., 2003).

In the fifth stage, the researcher should address response bias; that is, the tendency of respondents to respond systematically to questionnaire items on some other basis than the specific item content (Paulhus, 1991, p. 17). Whereas response bias is a common methodological issue that comes in many different forms (for an overview see Baumgartner & Steenkamp, 2001), the SD literature draws specific attention to contextual contamination (Landon, 1971; Osgood et al., 1957), or response bias due to anchoring effects (see Tversky & Kahneman, 1974), in which case respondents’ ratings on later scales in the SD are influenced by the beliefs that were rendered accessible when they responded to the preceding scales (Bickart, 1993, p. 52). Common explanations for contextual contamination include the affect heuristic, by which the difficulty of the first scales sets the mood of the respondents (Weinstein & Roediger, 2010, 2012); the primacy effect, which implies that respondents set the first scales as cognitive standards for answering later scales and therefore weight the first scales more heavily (Krosnick, 1999); and the recency effect, which holds that respondents continuously update and revise their beliefs and therefore arrive at more-thorough evaluations when answering later scales than first scales (Hogarth & Einhorn, 1992). Given that the SD typically consists of a set of scales grouped into one or more blocks, it is assumed to be relatively susceptible to contextual contamination (Weinstein & Roediger, 2012). Therefore, testing for contextual contamination is a necessity when applying the SD.

While it has been argued that a systematic approach to scale validation reduces the risks of anchoring effects (Burton-Jones, 2009; Podsakoff et al., 2003), such an approach does not fully rule out contextual contamination and its potential negative consequences (see Table 2). Therefore, empirical investigation of anchoring effects is recommended (Richardson, Simmering, & Sturman, 2009). Anchoring effects can be studied at the dimension or item level. For a SD intended to measure a multidimensional concept using different sets of bipolar scales for each dimension, researchers are advised to assess the degree to which the first measured dimension affects the subsequent evaluation of the other dimensions (Landon, 1971; Osgood et al., 1957). For a SD that is intended to measure a single-dimensional concept using multiple bipolar scales or a multidimensional concept using a single-item measure for each of the dimensions[5], contextual contamination could be tested on an item level instead of on a dimension level (see Landon, 1971). If a SD is found to be sensitive to contextual contamination, statistical solutions can be used in the final application of the SD that aim at filtering out any systematic order effects (for suggestions, see Podsakoff et al., 2003). Alternatively, randomizing the order of the bipolar scales for each respondent in order to minimize order bias may also be an appropriate answer to anchoring effects, making contextual contamination less likely to be an issue (cf. Krosnick, 1999).

Finally, as for any other measurement instrument, it is recommended to apply psychometric tests (see MacKenzie et al., 2011; Netemeyer et al., 2003) to establish whether the scales do indeed show the required psychometric qualities in the final data collection. If some scales do not pass these tests, the individual researchers will have to decide whether they wish to improve the psychometrics or model fit by removing the problematic scales, preferably in a test-retest setting with independent datasets, or want to retain these scales because removing them may limit content and face validity (Netemeyer et al., 2003). Re-using data collected in previous phases is a possibility but, since this is likely to induce bias, cross-validation is preferable (Cooil, Winer, & Rados, 1987). A rather robust cross-validation method is validity generalization (Murphy, 2003), a test of the weights (scale loadings; inter-item correlations; regression/path coefficients) found in one situation against those found for a second sample from the population. Given that semantic differentiation is a concept-tailored method, validity generalization across different concepts is not recommended. Rather, researchers should test the same concept across comparable subjects (cf. Wiggins & Fishbein, 1969) in a range of related situations relevant to what is being measured. As such, the comparability of the concept's meaning can be tested most directly (Murphy, 2003; Osgood et al., 1957).

4. Illustration Exercise: Assessing the Meaning of EMQ

Since “instrument validation may be best understood by seeing how validation can be applied to an actual MIS research problem” (Straub, 1989, p. 154), we describe the suggested application of each stage of the framework via an illustration exercise. The purpose of this exercise is to exemplify the tests that researchers may apply in each of the stages by making some procedural suggestions. The illustration exercise focuses on assessing the meaning of the concept electronic marketplace quality (EMQ) by developing and applying a SD. Rooted in an established and growing field of electronic marketplace (EM) studies (for an extensive overview, see Standing, Standing, and Love, 2010), EMQ refers to buyers' quality perceptions of consumer-to-consumer (C2C) EMs[6]. Drawing on works on website quality, EMQ is expected to be multidimensional and rather complex in nature (cf. Yang, Cai, Zhou, and Zhou, 2005), making the SD one of the most appropriate techniques of measurement.

C2C EMs have been the context of several empirical explorations (Verhagen, Meents, & Tan, 2006; Standing et al., 2010). Remarkably, a well-conceptualized and validated instrument to measure consumers' overall perception of an EM is lacking. We refer to this overall perception as EMQ, defined as a mixture of evaluative responses (beliefs) that are derived from all kinds of implicit (i.e., imperceptible, psychological) and explicit (i.e., observable, concrete) functions and services provided by the intermediary and the population of sellers (Sarkar, Butler, & Steinfield, 1995). Examples of functions and services made available by the intermediary include providing the technological infrastructure, logistical settlement, and control mechanisms. Sellers expand these functions and services by offering sales related functions such as product selection, product description, and provision of contact information. Given that it comprises multiple related perceptions, this research conceptualizes EMQ as a composite evaluative concept that is rather complex, multidimensional, and reflective in nature[7].

[5] In situations where a concept, in the raters' minds, refers to an easily and uniformly imagined object, has low complexity, and relates an easily and uniformly imagined characteristic to the object, single-item measures might be a viable alternative to multi-item measures (Christophersen & Konradt, 2011).

[6] We interpret a C2C EM here as an online environment with specific boundaries that is supported and enabled by a combination of IT and various services, procedures. and regulations offered by a third-party intermediary, in which consumers and sellers are matched, trust is provided, information about products and prices is exchanged, and transactions can be closed (O'Reilly & Finnegan, 2010; Standing et al., 2010).

[7] A reflective approach seemed most appropriate given the focus on EMQ as a well-developed mental evaluation by the respondent that can be expected to be already in existence prior to its measurement by the researcher (Bagozzi, 2007; Marakas, Johnson, & Clay, 2008).

Table 5 summarizes the specific actions we have taken in applying each of the stages of the framework. In the remainder of this section, we explore these actions and their empirical outcomes in the context of the EMQ example.

Table 5. Actions Taken in the Development of a SD to Measure the Meaning of EMQ

	Stage	Actions taken
1	Establishment of a sample of valid bipolar scales.	1. Literature study in the EM field and subjective content analysis of EMs, resulting in a preliminary set of 69 bipolar scales. 2. Selection of antonyms for each bipolar scale by making use of the SD literature and linguistic works. 3. Pretest with academic researchers and EM practitioners using: - A rating procedure and sum score technique in order to assess the applicability and relevance of the bipolar scales. - A free association technique in order to assess the domain coverage of the sample of bipolar scales. 4. Removing/adding/rewording of the scales based on the pretest, resulting in an updated list of 69 bipolar scales.
2	Linguistic test of semantic bipolarity.	1. Pretest for linguistic contrast with a sample of native speakers using a missing word technique. Each subject was confronted with a list of either the positive or the negative bipolar scale anchors and asked to fill in the missing linguistic opposites. The results did not indicate any need for refinement. 2. Pretest for psychological bipolarity with academic researchers and EM practitioners to judge the linguistic alignment of each bipolar scale in relation to the EMQ concept by using yes-no and open ended questioning. Some refinements in wording were made.
3	Linguistic test of SD wording.	1. Construction of draft questionnaire containing the 69 bipolar scales. 2. Translation of draft questionnaire into the language it is to be administered using back-translation and the parallel/ double translation technique. 3. Pretest with academic experts using: - A rating procedure and open-ended questioning technique to assess the extent to which each bipolar scale and its introduction is subject to commonly made faults in bipolar scale item wording. - Open-ended questioning to assess the extent to which the overall questionnaire may be subject to other faults threatening the understandability of the bipolar scales. 4. Editing/rewording of the questionnaire based on the pretest.
4	Pilot test of SD dimensionality.	1. Data collection through a pilot survey among a student sample. The students visited four different EMs and completed each visit with filling in an online questionnaire containing the preliminary 69 bipolar scales. 2. Item sorting exercise with academic researchers and EM practitioners to get a first understanding of the dimensionality of the SD. An eleven-dimensional SD to measure EMQ was suggested. 3. EFA (principal components analysis with varimax rotation) on the data aggregated across the four EMs. After dropping 12 bipolar scales, the outcome was a twelve-dimensional SD. 4. CFA (Amos 20 with maximum likelihood estimation) on the twelve-dimensional SD structured as correlated first-order model. After removing 7 bipolar scales, the results provided support for the unidimensionality, convergent validity, and discriminant validity of the twelve-dimensional SD consisting of 50 bipolar scales. 5. Reliability analyses. All Cronbach's Alpha's were acceptable.

Table 5. Actions Taken in the Development of a SD to Measure the Meaning of EMQ (cont.)

