VARIATIONS ON A RATING SCALE: THE EFFECT ON EXTREME RESPONSE TENDENCY IN PRODUCT RATINGS

Complete Research

Tsekouras, Dimitrios, Rotterdam School of Management, Erasmus University, Netherlands, dtsekouras@rsm.nl

Abstract

Product ratings have become an integral element of online businesses especially for experience goods, yet seem to be prone to biases that shift most of the distribution towards the extreme points of the scales. Response biases due to inherent traits (such as acquiescence or extreme response style) are widely investigated in survey design and marketing research, yet little is known about how the rating scale variations in user generated product evaluations influence their formation. More precisely, in an experimental study in the context of movie ratings, I show that the use of emotional labels attracts users to the endpoints of the rating scale but their responses are less susceptible to extreme response tendency when the size of the rating scale is increased. Also, simply priming the midpoint of the scale reduces extreme responses, though this effect is attenuated when emotional labels are used. Such effects remain consistent when I account for response styles, cultural dimensions and individual characteristics. The broad use of product ratings in generating personalized recommendations and predicting market performance necessitates a discussion on how to better account for potential distortions in these ratings due to variations in the rating scale.

Keywords: Word Of Mouth, Product Rating Scales, Response Styles, Response Bias.

1 Introduction

Companies traditionally reached consumers in order to gain insights on the performance and efficiency of their offerings. The emergence of online technologies has facilitated such exchange of information and, as a result, consumers have become increasingly active in voluntarily evaluating products online across a wide variety of industries (e.g. travel, electronics). In fact, there are many popular platforms that base their mere existence and business model on such user generated content (e.g. IMDB, Yelp). Such consumer ratings have become an integral element of online vendors and a considerable part of marketing budget is spent in maintaining and analysing such user-generated content (Chevalier and Mayzlin, 2006; Moe & Schweidel, 2011). Studies have showed that product ratings are a reflection of product quality (Hu et al., 2009) and a fairly good predictor of future sales (Duan et al., 2008a; Zhu & Zhang, 2010; Floyd et al., 2014). Further, product ratings are often used to further allocate marketing budget (when and where needed), to improve the products offered (e.g. hotel improvements after bad ratings), or even generate personalized recommendations for users (e.g. Amazon or Netflix recommendations). Also, consumers tend to trust more opinions derived from other customers of a product than information provided by the vendors themselves (Chevalier and Mayzlin, 2006; Dellarocas, 2003). The amount of (positive) product ratings increases the propensity of new consumers to contribute their own opinions about a product, potentially leading to an upward spiral of "positive buzz" which improves the product performance (Moe & Schweidel, 2011). In contrast, a negative turn in the product ratings can be detrimental for the market performance of a product.

User generated product evaluations can take the form of a rating (numerical evaluation) or a review (textual content) (Moe & Schweidel, 2011). Whereas reviews are more insightful, they require more effort to process. Therefore many customers use the numerical ratings as a proxy of the textual reviews. Despite the wide acceptance and popularity of consumer product ratings, it seems that there is no consensus on the format and the structure of the rating environment. There are many examples of websites that use various structures in their rating systems. CNET and iTunes ask their users to rate applications on a 5-points scale, YouTube uses a binary scale (like/dislike) and Winespectator uses a 100-points scale for the wine ratings. Amazon uses emotion related labels on the 5-star scale ("hated" versus "loved this product"). Although many differences stem from industry norms (as in the case of wines) or product characteristics, interestingly enough I observe within industry differences as well. In movie ratings, IMDB.com uses a 10-points scale whereas RottenTomatoes uses a 5-points scale, both with numerical labels. In contrast Netflix uses a 5-points fully labeled scale (from "Love" to "Hate"). Differences are also observed in restaurant ratings (e.g. Yelp uses 5-points scale, Iens.nl a 10-points scale and Zagat a 30-points scale). Websites have also experimented with different scales over the years. For example, Facebook used a 5-points hedonic scale ("really don't like it" to "Love it") before shifting to a 5-points scale with objective labels ("Very Poor" to "Excellent").

An interesting question that emerges from observing such variations in product rating scales is to what extent do these variations influence the way users respond in rating the given products. The relevance of such question becomes prevalent when considering systematic biases observed in online consumer product ratings. Online ratings are characterized by extreme responses, which form a j-shaped distribution due to purchasing and under-reporting biases (Hu et al., 2009). Consumers with more extreme ratings (very satisfied or dissatisfied) are more likely to rate and are therefore keener to inflate their ratings (Li and Hitt, 2008). Also, users are prone to social influence since their ratings most times are anchored by the existing ratings of others (Sridhar and Srinivasan, 2012). Such findings indicate that consumers' ratings reflect their true rating plus a systematic error (response bias). Numerous studies focused on response biases occurring in the context of questionnaire design and survey formats. Such biases might be related to inherent response styles (user traits) or response stimuli (rating scale) and can seriously distort the reliability and validity of the outcomes (Baumgartner and Steenkamp, 2001; Weijters et al., 2010). Social desirability, yeasaying or naysaying (tendency to always agree or disagree to the interviewer) and extreme response, are the most frequently identified response styles in questionnaire responding (Tellis and Chandrasekaran, 2010). Whereas response styles mostly depend on inherent traits (e.g. personality, culture, language) (De Jong et al., 2008; De Langhe et al., 2011), there is evidence that response styles can be influenced by response sets as well (e.g. size, existence of midpoint, order of answers) (Dawes, 2008; Greenleaf, 1992; Hui and Triandis, 1989; Weijters, 2006). However, little has been known regarding the effect of response sets in the context of online user generated product ratings (word-of-mouth). Product rating responses are expected to differ from survey responding in that: (a) product ratings are less prone to social desirability biases as they are user driven, (b) there is expected a lower level of acquiescence (tendency to agree to the experimenter) and (c) the lack of demand characteristics issue (participants respond in a particular way because they are observed) (Nichols and Maner, 2008; Tellis and Chandrasekaran, 2010).

In an online experiment, I investigate the effect of the variations in the rating scales in the context of experience products (movies). The context is very natural since consumers rely on those to base their purchase decisions (Godes and Mayzlin, 2004). More precisely, I focus on users' extreme response tendency and show that the emotionality of the rating scale labels influences their ratings in conjunction with the size of the scale and the explicit labeling of the scale midpoint. Such effects remain consistent when the cultural scores of individualism and uncertainty avoidance, as well as a general tendency to extreme response (captured by extreme response tendency in a series of unrelated questions) are taken into account. The study provides insights in the effects of rating scale formats on product ratings and explores the presence of certain biases in this context. As product ratings are often used to estimate future product performance and investments, a poorly applied rating scale can serious contamination of these figures.

2 Theoretical Background

2.1 User Generated Ratings

The technological advances in e-commerce have facilitated and popularized user generated product evaluations (Word-of-Mouth). Most retailing websites incorporate rating systems in order to encourage their users to share their post-consumption experiences (Chevalier and Mayzlin, 2006; Dellarocas, 2003). There is an increasing number of websites that serve as hosts of independent product ratings and build their business model around user generated content (e.g. IMDB, Yelp, Tripadvisor). Such content has become a fundamental component of online vendors in their process of identifying consumer attitudes and construct market estimates (Duan et al., 2008b; Moe & Schweidel, 2011) as well as for consumers in order to facilitate their purchase decisions (Chevalier and Mayzlin, 2006).

Consumers' product opinions can be distinguished between product reviews and product ratings. Product reviews consist of the textual evaluation of a product ratings are quantitative interpretation built on a scale with a limited amount of options (Moe & Schweidel, 2011). Though reviews are richer in content, they require effort to produce (and process), compared to the numerical ratings. As a result, the number of produced ratings is multiple compared to that of the reviews. Similarly, most consumers use the ratings as a fair proxy of product quality (Hu et al., 2009).

There is a growing stream of literature focusing on the antecedents and consequences of product ratings. Product ratings are prone to a purchasing an under-reporting bias that are reflected in a J-shaped distribution of ratings. Purchasing bias delves from the fact that those who bought a product are those who saw benefits in it and that is reflected in the extremely positive ratings. Under-reporting bias implies that those with moderate opinions are less likely to contribute their ratings (Hu et al., 2009). As a result, online ratings are sensitive to extreme response behaviour. Also, consumers are prone to social influences as the volume and valence of existing ratings affect users' decision of whether and how much to rate a product (Moe & Schweidel, 2011; Sridhar and Srinivasan, 2012). Finally, product ratings largely depend on the product category as well as on previous raters' and own expertise (Moe & Schweidel, 2011; Zhu and Zhang, 2010). Such findings indicate that consumers' ratings include a systematic error component next to the true rating score.