	Stage	Actions taken
5	Test of contextual contamination.	1. Data was collected through a survey among undergraduate students. The students visited an EM and completed this visit by filling in an online questionnaire. Three versions of the questionnaire containing the 50 bipolar scales were constructed and at random distributed among the students: - A version with the twelve EMQ dimensions and their items in the order as tested in the previous steps of the process. - A version with the twelve dimensions in a reversed order. - A version with the bipolar within each of the twelve dimensions in a reversed order. This resulted in three datasets that were used for further testing. 2. Measurement invariance testing: factor loadings. Differences in factor loadings across the three datasets were examined by using the coefficient of concordance. No indications for anchoring effects were found. 3. Measurement invariance testing: inter-item correlations. Differences across the three datasets between the inter-item correlation matrices within each of the twelve dimensions were examined by applying Fisher's z transformation test. No indications for anchoring effects were found.
6	Application of SD.	1. Data was collected via large-scale online surveys conducted among real visitors of two EMs: eBay.nl and the EM with the largest market share in The Netherlands. 2. EFA (principal components analysis with varimax rotation) on subsets of each of the two datasets. The results confirmed the unidimensionality of the twelve EMQ dimensions. 3. CFA (Amos 20 with maximum likelihood estimation) on the remaining data of each of the two EMs. Following the results of the EFA, a correlated first-order model consisting of the twelve EMQ dimensions was tested. After removing 13 bipolar scales the measurement model fitted well to the data. Two alternative model structures were also tested but rejected due to a lack of fit with the data. Overall, the results supported the unidimensionality of the EMQ dimensions and the multidimensionality of the 37-item EMQ scale. 4. Reliability and validity analyses. Using the data of the two EMs, support for the reliability, convergent validity, discriminant validity, and predictive validity of the EMQ scale was found. 5. Cross-validation. New data was collected via a third EM. Visitors of this EM were invited to participate an online survey at two different sections the EM. The two independent subsamples were used for CFA (Amos 20 with maximum likelihood estimation). The results reconfirmed the unidimensionality of the EMQ dimensions and the multidimensionality of the EMQ scale. Subsequent testing again provided supports for the reliability, convergent validity, discriminant validity, and predictive validity of the 37-item EMQ scale.

4.1. Establishment of a Sample of Valid Bipolar Scales

We derived a preliminary set of 69 bipolar scales from the literature (e.g., Lin, Janamanchi, & Huang, 2006; Pinker, Seidmann, & Vakrat, 2003) and subjective content analysis (cf. Mindak, 1961). We based antonyms on the list of 50 most frequently appearing antonyms in Osgood et al. (1957), with additional antonyms based on existing SD scales (e.g., Dickson & Albaum, 1977; Hawkins et al., 1974) and linguistic works such as Webster's Collegiate Thesaurus (Weir Kay, 1994).

The preliminary set of scales and antonyms was pretested among expert panels consisting of eleven researchers with relevant experience with scale development and linguistics, and seven practitioners working for two C2C EMs. In a formalized rating procedure (cf. Hambleton & Rogers, 1991), the experts judge the degree to which each pair of antonyms is: 1) applicable to the EMQ concept, 2) relevant to measuring EMQ, and 3) the degree to which the total item pool covers the domain of the EMQ concept (content validity) (cf. Netemeyer et al., 2003). The experts expressed their opinion on seven-point rating scales ("very inapplicable" to "very applicable" and "very irrelevant" to "very relevant"), explained their opinions, and suggested improvements. Based on the results, we considered items with an average applicability rating lower than 6 (i.e., "quite applicable"; "quite relevant") for rewording or deletion (cf. Bearden et al., 2001). We also took the explanations given by the experts into consideration in this decision. Finally, to judge the domain coverage of the pool of bipolar scales (cf. Netemeyer et al., 2003), we asked the experts to suggest additional items based on an overview of the preliminary items. We selected additional items if they were applicable to our definition of the EMQ concept and were mentioned multiple times by the experts. Based on the outcomes, we edited the items and reworded them where necessary. Of the initial item pool of 69 items, 9 items remained unedited, 23 items were reworded, 35 items were removed, and 35 new items were added. The result was an updated item pool of 69 items.

4.2. Linguistic Test of Semantic Bipolarity

We adopted the missing word technique (Dickson & Albaum, 1977) to test the linguistic contrast of the distinct scale anchors. We split the list of antonyms into one list of positive anchors and one list of negative anchors, and we subsequently asked subjects (a convenience sample of native English speakers) to fill in the missing linguistic opposites. Using native speakers ensured linguistic homogeneity so that respondents applied the same meaning to the antonyms being presented to them (Dickson & Albaum, 1977). Our friends and relatives in the USA were sent an email invitation with a hyperlink to a webpage that automatically redirected each respondent to one of the two online questionnaires. To ensure that the respondents had a higher than average intelligence (enabling them to better differentiate between the meaning of words) (cf. Osgood et al., 1957), we decided only to include respondents who had attended college. The final data set comprised 32 respondents. The results of the test confirmed the selected list of anchors and indicated that none of the anchors needed to be reworded.

Psychological bipolarity was pretested with the same expert panel used in stage 1. For each antonym pair in the preliminary list of 69 semantic phrases, the experts judged whether each of the two anchors aligned with the focal concept (cf. Heise, 2010) by filling in a “yes-no” question, and by additionally explaining their opinions and suggesting improvements. Based on their answers, we slightly modified the wording of some anchors.

4.3. Linguistic Test of Semantic Differential Wording

We combined the 69 bipolar scales in an American English draft questionnaire, which we subsequently had translated into Dutch, the language in which it was to be administered in the final data collection. We selected two translators with an academic background in both languages and an appropriate understanding of the research domain. A combination of the back-translation and the parallel or double translation technique (Malhotra, Agarwal, & Peterson, 1996) was used. First a bilingual speaker (native: Dutch) translated the English questionnaire into Dutch. A second bilingual speaker (native: English) compared this Dutch questionnaire to the original English questionnaire. Afterwards, both bilingual speakers discussed the appropriateness of the translation (cf. Hong & Tam, 2006) and agreed that the semantics were comparable to the original American English questionnaire.

To evaluate the wording of each bipolar scale and its introduction, we held a pretest using an expert panel (cf. Foddy, 2004a) of three academic experts with a background in both questionnaire design and e-commerce research. First, we asked the experts to evaluate the wording of each bipolar scale and its introduction. Based on a list of commonly made faults in item wording (e.g., Foddy, 2004b; Reynolds et al., 2004), we formulated four questions that probed whether a bipolar scale or its introduction: 1) contained words that would be incomprehensible to the intended respondents, 2) were worded simply enough, 3) were clear enough, and 4) referred to a clear context. The experts rated the scales on these criteria using the labels “certainly”, ”possibly” and ”no” (cf. Cannell, Fowler, Kalton, Oksenberg, & Bischoping, 2004). They also expressed their thoughts and considerations behind each rating in an open question. Secondly, we presented the overall questionnaire to the experts. We then asked them whether the questionnaire contained any other faults that could lead to incorrect interpretations or problems with completing the questionnaire. This led to some modifications: we changed some words into their synonyms and shortened some bipolar scales without changing their meanings. The modifications resulted in a preliminary measurement instrument consisting of 69 bipolar scales.

4.4. Pilot Test of Semantic Differential Dimensionality

We collected the data for the pilot study through a laboratory experiment using a convenience sample of 196 students in an IS course at a Dutch university. We instructed the participants to study four different EMs (eBay.nl and three Dutch C2C EMs) and to focus on the purchase of a digital camera, after which they filled in an online questionnaire addressing perceptions of EMQ.

There was no literature available to provide us with an initial clear indication of the dimensionality of the EMQ concept. In such cases, it is recommended to conduct an item-sorting exercise before starting with the factor analyses (Allport & Kerler, 2003). We asked the eighteen researchers and practitioners who previously evaluated the bipolar EMQ scales (see 3.1) to participate in a sorting exercise following the procedure described by Wolfinbarger and Gilly (2003). Based on our expertise and the statements from experts in earlier pretests, we prepared a preliminary classification of the EMQ bipolar scales into 11 dimensions. The experts then judged this preliminary classification for provisional construct validity by the experts, confirming the eleven dimensions.

Next, we ran principal components analysis with varimax rotation. Since analyzing the differences between the four selected EMs was not the objective of this particular study, the data set used in the analysis consisted of the scores of the respondents aggregated over the four studied EMs (cf. Srinivasan, Vanden Abeele, & Butaye, 1989). We excluded bipolar scales showing factor loadings of 0.40 or higher on more than one factor from subsequent re-specifications of the factor model in order to achieve an adequate level of preliminary unidimensionality (Hair, Anderson, Tatham, & Black, 1998; Ping, 2004). After removing 12 of such scales, the principal component analysis resulted in a preliminary twelve-dimensional solution of 57 bipolar scales (KMO MSA: 0.88; Bartlett's test of sphericity: 11699, p < .001), accounting for 74.82 percent of the variance. The factor solution reproduced 10 out of 11 dimensions from the item-sorting exercise. The results indicated that the one dimension that was not reproduced consisted of two separate dimensions. Then, we applied CFA (Amos 20, maximum likelihood) to the twelve-factor solution structured as a correlated first-order model (cf. Doll, Xia, & Torkzadeh, 1994; Yang et al., 2005); that is, the twelve extracted basic dimensions functioning as inter-correlated first-order factors. After deleting seven scales to improve the fit of the model to the data, the CFA showed an acceptable fit ($\chi 2$ = 1886.644, df = 1109, p < .001; $\chi 2$/df 1.701; GFI .90; AGFI .88; RMR .064; RMSEA .057; NFI .93; TLI .93; CFI: .94). This finding confirmed the twelve-factor conceptualization of EMQ, the unidimensionality of each underlying dimension, and provided first evidence of convergent and discriminant validity (Netemeyer et al., 2003; Ping, 2004). Computation of Cronbach's alphas substantiated the reliability of each dimension and indicated that item redundancy was very unlikely to be an issue because all alpha's surpassed the value of 0.70 but did not exceed the value of 0.95. The end result of this phase was a preliminary SD of 50 scales.

4.5. Test of Contextual Contamination

To test for contextual contamination, we collected data using a quasi-experimental design with 192 students in an IS course at a Dutch university. The tasks in the experiment consisted of studying an EM and completing an online questionnaire addressing perceptions of EMQ. We selected eBay.nl and the Dutch EM, which have the largest market share in the Netherlands. We directed the participants at random to one of the two EMs. Three different versions of the questionnaire were randomly assigned to the participants. Following Landon (1971), the first version of the questionnaire presented the twelve EMQ dimensions and their items in the same sequence as previously identified. In the second list, the twelve dimensions were presented in reversed order. The third version of the questionnaire reversed the order of the items in each dimension. For each of these three versions of the questionnaire we aggregated the scores of the respondents over the two studied EMs. This resulted in three independent datasets.

To assess the sensitivity of the EMQ scale to contextual contamination, we conducted a measurement invariance test (cf. Nye, Roberts, Saucier, & Zhou, 2008; Rigdon, Ringle, & Sarstedt, 2010; Sarstedt, Hensley, & Ringle, 2011) by testing for differences in the factor loadings and inter-item correlations across the three datasets (see Carte & Russell, 2003). We used partial least squares modeling (software package SmartPLS, see Ringle, Wende, & Will, 2005) to compute the factor loadings of the twelve dimensional measurement model (see Appendix A, Table A-1.). Kendall's coefficient of concordance (W) (for formulas see Legendre, 2005, 2010) was 0.77 (Chi-square 112.68, df= 49, p < 0.001)[8], which suggests an acceptable level of concordance between the factor loadings across the three datasets (cf. McBride & Wolf, 2007). Inter-item correlations (Pearson's *r*) were computed in each dimension across the three datasets[9]. We then used Fisher's Z transformation tests (for formulas and procedures see Lomax, 2001; Meng, Rosenthal, & Rubin, 1992) to test for any statistical significance[10]. For each inter-item correlation in the dataset collected with the first version of the questionnaire, we conducted a Z-test with its equivalent as computed with a) the data collected with the dimension-reversed questionnaire and b) the data collected with the item-reversed questionnaire. All computed Z-values were below the recommended critical value of 1.96 (p < .05), which suggests that no significant differences between the inter-item correlations across the three datasets were found. Because there were no significant differences in factor loadings or inter-item correlations between the different datasets, the test did not reveal potential contextual contamination weaknesses in the SD.

4.6. Application of Semantic Differential

We applied the EMQ scale consisting of the 50 bipolar scales (see appendix A, Table A-1.) on two independent samples: 1428 visitors of eBay.nl (EM1), the Dutch version of eBay.com; and 1051 visitors of the Dutch EM with the largest market share of EMs in the Netherlands (EM2)[11]. Visitors were invited to participate in the survey through banners in the digital camera section of each EM, which suggests a product focus similar to our pilot testing.

An initial test of the scale dimensionality using principal components analysis with varimax rotation, on subsets of 500 respondents from each of the datasets, provided preliminary though strong evidence for unidimensionality. To further validate the extracted latent structure, we applied CFA (Amos 20, maximum likelihood) to the remaining data on each of the two EMs (sample 1: n = 928; sample 2: n = 551). The fit indices of the initial solutions of the correlated first-order model highlighted the need for model improvement. Following suggestions made in the literature on SEM (e.g., Evermann & Tate, 2009; Gerbing & Anderson, 1988), we then studied the pattern of residuals to locate misspecifications. Thirteen items (see Appendix A, Table A-1.) shared large positive and negative residuals with items of other factors. Acknowledging that item deletion should not solely be based on statistical grounds (Byrne, 2010), we also took the face validity and content validity of these items into consideration. Based on both this relatively subjective analysis and the size of the residuals, we decided to delete all 13 items and re-estimate the model. The re-estimated model not only showed a good fit with the data (sample 1: $\chi 2$ = 1226.605, p < 0.001; CMIN/df: 2.183; GFI: 0.93; AGFI: 0.92; RMR: 0.059; RMSEA: 0.036; NFI: 0.96; TLI: 0.97; CFI: 0.98; sample 2: $\chi 2$ = 1154.230, p < 0.001; CMIN/df: 2.054; GFI: 0.90; AGFI: 0.87; RMR: 0.087; RMSEA: 0.044; NFI: 0.94; TLI: 0.96; CFI: 0.97), but also outperformed two alternative models that we tested (see Appendix B). The outcomes supported the unidimensionality of the factors (Ping, 2004) and the multidimensionality of the EMQ scale. Table 6 shows the resulting twelve EMQ dimensions.

[8] The significance of the chi-square test suggests that not all observed factor loadings may be concordant with each other (Legendre, 2005). It is well known, however, that the Chi-square test is a relative conservative test, which is relatively susceptible to sample size (Legendre, 2010). Furthermore, following Siegel and Castellan (1988), either a high or significant value of the coefficient of concordance implies that an acceptable level of overall concordance has been reached.

[9] For illustration, Table A-2 shows these inter-item correlations for the first (layout) and last (meeting sellers) dimension of the EMQ concept.

[10] We did not use Hotelling's *t*-test, proposed by Carte and Russell (2003), to test for differences in correlation coefficients due to its relative sensitivity to type I errors (see Meng et al., 1992).

[11] The sample characteristics of both samples are available from the authors on request.

Table 6. Overview of EMQ Dimensions and Their Definitions

EMQ dimension	Definition
Layout	The buyer's experience of the layout of the EM as being attractive and up-to-date.
Ease of use	The perceived usability of the EM, including navigation options, site structures and ease of learning how to use it.
Contacting the intermediary	Perceptions of the amount of information and options provided at the EM that enable buyers to get in touch easily with the intermediary facilitating the EM.
Institutional control	Perceptions of the measures applied by the intermediary, such as guarantees, privacy policy, and rules, to protect buyers and regulate the EM.
Community	The perceived ability of buyers to share their experiences and communicate with other buyers.
Contacting sellers	Perceptions of the amount of information and options provided at the EM that enable buyers to get in touch with sellers easily.
Seller information	The perceived amount and clarity of the information provided about sellers and their reputation.
Product information	The impression a buyer has about the way sellers describe and represent the products offered at the EM.
Pricing mechanisms	The perceived clarity and convenience of the mechanism that is used to establish and communicate prices at the EM.
Assortment	Overall buyer's perception of the assortment at the EM, including a) the size of the assortment and b) alignment of the assortment with one's interests.
Settlement	The ease and clarity of methods used for paying and receiving products bought at the EM, as perceived by buyers.
Meeting sellers	The buyer's perceived ease of meeting sellers in offline settings to inspect, pay for and pick up products.

Subsequent quantitative tests supported the reliability and the convergent and discriminant validity of the 37-item SD (appendix C). Using the behavioral variable attitude toward purchasing as the dependent variable in multiple regression and ridge regression, we also confirmed predictive validity (Table C-1 and C-2 in Appendix C). Overall, all resulting statistics support the applicability of the EMQ scale.

Finally, to further cross-validate the SD, we collected two new independent datasets in a third C2C EM, a Dutch EM facilitated by a newspaper publisher. Banners that invited visitors to complete an online questionnaire were placed in the automobile section and the study books section of the EM. The questionnaire addressed the 37 bipolar EMQ scales and the attitude toward purchasing (cf. data collection EM1 and EM2), while we added the intention toward purchasing, website satisfaction, and loyalty intentions for the purpose of extended predictive validity testing (cf. Wolfinbarger and Gilly, 2003). The final samples included 863 visitors of the automobile section (sample 3), and 590 visitors of the study books section (sample 4)[12]. We used CFA (Amos 20, maximum likelihood) to re-assess the dimensionality of the correlated first-order model. Except for the chi-square tests, all fit indices demonstrated very good fit with the data for both samples (Sample 3: $\chi2$ = 1352,609, df = 562, p < .001; $\chi2$/df 2.407; GFI .92; AGFI .90; RMR .062; RMSEA .040; NFI .96; TLI .97; CFI: .98. Sample 4: $\chi2$ = 1039,206; df = 562; p < .001; $\chi2$/df 1.849; GFI .91; AGFI .89; RMR .049; RMSEA .038; NFI .95; TLI .97; CFI: .98). As such, the results strongly reconfirmed the twelve-dimensional meaning of the EMQ scale. We then tested for reliability and validity following the exact procedures applied for sample 1 and 2. The results again demonstrated the reliability and the convergent, discriminant, and predictive validity of the EMQ scale (see Appendix D).

[12] The sample characteristics of both samples are available from the authors upon request.

5. Discussion and Conclusion

During the last two decades, scholars have put considerable effort into increasing the methodological rigor of IS research; for example, via methodological publications on measurement validation in general (e.g., Lewis et al., 2005; MacKenzie et al., 2011; Straub et al., 2004) and the application of the SD in particular (e.g., Chin et al., 2008). Despite this cumulative effort and all the new insights it has generated, it appears that many applications of the SD in IS research do not conform to the basic principles underlying this common measurement technique. Accordingly, we make scholars aware of these principles by showing them the relevance of adhering to these principles, and providing suggestions for how they could apply the SD more adequately. Based on the SD literature and recent methodological insights, we propose and illustrate an integrated framework of suggestions for developing and applying SDs.

The framework highlights the need for adopting and extending established guidelines for measurement scaling by emphasizing the establishment of a set of relevant bipolar scales (stage 1) and dimensionality testing (stage 4), by adding a novel stage of semantic bipolarity testing (stage 2), and by proposing distinct stages for wording (stage 3) and contextual contamination (stage 5). As such, it adds to the calls of Straub, Hoffman, Weber, and Steinfield (2002) and MacKenzie et al. (2011) to extend measurement methods in the academic IS community.

5.1. Contributions

A significant contribution of this study is that we clarify the requirements of semantic differentiation and the provision of corresponding suggested actions. Since linguistics forms the heart of semantic differentiation (Osgood et al., 1957), the study underlines the key role of bipolar scale selection, particularly because the scales determine the axis of the semantic continuum or the axes of the semantic space that is used to measure the meaning of the concept. Accordingly, we provide suggestions for bipolar scale selection and judgment. Moreover, we draw attention to the semantic relevance of bipolarity and clarity of wording combinations; we propose and empirically illustrate directions for testing linguistic and psychological bipolarity and within-scale wording. Finally, to define the axis of the semantic continuum or the axes of the semantic space, and in the case of the latter to ensure the distinctiveness of the factors representing the axes, the framework suggests the adoption of dimensionality pilot testing and testing of contextual contamination.