Product ratings are a reflection of product quality (Hu et al 2009) but also there is empirical evidence across various contexts regarding their effectiveness in influencing and predicting product sales (Clemons et al., 2006; Dellarocas et al., 2007; Duan et al., 2008a; Forman et al., 2008; Floyd et al., 2014). However, the negative ratings show more predictive power than the respective positive ones (Chevelier & Mayzlin, 2006; Mudambi and Schuff, 2010). Caution has been also raised to a self-selection bias, in that there is a systematic component in those who decide to contribute the early ratings, which in turn influence subsequent ratings (Li and Hitt, 2008). Product ratings are also used to make investment decisions (e.g. marketing budget), to improve products, or generate personalized recommendations for raters (e.g. Netflix or Amazon). Since companies and consumers form decisions on the basis of such ratings, it is very important to understand the extent to which ratings can be inflated or deflated from their real value. Product ratings seem to be prone to various response biases, however this study attempts to investigate how the ratings can systematically be attributed to the structure of the rating system. Such an approach has a rich background in response theories (Baumgartner and Steenkamp, 2001) but has been under-researched in the context of product ratings.

2.2 Response Sets and Styles

In various fields, such as marketing and sociology, scales are applied to measure individuals' attitude and opinion about a wide range of issues. Typically respondents are asked to make a choice out of an ordered set of available options (Friedman & Amoo, 1999). Companies traditionally used such scales when surveying consumers to gain insights on the performance and efficiency of their offerings. In fact, consumers respond to numerous surveys regarding their behavioral intentions and opinions about a variety of topics (Tellis and Chandrasekaran, 2010).

However, there is evidence that next to respondents' true answers a systematic and a random error component are entailed (De Jong et al. 2008). Whereas the random error is unobserved, the systematic error (response bias) reflects the responder's tendency to systematically answer questions in a certain pattern irrespective of what the questions intend to measure (Baumgartner and Steenkamp, 2001). Such a response bias may be a function of individual characteristics (response styles) or stimuli (response set) (Kieruj and Moors, 2010; Paulhus, 1991; Weijter et al., 2010). Whereas response styles are respondent specific, response sets assume that different question formats yield responses that are substantially different in irrespective of the content (Greenleaf, 1992; Hui and Triandis, 1989; Watkins and Cheung, 1995). The response biases distort the outcomes of a survey in terms of reliability and validity of the responses.

The main response styles in sociological and psychological literature are the social desirability, (dis-) acquiescence, extreme and middle response bias (Greenleaf, 1992; Paulhus, 1991; van Vaerenbergh & Thomas, 2013). Social desirability is the tendency to respond in an acceptable way (based on cultural norms) regardless the true answer. Relatedly respondents tend to shift their behavior due to the fact that they are being observed, hence distorting their true answers (Nichols and Maner, 2008). Acquiescence is the tendency to disproportionally agree rather than disagree (and vice versa for disacquiescence) with items, regardless of the question. A related response style is the net acquiescence, which is the pattern of showing greater acquiescence than disacquiescence (Baumgartner and Steenkamp, 2001). Extreme response bias is the tendency to systematically choose the end-points of a scale (either negative or positive) whereas middle response bias refers to an excessive choice of the middle point, regardless of the question's content. Finally, respondents might use a narrow range of response categories around the mean (Greenleaf, 1992). Response styles are mostly a function of respondents' demographic (e.g. age, gender, education) or cultural characteristics (Baumgartner and Steenkamp 2001; De Jong et al., 2008). Experimenters should take into account these potential distortions in their collected responses and correct them accordingly (latent-class confirmatory factor analysis, independent scale usage scores, reverse-coding) (Kieruj and Moors, 2013; Tellis and Chandrasekaran, 2010).

Whereas response styles vary at the respondent level, there is evidence that there are systematic response differences that are associated with scale-related stimuli. Such systematic errors influence the response behavior (measured in terms of the different aforementioned response styles). First, there is evidence of an order effect in that the direction of ranking (left or right) influences the distribution of responses (if negative items are presented first they received more votes and vice versa) (Hartley and Betts 2010). Relatedly, respondents show some inertia as their responses highly depend on previous answers they gave (De Jong et al. 2012). The scale format influences response behavior. Two stage questions show more extreme response tendency than one-stage questions (Albaum et al., 2007). Fully labeled scales reduce the extreme response whereas more extended response scales are decreasing extreme response yet increase acquiescence (Dawes, 2008; Weijters et al. 2010). Also, scales including a middle option (usually depicted by the odd or even number of categories) reduce extreme response

(Kalton et al., 1980). There are also presentational elements of a scale that influence responses (e.g. color, screen position, use of emoticons) (Tourangeau et. al., 2007; Tourangeau et. al., 2013). Finally, recent studies showed mixed evidence on the effect of scale language (native versus second) on extreme response (De Langhe et. al., 2011; Harzing, 2006).