The framework we developed is quite extensive and need not always be applied in full. The full framework is especially relevant when a new SD is developed, but may be more selectively applied in studies with a more confirmatory nature. For developing new SD scales in relatively new or unexplored research fields, we strongly recommend fully adopting our framework. Still, in practice, researchers may obviously also conduct confirmatory studies by using an existing SD to investigate the same concept in the same phenomenal context (i.e., population in terms of socio-demographics, nationality, and cultural background; see Berthon et al., 2002) as in the original study, either to replicate or theoretically extend existing nomological structures. In such cases, no further testing of whether the SD requirements have been met is necessary if such testing was already done in the study in which the particular SD originated. Applying an existing SD to investigate the same concept, though in a phenomenal context that differs from the one studied in the original article, is another common practice. If the original paper confirms that the focal SD has already passed content validity tests, the available sample of bipolar scales can be considered representative and relevant. The SD, however, remains a context-specific technique. Therefore, we do recommend researchers conducting contextual extensions to pay attention to bipolarity, wording clarity, unidimensionality, and contextual contamination. Bipolarity and wording pretesting could be added rather easily to pretests already planned in a research project. Dimensionality and contextual contamination could be tested in a small-scale pilot. If for any reason such pilot testing is difficult (e.g., constraints in time, budget, or sample availability), researchers could choose to make use of the data of the final data collection. Scholars considering this approach have to weigh the benefit of increased efficiency against the risks of finding factor solutions that differ from the prespecified theoretical conceptualization and of being confronted with order biases that do demand additional post-hoc statistical remedies (cf. Podsakoff et al., 2003).

5.2. Limitations and Future Recommendations

Due to the focused scope of our work, it has several limitations. First, our framework is intended as a set of procedural guidelines and not as an exhaustive, detailed overview of methodological techniques and accompanying instructions for each of the described stages. Other IS researchers are invited to add to or refine some of the exemplified techniques for each of the framework's stages. One possible extension of our framework concerns further bipolarity testing. By addressing linguistic and psychological bipolarity, the framework addresses two important aspects of bipolarity. Still, bipolarity could be investigated even more rigorously by also adding metric testing of: 1) the midpoint of each semantic scale pair (see Cogliser & Schriesheim, 1994), and 2) the equidistance of the value scales to their midpoint (see Schriesheim & Klich, 1991). Another possible extension concerns the inclusion of end users in testing the wording of the SD (stage 3). Following the established pretest literature (e.g., Foddy, 2004a), we suggested and made use of an expert panel to conduct the linguistic testing. Still, given that the interaction of end users with the IS artifact under study belongs to the core properties of the IS field (Benbasat & Zmud, 2003), the use of end users could be a valuable extension.

Second, based on existing evidence that SDs have the potential to outperform other commonly used scaling types (Chin et al., 2008) and given the apparent room for improvement in semantic differentiation in IS research, we focused on guiding IS researchers on how to go about developing and using SDs. Accordingly, an empirical comparison of the psychometric properties yielded by the SD versus other commonly used types of measurement instruments in IS research (e.g., Likert scaling) fell outside our scope. This is not to say that such an empirical comparison would be of no interest—on the contrary. Weaker psychometric properties of certain scaling methods typically result in systematic measurement error, which reduces the amount of explained trait variance (Cote & Buckley, 1987), and thus has a substantial impact on research outcomes. Future research could shed light on how certain types of measurement instruments compare in terms of psychometrics and explained variance when measuring particular concepts in the IS research field.

Third, while the framework can be applied directly to the vast majority of concepts in the IS research domain because these concepts can be measured on scales consisting of opposite states that are psychologically meaningful (Cenfetelli, 2004), some caution seems required for so-called dual-factored concepts. These concepts consist of two separate though closely related constructs that are more than just the opposite of each other (Cenfetelli & Schwarz, 2011). A typical example of a dual-factored concept is positive affect/negative affect. Both constructs are considered as two independent parts of the concept "emotion" (Laros & Steenkamp, 2005) that do coexist next to each other, and may have asymmetric effects on cognitive processing and decision making (Bagozzi, Gürhan-Canli, & Priester, 2002; Griskevicius, Shiota, & Neufeld, 2010). Dual-factored concepts such as this should be conceptualized as two different and separate constructs before the SD framework is applied to each of them. Conceptualizing the two constructs as one bipolar concept and using the two factors as antonym pairs in a bipolar scale to measure this concept would not only make little sense from a theoretical perspective, it would most likely translate into a semantic continuum that has an arbitrary midpoint and as such could induce considerable systematic measurement error.

Fourth, we introduce the guidelines for adequate semantic differentiation suggested in the framework to deal with some of the major shortcomings in SD use as observed in the IS field. As such, these guidelines specifically tap into the characteristics of the SD. This is not to say, however, that paying close attention to some facets that we discuss would not benefit more-appropriate use of other measurement instruments as well (e.g., Likert scaling). For example, we believe that testing the content validity of measurement items in relation to the concept under study is something IS researchers always should do in their measurement practices. Also, testing the linguistic properties of measurement items in relation to the concept under study seems a good measurement practice. In general, we recommend IS researchers to pay attention to and test the alignment of their measurement instruments to the research context because this will reduce the likelihood of systematic measurement error and contribute to the psychometric quality of measurement in our research field.

Fifth, following the widespread view in the academic literature that the SD has been put forward to measure the meaning of psychological constructs (Heise, 2010), the suggested framework in this paper predicates on a reflective measurement approach. This is not to say that semantic differentiation could not be of use when applying formative measurement designs (see Jarvis, MacKenzie, & Podsakoff, 2003). In such cases, linguistic tests of bipolarity (stage 2) and wording (stage 3) and tests of contextual contamination (stage 5) seem to apply rather directly. When considering the selection of a sample of bipolar scales (stage 1), the testing of dimensionality (stage 4), and the final application of the SD (stage 6), however, adapting the suggested actions for scale selection and scale validation is recommended (e.g., see Dickinger & Stangl, 2013; MacKenzie et al., 2011). Follow-up study could extend our work across formative measurement practices.

Sixth, whereas we conceptually back up our decision to include the attitude toward purchasing, intention toward purchasing, website satisfaction, and loyalty intentions as dependent variables in the application of the EMQ scale, the results suggest some caution because not all EMQ dimensions contributed significantly to the explanation of variance in each of these constructs. Falling outside the objective of this paper, we suggest future research to shed more light on the significance and magnitude of the influence of the single EMQ dimensions. Such enquiry could expand the scope from our methodological illustration to a study that taps more explicitly into the theoretical reasons underlying the influence of the EMQ dimensions on different behavioral variables.

Seventh and finally, in this paper, we use the procedure of calculating the mean of a set of bipolar scales to obtain a respondent's overall score on the underlying concept (i.e., for each EMQ dimension). Even though this is a common procedure described in classical test theory (CTT) (see Furr and Bacharach, 2008), more-advanced calculation procedures may offer further refinement. A possible direction for such refinement may be found in item response theory (IRT) (Embretson & Reise, 2000). One of the applications of IRT is the assessment of item difficulty; that is, the difficulty respondents have to understand the measurement item and thus correctly answer it. IRT suggests options to measure item difficulty and account for it in the calculation of item scores (see Furr and Bacharach, 2008; Raykov and Marcoulides, 2011). Such advanced calculation may benefit the measurement practices of the IS researcher applying the SD and, therefore, it seems an interesting avenue for future research.

5.3. Conclusion

Building on the relevance of the semantic differential (SD) for the IS field, we overview the basic principles of conducting IS research using the SD. Based on an analysis of the extent to which these principles are adhered to in IS research and of the consequences of not adhering to these principles, we develop a set of procedural guidelines for developing and applying SD in IS research. By showing how these guidelines were applied in the practical example of studying electronic marketplace quality, we provide a concrete insight into what activities are required to meet each of the guidelines, and how these activities contribute to validity and reliability of SD-based research in our field. Thus, our paper serves to make the IS field aware of the principles of SD research and their relevance, and to provide the field with a framework that can enhance the quality of IS research using this measurement technique.

References

Allport, C. D., & Kerler, W. A. (2003). A research note regarding the development of the consensus on appropriation scale. *Information Systems Research*, *14*(4), 356-359.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bagozzi, R. P. (2007). On the meaning of formative measurement and how it differs from reflective measurement: Comment on Howell, Breivik, and Wilcox. *Psychological Methods*, *12*(2), 229-237.

Bagozzi, R. P., Gürhan-Canli, Z., & Priester, J. R. (2002). *The social psychology of consumer behavior*. Buckingham, UK: Open University Press.

Barrett, L. F., & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology*, *74*(4), 967-984.

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

Bearden, W. O., Hardesty, D. M., & Rose, R. L. (2001). Consumer self-confidence: Refinements in conceptualization and measurement. *Journal of Consumer Research*, *28*(1), 121-134.

Bearden, W. O., Netemeyer, R. G., & Haws, K.L. (2011). *Handbook of marketing scales: Multi-item measures for marketing and consumer behavior research*. Thousands Oaks, CA: Sage.

Benbasat, I., & Zmud, R. W. (2003). The identity crisis within the IS discipline: Defining and communicating the discipline's core properties. *MIS Quarterly*, *27*(2), 183-194.

Berthon, P., Pitt, L., Ewing, M., & Carr, L. (2002). Potential research space in MIS: A framework for envisioning and evaluating research replication, extension, and replication. *Information Systems Research*, *13*(4), 416-427.

Bhattacherjee, A., & Premkumar, G. (2004). Understanding changes in belief and attitude toward information technology Usage: A theoretical model and longitudinal test. *MIS Quarterly*, *28*(2), 229-254.

Bickart, B.A. (1993). Carryover and backfire effects in marketing research. *Journal of Marketing Research*, *30*(1), 52-62.

Blair, J., & Presser, S. (1992). An experimental comparison of alternative pretest techniques: A note on preliminary findings. *Journal of Advertising Research*, *32*(2), RC-2-RC-3.

Burton-Jones, A. (2009). Minimizing method bias through programmatic research. *MIS Quarterly*, *33*(3), 445-471.

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. New York: Routledge.

Cacioppo, J. T., & Berntson, G. G. (1994). Relationships between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin*, *115*(3), 401-423.

Cannell, C. F., Fowler, J. R., Kalton, G., Oksenberg, L., & Bischoping, K. (2004). New quantitative techniques for pretesting survey questions. In M. Bulmer (Ed.), *Questionnaires: Four-volume set* (pp. 187-201). London: Sage.

Carroll, J. B. (1959). Review of the measurement of meaning. *Language*, *35*, 58-77.

Carte, T. A., & Russell, C. J. (2003). In pursuit of moderation: Nine common errors and their solutions. *MIS Quarterly*, *27*(3), 479-501.

Cenfetelli, R. T. (2004). Inhibitors and enablers as dual factor concepts in technology usage. *Journal of the Association for Information Systems*, *5*(11-12), 472-492.

Cenfetelli, R. T., & Schwarz, A. (2011). Identifying and testing the inhibitors of technology usage intentions. *Information Systems Research*, *22*(4), 808-823.

Chin, W. W., Johnson, N., & Schwarts, A. (2008). A fast form approach to measuring technology acceptance and other constructs. *MIS Quarterly*, *32*(4), 687-704.

Christophersen, T., & Konradt, U. (2011). Reliability, validity, and sensitivity of a single-item measure of online store usability. *International Journal of Human-Computer Studies*, *69*(4), 269-280.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in scale development. *Psychological Assessment*, *7*(3), 309-319.

Cliff, N. (1959). Adverbs as multipliers. *Psychological Review*, *66*(1), 27-44.

Coaley, K. (2010). *An introduction to psychological assessment and psychometrics*. London: Sage.

Cogliser, C. C., & Schriesheim, C.A. (1994). Development and application of a new approach to testing the bipolarity of semantic differential items. *Educational and Psychological Measurement*, *54*(3), 594-605.

Cooil, B., Winer, R. S., & Rados, D. L. (1987). Cross-validation for prediction. *Journal of Marketing Research*, *24*(3), 271-279.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98-104.

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

Deese, J. (1964). The associative structure of some common English adjectives. *Journal of Verbal Learning and Verbal Behavior*, *3*(5), 347-357.

Devellis, R. F. (2012). *Scale development: Theory and applications.* Thousand Oaks, CA: Sage.

Dickinger, A., & Stangl, B. (2013). Website performance and behavioral consequences: A formative measurement approach. *Journal of Business Research*, *66*(6), 771-777.

Dickson, J., & Albaum, G. (1977). A method for developing tailormade semantic differentials for specific marketing content areas. *Journal of Marketing Research*, *14*(1), 87-91.

Dillman, D. A. (2007). *Mail and internet surveys: The tailored design method*. New York: John Wiley & Sons.

Doherty, N. F., Marples, C. G., & Suhaimi, A. (1999). The relative success of alternative approaches to strategic information systems planning: An empirical analysis*. Journal of Strategic Information Systems*, *8*(3), 263-283.

Doll, W. J., Xia, W., & Torkzadeh, G. (1994). A confirmatory factor analysis of the end-user computing satisfaction instrument. *MIS Quarterly*, *18*(4), 453-461.

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

Eggins, S. (2004). An introduction to systematic functional linguistics. London: Bloomsbury Academic.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum.

Evermann, J., & Tate, M. (2009). Building theory from quantitative studies, or, how to fit SEM models. *Proceedings of the International Conference on Information Systems*.

Falthzik, A. M., & Johnson, M. A. (1974). Statement polarity in attitude studies. *Journal of Marketing Research, 11*(1), 102-105.

Foddy, W. (2004a). An empirical evaluation of in-depth probes used to pretest survey questions. In M. Bulmer (Ed.), *Questionnaires: Four-volume set* (pp. 217-242). London: Sage.

Foddy, W. (2004b). Checks to ensure that questions work as intended. In M. Bulmer (Ed.), *Questionnaires: Four-volume set* (pp. 119-127). London: Sage.

Fowler, F. J., Jr. (2009). *Survey research methods.* Thousand Oaks, CA: Sage.

Friedmann, R., & Zimmer, M. R. (1988). The role of psychological meaning in advertising. *Journal of Advertising*, *17*(1), 31-40.

Funke, F., & Reips, U.-D. (2012). Why semantic differentials in web-based research should be made from visual analogue scales and not from 5-point scales. *Field methods*, *24*(3), 310-327.

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics: An introduction*. Thousand Oaks, CA: Sage.

Gerbing, D. W., & Anderson, J. C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research*, *25*(2), 186-192.

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling*, *3*(1), 62-72.

Green, R. F., Goldman, S. L., & Salovey, P. (1993). Measurement error masks bipolarity in affect ratings. *Journal of Personality and Social Psychology*, *64*(6), 1029-1041.

Griskevicius, V., Shiota, M. N., & Neufeld, S. L. (2010). Influence of different positive emotions on persuasion processing: A functional evolutionary approach. *Emotion*, *10*(2), 190-206.

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology.* Hoboken, NJ: John Wiley & Sons.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W.C. (1998). *Multivariate data analysis*. Upper Saddle River, NJ: Prentice-Hall.

Hambleton, R.K., & Rogers, J.H. (1991). Advances in criterion-referenced measurement. In R.K. Hambleton & J.N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications* (pp. 3-43). Dordrecht: Kluwer Academic Publishers.

Handling, A. H. (1994). Response and order effects in referendum voting: Exploring the influence of contextual bias on public policy. *Journal of Business Research*, *30*(1), 95-109.

Hardesty, D. M., & Bearden, W. O. (2004). The use of expert judges in scale development: Implications for improving face validity of measures of unobservable constructs. *Journal of Business Research*, *57*(2), 98-107.

Hardin, A. M., Chang, J. C-.J., & Fuller, M. A. (2008). Formative vs. reflective measurement: Comment on Marakas, Johnson, and Clay. *Journal of the Association for Information Systems*, *9*(9), 519-534.

Hawkins, D. I., Albaum, G., & Best, R. (1974). Stapel scale or semantic differential in marketing research? *Journal of Marketing Research*, *11*(3), 318-322.

Heise, D. R. (2010). *Surveying cultures: Discovering shared conceptions and sentiments*. Hoboken, NJ: John Wiley & Sons.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1-55.

Hong, S.-J., & Tam, K. Y. (2006). Understanding the adoption of multipurpose information appliances: The case of mobile data services. *Information Systems Research*, *17*(2), 162-179.

Hong, S.-J., Thong, J. Y. L., & Tam, K. Y. (2006). Understanding continued information technology usage behavior: A comparison of three models in the context of mobile internet. *Decision Support Systems*, *42*(3), 1819-1834.

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods*, *12*(2), 205-218.

Huang, M.-H. (2005). Web performance scale. *Information & Management*, *42*(6), 841-852.

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

Jayanti, R. K., & Burns, A. C. (1998). The antecedents of preventive health care behavior: An empirical study. *Journal of the Academic Marketing Science*, *26*(1), 6-15.

Kahn, R. L., & Cannell, C. F. (2004). The formulation of questions. In M. Bulmer (Ed.), *Questionnaires: Four-volume set* (pp. 55-78). London: Sage.

Kelly, R. F., & Stephenson, R. (1967). The semantic differential: An information source for designing retail patronage appeals. *Journal of Marketing*, *31*(4), 43-47.

Kerlinger, F. N. (1973). *Foundations of behavorial research*. New York: Holt, Rinehart and Winston.

Krosnick, J. A. (1999). Maximizing questionnaire quality. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of political attitudes* (pp.37–57). San Diego, CA: Academic Press.

Landon, E. L., Jr. (1971). Order bias, the ideal rating, and the semantic differential. *Journal of Marketing Research*, *8*(3), 375-378.

Laros, F. J. M., & Steenkamp, J.-B. E. M. (2005). Emotions in consumer behavior: A hierarchical approach. *Journal of Business Research*, *58*(10), 1437-1445.

Lasorsa, D. L. (2003). Question-order effects in surveys: The case of political interest, news attention, and knowledge. *Journalism and Mass Communication Quarterly*, *80*(3), 499-512.

Legendre, P. (2005). Species associations: The Kendall coefficient of concordance revisited. *Journal of Agricultural, Biological, and Environmental Statistics*, *10*(2), 226-245.

Legendre, P. (2010). Coefficient of concordance. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 164-169). Los Angeles: Sage Publications.

Lewis, B. R., Templeton, G. F., & Byrd, T. A. (2005). A methodology for construct development in MIS research. *European Journal of Information Systems*, *14*(4), 388-400.

Lin, Z., Janamanchi, B., & Huang, W. (2006). Reputation distribution and consumer-to-consumer online auction market structure: An exploratory study. *Decision Support Systems*, *41*(2), 435-448.

Lomax, R. G. (2001). *An introduction to statistical concepts for education and behavioral science*s. Mahwah, NJ: Lawrence Erlbaum Associates.

Luo, M. M., Chea, S., & Chen, J.-S. (2011). Web-based information service adoption: A comparison of the motivational model and the uses and gratifications theory. *Decision Support Systems, 51*(1), 21-30.

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

Malhotra, N. K., Agarwal, J., & Peterson, M. (1996). Methodological issues in cross-cultural marketing research: A state-of-the-art review. *International Marketing Review*, *13*(5), 7-43.

Marakas, G. M., Johnson, R. D., & Clay, P. F. (2008). Formative vs. reflective measurement: A reply to Hardin, Chang, and Fuller. *Journal of the Association for Information Systems*, *9*(9), 535-554.

McBride, S. D. & Wolf, B. (2007). Using multivariate statistical analysis to measure ovine temperament: Stability of factor construction over time and between groups of animals. *Applied Animal Behaviour Science*, *103*(1-2), 45-58.

McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews*: *Computational Statistics*, *1*(1), 93-100.

Menezes, D., & Elbert, N. F. (1979). Alternative semantic scaling formats for measuring store image: An evaluation. *Journal of Marketing Research*, *16*(1), 80-87.

Meng, X.-L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*(1), 172-175.

Mindak, W. A. (1961). Fitting the semantic differential to the marketing problem. *Journal of Marketing*, *25*(4), 28-33.

Murphy, K. R. (2003). *Validity generalization: A critical review*. Mahwah, NJ: Lawrence Erlbaum Associates.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications.* Thousand Oaks, CA: Sage.

Neuberg, S. L., West, S. G., Judice, T. N., & Thompson, M. M. (1997). On dimensionality, discriminant validity, and the role of psychometric analyses in personality theory and measurement: Reply to Kruglanski et al.'s (1997) defense of the need for closure scale. *Journal of Personality and Social Psychology*, *73*(5), 1017-1029.

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

Nye, C. D., Roberts, B. W., Saucier, G., & Zhou, X. (2008). Testing the measurement equivalence of personality adjective items across cultures. *Journal of Research in Personality, 42*(6), 1524-1536.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity, 41*(5), 673-690.

O'Reilly, P., & Finnegan, P. (2010). Intermediaries in inter-organisational networks: building a theory of electronic marketplace performance. *European Journal of Information Systems*, *19*, 462-480.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wright (Eds.), *Measures of personality and social psychological attitudes* (Vol. 1, pp. 17-51). San Diego, CA: Academic Press.

Ping, R. A., Jr. (2004). On assuring valid measures for theoretical models using survey data. *Journal of Business Research*, *57*(2), 125-141.

Pinker, E. J., Seidmann, A., & Vakrat, Y. (2003). Managing online auctions: Current business and research issues. *Management Science*, *49*(11), 1457-1484.

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

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York: Routledge.

Reynolds, N., Diamantopoulos, A., & Schlegelmilch, B. (2004). Pretesting in questionnaire design: A review of the literature and suggestions for further research. In M. Bulmer (Ed.), *Questionnaires: Four-volume set* (pp. 203-216). London: Sage.

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

Rigdon, E. E., Ringle, C. M., & Sarstedt, M. (2010). Structural modeling of heterogeneous data with partial least squares. *Review of Marketing Research*, *7*, 255-296.

Ringle, C. M., Wende, S., & Will, A. (2005). *SmartPLS 2.0 (M3) beta*. Retrieved from http://www.smartpls.de

Sarkar, M. B., Butler, B., & Steinfield, C. (1995). Intermediaries and cybermediaries: A continuing role for mediating players in the electronic marketplace. *Journal of Computer-Mediated Communication*, *1*(3), 1-14.

Sarstedt, M., Hensley, J., & Ringle, C. M. (2011). Multigroup analysis in partial least squares (PLS) path modeling: Alternative methods and empirical results. In M. Sarstedt, M. Schwaiger, & C.R. Taylor (Eds.) *Advances in international marketing volume 22: Measurement and research methods in international marketing* (pp. 195-218). Howard House (Bingley): Emerald Group Publishing.

Schriesheim, C. A. (1981). Leniency effects on convergent and discriminant validity for grouped questionnaire items: A further investigation. *Educational and Psychological Measurement*, *41*, 1093-1099.

Schriesheim, C. A., & Klich, N. R. (1991). Fiedler's least preferred coworker (LPC) instrument: An investigation of its true bipolarity. *Educational and Psychological Measurement*, *51*(2), 305-315.

Schriesheim, C. A., Solomon, E., & Kopelman, R.E. (1989). Grouped versus randomized format: An investigation of scale convergent and discriminant validity using LISREL confirmatory factor analysis. *Applied Psychological Measurement*, *13*(1), 19-32.

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Thousand Oaks, CA: Sage.

Sekaran, U. (1983). Methodological and theoretical issues and advancements in cross-cultural research. *Journal of International Business Studies*, *14*(2), 61-73.

Sharpe, L. K., & Anderson, W.T. (1972). Concept-scale interaction in the semantic differential. *Journal of Marketing Research*, *9*(4), 432-434.

Siegel, S., & Castellan, N. J. Jr. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Sirdeshmukh, D., Singh, J., & Sabol, B. (2002). Consumer trust, value, and loyalty in relational exchanges. *Journal of Marketing*, *66*(1), 15-37.

Snider, J. G., & Osgood, C. E. (Eds.). (1969). *Semantic differential technique: A sourcebook*. Chicago: Aldine Publishing Company.

Srinivasan, V., Vanden Abeele, P., & Butaye, I. (1989). The factor structure of multidimensional response to marketing stimuli: A comparison of two approaches. *Marketing Science*, *8*(1), 78-88.

Standing, S., Standing, C., & Love, P. E. D. (2010). A review of research on e-marketplaces 1997-2008. *Decision Support Systems*, *49*(1), 41-51.

Straub, D. W. (1989). Validating instruments in MIS research. *MIS Quarterly*, *13*(2), 147-169.

Straub, D. W., Boudreau, M.-C., and Gefen, D. (2004). Validation guidelines for IS positivist research. *Communications of the Association for Information Systems*, *13*(1), 380-427.

Straub, D. W., Hoffman, D. L., Weber, B. W., & Steinfield, C. (2002). Toward new metrics for net-enhanced organizations. *Information Systems Research*, *13*(3), 227-238.

Szymanski, D. M., & Hise, R. T. (2000). E-satisfaction: An initial examination. *Journal of Retailing*, *76*(3), 309-322.

Torkzadeh, G., & Dhillon, G. (2002). Measuring factors that influence the success of Internet commerce. *Information Systems Research*, *13*(2), 187-204.

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

Treiblmaier, H., & Filzmoser, P. (2010). Exploratory factor analysis revisited: How robust methods support the detection of hidden multivariate data structures in IS research. *Information & Management*, *47*(4), 197-207.

Treiblmaier, H., & Filzmoser, P. (2011). *Benefits from using continuous rating scales in online survey research.* Paper presented at the International Conference on Information Systems, Shanghai, China.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131.

Van Auken, S., & Barry, T. E. (1995). An assessment of the trait validity of cognitive age measures. *Journal of Consumer Psychology*, *4*(2), 107-132.

Van der Heijden, H., Verhagen, T., & Creemers, M. (2003). Understanding online purchase intentions: Contributions from technology and trust perspectives. *European Journal of Information Systems*, *12*, 41-48.

Verhagen, T., Meents, S., & Tan, Y-.H. (2006). Perceived risk and trust associated with purchasing at electronic marketplaces. *European Journal of Information Systems*, *15*, 542-555.

Verhagen, T., & Van Dolen, W. (2009). Online purchase intentions: A multi-channel store image perspective. *Information & Management*, *46*(2), 77-82.

Weinstein, Y., & Roediger, H. L. (2010). Retrospective bias in test performance: Providing easy items at the beginning of a test makes students believe they did better on it. *Memory & Cognition*, *38*(3), 366-376.

Weinstein, Y., & Roediger, H. L. (2012). The effect of question order on evaluations of test performance: how does the bias evolve? *Memory & Cognition*, *40*(5), 727-735.

Weir Kay, M. (Ed.). (1994). *Webster's collegiate thesaurus*. Springfield, MA: Merriam-Webster.

Wiggins, N., & Fishbein, M. (1969). Dimensions of semantic space: A problem of individual differences. In J.G. Snider & C.E. Osgood (eds.). *Semantic differential technique: A sourcebook* (pp. 183-193). Chicago: Aldine Publishing Company.

Wirtz, J., & Lee, M. C. (2003). An examination of the quality and context-specific applicability of commonly used customer satisfaction measures. *Journal of Service Research*, *5*(4), 345-355.

Wolfinbarger, M., & Gilly, M. C. (2003). eTailQ: Dimensionalizing, measuring and predicting etail quality. *Journal of Retailing*, *79*(3), 183-198.

Xiong, M. J., Franks, J. J., & Logan, G. D. (2003). Repetition priming mediated by task similarity in semantic classification. *Memory & Cognition*, *31*(7), 1009-1020.

Xiong, M. J., Logan, G. D., & Franks, J. J. (2006). Testing the semantic differential as a model of task processes with the implicit association test. *Memory & Cognition*, *34*(7), 1452-1463.

Xue, Y., Liang, H., & Wu, L. (2011). Punishment, justice, and compliance in mandatory IT settings. *Information Systems Research*, *22*(2), 400-414.

Yang, Z., Cai, S., Zhou, Z., & Zhou, N. (2005). Development and validation of an instrument to measure user perceived service quality of information presenting Web portals. *Information & Management*, *42*(4), 575-589.

Yi, M. Y., & Davis, F.D. (2003). Developing and validating an observational learning model of computer software training and skill acquisition. *Information Systems Research*, *14*(2), 146-169.

Yorke, M. (2001). Bipolarity…or not? Some conceptual problems relating to bipolar rating scales. *British Educational Research Journal*, *27*(2), 171-186.