Response biases are expected to be predominant also in the context of user generated product ratings. A clear indication of such biases is the observed J-shaped distribution of product ratings (inflating the extreme responses of the scale) (Hu et al., 2009). Therefore, this study focuses on extreme response bias, and more precisely, the distance of a rating from the midpoint of the scale. The main response set elements that are tested in this study and are expected to influence responses are the type of labels at the endpoints of a scale, the size of the scale and the midpoint of a scale. This is one of the first studies that delve into the scale's effect on extreme product rating in the context of word-of-mouth. Also, most studies regarding these scale elements (mostly in the context of survey responses) do not account for interactions that emerge from the variations in the scales.

2.3 Labels in a rating scale

A very important factor that influences how users respond in a given scale is the label format of the scale. The labels attached to various scale options are particularly vulnerable to biases that can violate the reliability and distribution of the responses (Myers & Warner, 1968). Numerically labeled scales (1-5 scales) lead to higher extreme response compared to verbal scales (disagree to agree scales) (O'Muircheartaigh et al., 1995). However, there are many variations in the operationalization of verbal scales and the wording of the categories may influence individuals' responses (Weijters et al., 2013; Wyatt and Meyers 1987). Studies also showed that fully verbally labeled scales reduce the extreme response compared to scales labeled only at the endpoints (Weijters et al., 2010). Users assign a certain psychological value to all options in a scale as a reference to the endpoint labels and therefore might neglect all intermediate points (Wildt and Mazis, 1978). Also, the direction of the labels influences the distribution of responses as users anchor themselves to the first labels that appear in the scale (Hartley and Betts, 2010).

The way consumers evaluate products can considerably fluctuate to the extent they tap into emotional concepts (Voss et al., 2003). Emotions reflect the users' experience of various feelings and are indicated as a crucial driver of behavior as well as of how users evaluate existing reviews (Yin et al., 2014). Respectively, many companies make use of emotional cues in an attempt to influence product perceptions (Ludwig et. al., 2013). The persuasive power of emotional words is expected to apply in product rating scales[1] as well; yet, evidence for the direction of such effects remains contradictory. An intensity related hypothesis posits that more intense verbal labels at the endpoints of a rating scale may lead respondents to move away from the ends of the scale (Ostrom, 1966; Upshaw, 1965; Weijters et al., 2013). Also, as emotional labels ignite the arousal of users, users become more activated in the rating process and widen the range of their responses. Finally, as non-emotional labels are often considered as less strong, they are expected to increase extreme response as true answers may be beyond these labels (Wyatt & Meyers 1987). However, the intensity of emotion-related words largely depends on the applied language of the scale (i.e. words in native language are perceived as more intense than in second language) (De Langhe et al., 2011).

The polar emotions found in emotional label of a rating scale may cause two distinctive memory sys-

[1] In this study, emotional labels in a rating scale refer to e.g. "hate" to "love" a product (see Amazon and Netflix) whereas non-emotional labels (informational) refer to e.g. "Very Poor" to "Excellent" products (see iTunes and Facebook).

tems being activated in the brain. Such activation lets users experience opposing emotions, which in turn drive behavior (Lau-Gesk & Meyers-Levy, 2009; Lench et al., 2011). However, this clash in emotions will increase the cognitive load of users, which results in the use of cognitive shortcuts in their decision making and as a result, more extreme responses (Tversky and Kahneman, 1974; Knowles and Condon, 1999). Additionally, in the context of experience products (such as the context of this study; movies), emotional labels are easier to assess since there is a good fit with a subjective emotional evaluation of the consumption experience of such a product. As a result, users are more familiar with evaluating a movie based on how they experience it rather than assigning an objective value for its quality. Accordingly, the familiarity hypothesis showed that are attached to labels they are familiar to and therefore are more likely to choose them (Weijters et al., 2013). Lastly, non-emotional labels are harder to evaluate, resulting in increased cognitive load, which in turn makes users more prone to response biases. Therefore, it is expected that:

H1: A scale with emotional labels at the end points will increase extreme response compared to a scale with non-emotional labels.