Appendices

Appendix A: Factor loadings and Inter-Item Correlations

Table A-1. Item Overview and Factor Loadings Across the Three Datasets

Dimension	**Item**	**Item wording**	**Factor loadings: first version (n = 66)**	**Factor loadings: dimension-reversed (n = 64)**	**Factor loadings: item-reversed (n = 62)**
Layout	Lay1	Unattractive website layout – attractive website layout	0.83	0.89	0.89
	Lay2	Outdated website layout – up to date website layout	0.88	0.91	0.82
	Lay3	Boring website layout – interesting website layout	0.89	0.84	0.72
Ease of use	Ease1	Difficult to navigate website – easy to navigate website	0.90	0.89	0.85
	Ease2	Unclear website structure - clear website structure	0.88	0.92	0.87
	Ease3*	Difficult to search on the website – easy to search on the website	0.70	0.66	0.69
	Ease4	Difficult to learn how to use the website - easy to learn how to use the website	0.95	0.97	0.90
Contacting the intermediary	Contmed1	Insufficient information to contact <name intermediary> - sufficient information to contact <name intermediary>	0.87	0.95	0.91
	Contmed2	Difficult to contact <name intermediary> via the website – easy to contact <name intermediary> via the website	0.89	0.82	0.91
	Contmed3	Insufficient options to contact <name intermediary> - sufficient options to contact <name intermediary>	0.81	0.91	0.87
Institutional control	Instit1*	Insufficient guarantees – sufficient guarantees	0.56	0.66	0.49
	Instit2	Unclear information about guarantees – clear information about guarantees	0.72	0.79	0.63
	Instit3	Insufficient information about the privacy policy – sufficient information about the privacy policy	0.65	0.74	0.74
	Instit4	Insufficient privacy protection - sufficient privacy protection	0.73	0.78	0.73
	Instit5	Unclear information about the rules on <name EM> – clear information about the rules on <name EM>	0.78	0.73	0.70
	Instit6*	Insufficient rules that protect me on <name EM> – sufficient rules that protect me on <name EM>	0.56	0.64	0.62
	Instit7*	Weak website security – strong website security	0.70	0.76	0.65
	Instit8*	Insufficient monitoring of sellers - sufficient monitoring of sellers	0.82	0.84	0.60

Table A-1. Item Overview and Factor Loadings Across the Three Datasets (cont.)

Dimension	Item	Item wording	Factor loadings: first version (n = 66)	Factor loadings: dimension-reversed (n = 64)	Factor loadings: item-reversed (n = 62)
	Instit9*	Passive in removing swindlers – active in removing swindlers	0.66	0.64	0.71
Community	Commu1*	Difficult to contact other buyers – easy to contact other buyers	0.67	0.82	0.74
	Commu2	Difficult to share experiences with other buyers – easy to share experiences with other buyers	0.87	0.90	0.78
	Commu3	Few buyers sharing their experiences on <name EM> - many buyers sharing their experiences on <name EM>	0.86	0.83	0.70
	Commu4	Insufficient options to communicate with other buyers – sufficient options to communicate with other buyers	0.83	0.85	0.64
	Commu5*	Weak common bond between buyers – strong common bond between buyers	0.70	0.61	0.67
Contacting sellers	Contsel1	Insufficient information to contact sellers – sufficient information to contact sellers	0.93	0.94	0.88
	Contsel2	Difficult to contact sellers via the website – easy to contact sellers via the website	0.88	0.84	0.90
	Contsel3	Insufficient options to contact sellers – sufficient options to contact sellers	0.75	0.80	0.84
Seller information	Infsel1	Insufficient information about sellers – sufficient information about sellers	0.97	0.94	0.96
	Infsel2	Unclear indication of sellers' reputation – clear indication of sellers' reputation	0.82	0.78	0.95
	Infsel3	Insufficient information about sellers' reputation - sufficient information about sellers' reputation	0.89	0.77	0.86
Product information	Prodinf1 *	Unclear descriptions of <name products> - clear descriptions of <name products>	0.72	0.68	0.75
	Prodinf2	Incorrect descriptions of <name products> – correct descriptions of <name products>	0.88	0.82	0.79
	Prodinf3	Bad representation of <name products> (images/photos) – good representation of <name products> (images/photos)	0.76	0.78	0.80
	Prodinf4*	Difficult to assess the quality of <name products> - easy to assess the quality of <name products>	0.80	0.79	0.77
	Prodinf5*	Insufficient product photos of <name products> – sufficient product photos of <name products>	0.78	0.86	0.81
	Prodinf6*	Unclear whether <name products> are used - clear whether <name products> are used	0.73	0.83	0.76
	Prodinf7	Unclear condition of <name products> – clear condition of <name products>	0.74	0.84	0.87

X

Table A-1. Item Overview and Factor Loadings Across the Three Datasets (cont.)

Dimension	Item	Item wording	Factor loadings: first version (n = 66)	Factor loadings: dimension-reversed (n = 64)	Factor loadings: item-reversed (n = 62)
Pricing mechanisms	Pricing1	Unclear how final prices are effected – clear how final prices are effected	0.96	0.87	0.81
	Pricing2	Inconvenient pricing method – convenient pricing method	0.96	0.89	0.78
	Pricing3	Unclear what final price to pay – clear what final price to pay	0.74	0.63	0.85
Assortment	Assor1	Few interesting <name products> – many interesting <name products>	0.94	0.84	0.89
	Assor2	Limited range of <name products> – wide range of <name products>	0.96	0.87	0.96
	Assor3	Insufficient number of <name products> - sufficient number of <name products>	0.92	0.90	0.84
Settlement	Settl1	Unclear how to pay for <name products> – clear how to pay for <name products>	0.92	0.87	0.75
	Settl2	Difficult to pay for <name products> - easy to pay for <name products>	0.87	0.95	0.78
	Settl3	Unclear how to receive <name products> – clear how to receive <name products>	0.84	0.80	0.70
	Settl4*	Difficult to receive <name products> – easy to receive <name products>	0.78	0.62	0.75
Meeting sellers	Meet1	Difficult to meet sellers and evaluate <name products> before you buy - easy to meet sellers and evaluate <name products> before you buy	0.98	0.91	0.89
	Meet2	Difficult to meet sellers and pay them - easy to meet sellers and pay them	0.88	0.88	0.80
	Meet3	Difficult to pick up <name products> at the sellers' location - easy to pick up <name products> at the sellers' location	0.63	0.69	0.73

Table A-2. Inter-Item Correlations Across the Three Datasets

	First version (n = 66)			Dimension-reversed (n = 64)			Item-reversed (n = 62)		
Items	Lay1	Lay2	Lay3	**Items**	Lay1	Lay2	Lay3	**Items**	Lay1
Lay1	-			Lay1	-			Lay1	-
Lay2	0.59	-		Lay2	0.59	-		Lay2	0.59
Lay3	0.60	0.70	-	Lay3	0.60	0.70	-	Lay3	0.60
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
Items	Meet1	Meet2	Meet3	**Items**	Meet1	Meet2	Meet3	**Items**	Meet1
Meet1	-			Meet1	-			Meet1	-
Meet2	0.81	-		Meet2	0.81	-		Meet2	0.81
Meet3	0.62	0.71	-	Meet3	0.62	0.71	-	Meet3	0.62

Appendix B: CFA Alternative Model Testing Sample 1 and 2

To further investigate the dimensionality of the 37-item EMQ instrument, we tested two alternative models. Drawing on Doll et al. (1994), we tested a model consisting of twelve uncorrelated first-order factors and an one-factor model relating all single items to one first-order EMQ factor. We tested the twelve uncorrelated first-order factor model to evaluate the likelihood that the twelve dimensions functioned as unrelated dimensions. If demonstrated, this would imply that the twelve dimensions should be treated separately rather than as dimensions of the same underlying concept; that is, EMQ. We tested the one-factor model to assess the likelihood that the 37-item EMQ instrument reflected one dimension instead of the proposed twelve dimensions. If demonstrated, this would refute the notion that the EMQ concept is multidimensional in nature. The results (Table B-1) indicated unacceptable fit for the two alternative models. Therefore, we concluded that the correlated twelve first-order factors model is most applicable to model EMQ.

Table B-1. Shortcomings in Semantic Differentiation and Measurement Consequences

Model	χ2	*Df*	*χ2/df*	GFI	AGFI	RMR	RMSEA	NFI	TLI	CFI
Sample 1 (n = 928)										
Twelve first-order factors (correlated)	1226.605 (p < .001)	562	2.183	.93	.92	.059	.036	.96	.97	.98
Twelve first-order factors (uncorr.)	4418.899 (p < .001)	629	7.025	.68	.64	.485	.081	.85	.86	.87
One first-order factor	18256.962 (p < .001)	629	29.025	.47	.40	.228	.174	.36	.33	.37
Sample 2 (n = 551)										
Twelve first-order factors (correlated)	1154.230 (p < .001)	562	2.054	.90	.87	.087	.044	.94	.96	.97
Twelve first-order factors (uncorr.)	3058.767 (p < .001)	629	4.863	.67	.63	.487	.084	.83	.85	.86
One first-order factor	12004.672 (p < .001)	629	19.085	.42	.35	.258	.181	.32	.29	.33

Appendix C: Reliability and Validity Testing Sample 1 and 2

The reliability statistics indicated good reliability for the twelve EMQ dimensions. Except for the dimension product information (α = 0.77, sample 2), all Cronbach's alphas surpassed the 0.80 level. Since none of the Cronbach's alphas surpassed the 0.95 level, and given that we used relatively few bipolar scales to measure each dimension, we obtained no indications for item redundancy (also see Bearden et al., 2011). For all dimensions, the average variance extracted (AVE) exceeded the 0.50 thresholds prescribed in the literature (e.g., Ping, 2004). We then tested for convergent, discriminant, and predictive validity. We assessed convergent validity by AVE's, Cronbach's alphas, and minimum item-to-total correlations. All AVE's were above the recommended level of 0.50 (Yi & Davis, 2003). The minimum item-to-total correlations revealed high correlations, all exceeding the criterion of 0.40 (see Jayanti & Burns, 1998), and thereby providing strong additional support for convergent validity.

To test for discriminant validity, we studied the within-construct item correlations for each of the twelve EMQ dimensions and compared these loadings with cross-loadings on items of other dimensions. All within-construct item loadings were higher than their cross-loadings, and we observed no cross-loadings above 0.70, which suggests discriminant validity (Ping, 2004). To further assess discriminant validity, we measured the differences between the squared correlations between dimensions and their individual AVE. Since the value of squared correlations was less than either of their individual AVE's for all pairs of dimensions we tested for, discriminant validity was confirmed (Yi & Davis, 2003). Finally, we assessed the predictive validity of the EMQ scale. For both samples, the attitude toward purchasing was regressed on the EMQ dimensions (scales derived from Van der Heijden, Verhagen, & Creemers, 2003). Table C-1 reports the results (All VIF scores < 10).

Table C-1. Standardized Regression Coefficients of EMQ Dimensions on Attitude (Multiple regression)

	Sample 1 (n = 928)	Sample 2 (n = 551)
Dimension	**Attitude**	**Attitude**
Layout	-.01	.00
Ease of use	-.03	.02
Contacting the intermediary	-.00	-.06
Institutional control	08	-.02
Community	-.06	.03
Contacting sellers	-.07	-.12
Seller information	.02	.07
Product information	**.28 ***	.11
Pricing mechanisms	**.11 ****	.01
Assortment	**.16***	**.24***
Settlement	.03	.08
Meeting sellers	**.16***	**.25***
R^2	.27	.26
Adjusted R^2	.26	.24

* significant at p < .001, ** significant at p < .01.

For both samples, the EMQ scale explained around 25 percent of the variance of the attitude toward purchasing. Given the level of target specificity of the attitude construct, these findings were encouraging. The number of dimensions that contributed significantly to this variance, however, seemed slightly below our expectations. Four dimensions for sample 1 and two dimensions for sample 2 had significant influences on the attitude. Given the large number of dimensions that constitute the EMQ scale. it seemed reasonable to explore whether the analyses had been affected by multi-collinearity. Even though the VIF scores demonstrated that the level of multi-collinearity was rather low, there could have been other factors such as the sample size, the proportion of the variance in the dependent variable associated with the independent variables, and the variance of the independent variables that boosted the influence of multi-collinearity and resulted in a deflation of the variance of the regression coefficients (O'Brien, 2007, p. 675). We therefore decided to run ridge regression (SPSS statistics, version 20), a technique that uses biased estimation to remove the effects of multi-collinearity and arrive at a more pure estimation of the regression coefficients (for a detailed discussion, see McDonald, 2009). Table C-2 shows the results.

Table C-2. Standardized Regression Coefficients of EMQ Dimensions on Attitude (Ridge regression)

	Sample 1 (n = 928)	**Sample 2 (n = 551)**
Dimension	**Attitude**	**Attitude**
Layout	.02	-.01
Ease of use	.03	.01
Contacting the intermediary	.03	-.03
Institutional control	**.05 ***	.02
Community	.01	.05
Contacting sellers	-.01	-.03
Seller information	.03	.02
Product information	**.12 ***	**.08 ***
Pricing mechanisms	**.08 ***	.02
Assortment	**.13 ***	**.18 ***
Settlement	.04	**.07 ***
Meeting sellers	**.12 ***	**.18 ***
R^2	.30	.36
Adjusted R^2	.28	.33

* significant at p < .001.

The results of the ridge regression show that the EMQ scale explained 28 percent (sample 1) and 33 percent (sample 2) of the variance of the attitude. Five dimensions contributed significantly to the attitude for Sample 1; four dimensions contributed significantly to the attitude for sample 2. Overall, these results reconfirm the predictive validity of the EMQ scale as a whole and imply that multiple dimensions may function as significant direct predictor of the attitude.

Appendix D: Reliability and Validity Testing Sample 3 and 4

We used Cronbach's alpha to re-assess the reliability and validity of the SD. All Cronbach's alphas fell within the range of 0.86-0.95, which suggests a high reliability. Because none of the Cronbach's alphas surpassed the 0.95 level and because we used relatively few bipolar scales per dimension, it seemed safe to state that item redundancy was unlikely to be an issue (see Netemeyer et al., 2003). All AVE's were above 61 and all minimum item-to-total correlations were above the value of .705, which reconfirmed the reliability and strongly confirming the convergent validity of the twelve EMQ dimensions. Applying the same procedures as for sample 1 and 2, discriminant validity was also strongly confirmed.

Next, we assessed the predictive validity of the EMQ scale by regressing the online purchase attitude, online purchase intention (measure: Van der Heijden et al., 2003), website satisfaction (measure: Szymanski & Hise, 2000), and loyalty intention (measure: Sirdeshmukh, Singh, & Sabol, 2002) on the twelve EMQ dimensions. The results are reported below (all VIF's < 10). The amount of explained variance and the significance of the coefficients for both samples supported the predictive validity of the EMQ scale as a whole; of the twelve EMQ dimensions, eight dimensions directly contributed to the variance of at least one of the four dependents across the two datasets (see Table D-1).

Table D-1. Standardized Regression Coefficients of EMQ Dimensions on Attitude, Intention, E-Satisfaction, and E-Loyalty (Multiple Regression).

	Sample 3 (*n* = 863)				**Sample 4 (*n* = 590)**			
Dimension	***Attitude***	***Intention***	***e-Satis.***	***e-Loy.***	***Attitude***	***Intention***	***e-Satis.***	***e-Loy.***
Layout	.04	.06	**.13****	.08	.04	-.04	**.18***	.09
Ease of use	-.00	-.01	**.23***	.04	**.14****	.11	**.25***	.04
Contacting the intermediary	-.00	.03	.02	.00	.07	.06	.03	.07
Institutional control	.05	.10	**.17***	.10	-.03	-.07	-.01	.02
Community	-.03	-.00	.07	-.00	-.09	-.07	.02	-.01
Contacting sellers	.01	-.02	.06	.03	.11	.09	**.12****	.09
Seller information	.01	-.02	-.09	-.02	-.00	-.01	.08	.03
Product information	.04	.02	.05	.02	**.12****	.07	-.03	.06
Pricing mechanisms	.09	.02	-.08	.05	-.03	.03	.04	.02
Assortment	**.20***	**.22***	.09	**.26***	**.14****	**.18***	**.15***	**.31***
Settlement	**.14****	.09	.07	**.13****	**.26***	**.29***	.06	.08
Meeting sellers	**.19***	**.17***	**.10****	**.13****	-.00	-.04	.02	.02
R squared	*.33*	*.26*	*.40*	*.37*	*.30*	*.24*	*.40*	*.35*
Adjusted R squared	*.32*	*.25*	*.39*	*.36*	*.28*	*.22*	*.39*	*.33*

* significant at p < .001, ** significant at p < .01.

Finally, following the rational and procedures applied in the analysis of the data of samples 1 and 2, we decided to rerun the analyses with ridge regression (SPSS statistics, version 20) in order to rule out possible disturbing effects of multi-collinearity. Table D-2 shows the results.

Table D-2. Standardized Regression Coefficients of EMQ Dimensions on Attitude, Intention, E-Satisfaction and E-Loyalty (Ridge Regression)

	Sample 3 (n = 863)				Sample 4 (n = 590)			
Dimension	Attitude	Intention	e-Satis.	e-Loy.	Attitude	Intention	e-Satis.	e-Loy.
Layout	.04	.04	.10 *	.06 *	.05	-.01	.12 *	.05 *
Ease of use	.04	.02	.11 *	.05 *	.09 *	.05 *	.12 *	.04
Contacting the intermediary	.04	.05 *	.07 *	.05 **	.07 *	.06 **	.06	.05
Institutional control	.04	.05 *	.07 *	.07 *	-.02	-.01	.05	.07
Community	-.01	.03	.07 *	-.01	-.02	-.04	.05	-.01
Contacting sellers	.05	.03	.05 *	.04	.09 *	.06 *	.09 *	.08 *
Seller information	.03	.00	-.01	.03	.01	-.01	.04	.05
Product information	.07 *	.05	.05 *	.07 *	.08 *	.07 *	.03	.11 *
Pricing mechanisms	.07 **	.04	.01	.06 *	.00	.05	.06	.04
Assortment	.11 *	.11 *	.08 *	.12 *	.15 *	.15 *	.09 *	.16 *
Settlement	.10 *	.07 *	.06 *	.08 *	.13 *	.12 *	.06 *	.06 **
Meeting sellers	.11 *	.09 *	.06 *	.09 *	-.03	-.04	.05	.06
R squared	*.35*	*.25*	*.39*	*.38*	*.37*	*.28*	*.44*	*.42*
Adjusted R squared	*.33*	*.22*	*.37*	*.36*	*.34*	*.25*	*.41*	*.39*

* significant at p < .001, ** significant at p < .01.

The results strongly reconfirm the predictive validity of the EMQ scale as a whole and show that eleven out of twelve EMQ dimensions have significant direct influences on the behavioral variables included in the analyses.

About the Authors

Tibert VERHAGEN is Associate Professor of E-business at the VU University Amsterdam (Faculty of Economics and Business Administration) in the Netherlands. He has a PhD in online consumer behavior. His research interests include the influence of emerging forms of IT on online consumer decision-making; online marketing; and research methodology. His work has been published in journals such as *European Journal of Information Systems*, *Information & Management*, *Computers in Human Behavior*, *Journal of Computer-Mediated Communication*, *New Media & Society*, and *International Journal of Information Management*.

Bart van den HOOFF is Professor of Organizational Communication and Information Systems at the VU University Amsterdam (Faculty of Economics and Business Administration) in the Netherlands. He has a PhD (with honors) in Communication from the University of Amsterdam. Before coming to the VU University, he worked in consultancy (M&I/Partners), at the Delft University of Technology and the University of Amsterdam. His research interests include the interaction between ICT, organization and individuals; Enterprise Systems; online interaction and knowledge coordination. His work has been presented at international conferences and published in (among others) Jo*urnal of Management Studies*, *Organization Studies*, *Journal of Information Technology*, *Information & Management*, *Communication Research*, *European Journal of Information Systems,* and *Human Communication Research*.

Selmar MEENTS holds a position as lecturer at the International Business School of the Amsterdam University of Applied Sciences. He received his PhD in e-business from the VU University Amsterdam. Using his practical experience as a consultant and research manager, he gives courses focusing on the management of buyer-seller relationships and on the design and execution of applied business research. Among his research interests are topics such as e-business, emerging technologies, buyer-seller relationships, and research design. His papers have been published in several conference proceedings, and have appeared in journals such as *Computers in Human Behavior*, *European Journal of Information Systems*, and *Information & Management*.