2.4 Midpoint of the Scale

Studies show that adding a midpoint to a rating scale may decrease the extreme response of users (Weijters et. al., 2010). Such an effect may be attributed to several reasons. First, users are more attached to labeled items and therefore are more likely to move towards the midpoint compared to when only the endpoints are labeled (Krosnick and Fabrigar, 1997). Relatedly, more people tend to select the midpoint in only endpoint labeled scales compared to fully labeled scales as intermediate responses become then more salient (Dillman and Christian, 2012; Weijters et al., 2010). Also, the presence of a midpoint increases the social desirability bias, since respondents select it more often in an attempt to not displease the interviewer (Garland, 1998). Moreover, the inclusion of a midpoint is related to the forced-choice bias, as ambivalent or indecisive respondents would be attracted to opt for such a neutral choice (Friedman & Amoo, 1999; Nowlis et al., 2002; Schaeffer and Presser, 2003). The increased use of the midpoint when it is offered is especially prevalent for verbal labels (compared to numerical labels) (O'Muircheartaigh et al., 1995).

As most studies in the past, prove the midpoint effect by introducing scales where a midpoint is not available (even numbered scales), such a structure is not widely used in rating websites. In contrary, I observe midpoint including odd-numbered scales in most websites; either 1-5 (5 points) or 1-10 (9 point) scales. The focus of this study is on the effect of labeling the midpoint compared to only labeling the endpoints (in that way the midpoint option is always available). I expect that the abovementioned evidence would be applied even in the more conservative implementation of merely adding a label on the midpoint. Priming the midpoint would make it more salient to raters (Dillman and Christian, 2012). This new reference point may cause adjustments in users' ratings and anchor them towards the midpoint (Schwarz et al., 1991; Tversky & Kahneman, 1974). Therefore:

H2: The presence of a neutral midpoint label in the rating scale will decrease extreme response.

Anchoring effects in a rating scale suggests that respondents focus first on the labeled categories then make adjustments in their product rating towards their final judgment. It was shown that respondents shifted to the midpoint more frequently when less intense labels were displayed ("agree" to "disagree" compared to "strongly agree" to "strongly disagree") (Dolnicar and Grun, 2013). As emotional labels are more familiar and representative to users in the context of experiential products (and hence attract responses), they would become stronger anchors in the rating process (Weijters et al., 2013) and therefore attenuate the effect of priming the midpoint of the scale. Additional studies found a U-shape relationship between emotional intensity and anchoring effect in that extreme (high or low) emotionality increases bidders' anchoring to the existing bid (Arana & Leon, 2008). Accordingly emotional labels in the endpoints activate such anchors and drive ratings away from the midpoint. Therefore:

H3: The negative effect of the presence of a neutral midpoint label on extreme response will be smaller for a scale with emotional labels at the end points (compared to non-emotional labels).

2.5 Size of the Rating Scale

Jacoby & Matell (1971) found that reliability of responses is independent of the size of the scale used. However, when the available options are fewer, respondents are forced to rate their continuously distributed perceptions on a rather discrete scale (Lehmann & Hulbert, 1972). Accordingly, increasing the number of response categories improves response reliability but such an improvement stagnates beyond a 5-points scale (Lissitz & Green, 1975). However there is evidence that the size of a rating scale influence the actual responses of raters. Larger scales increase the variance and decrease the average rating (Dawes, 2008; Preston and Colman, 2000). Rating scales with more points transmit more information to its respondents and (though increasing cognitive choice load) allow them to convey more easily their true answers (Cox 1980). As a result, a greater spread of answers decreases response biases. Correspondingly, Hui & Triandis (1989) found a lower extreme response tendency when rating scales were larger. Recent studies showed that the number of gradations of agreement decreases the extreme response in a response scale (Weijters et al., 2010) as well as increases the midpoint response style (Kieruj & Moors, 2010). Therefore:

H4: A larger size of the rating scale will decrease extreme response.

As emotional cues stimulate individual's memory systems by which felt emotions are being activated, it is expected that respondents confronted with an emotional rating scale apply a more intuitive rating compared to a non-emotional scale and therefore are more prone to an extreme response tendency. As a higher extreme response tendency would signify a shift away from the midpoint, shifting one gradation from the midpoint in a larger scale (e.g. 1-9 scale) would mean a smaller distance (1 out of 4 available gradation points) compared to one gradation shift in a 1-5 scale (1 out of 2 available gradation points)[2]. Alternatively, a non-emotional scale would lead users away from the extreme points. Therefore, one gradation shift from the extremes would move closer to the midpoint (and hence decrease the extreme response tendency) in a small scale, compared to a large scale. Therefore:

H5: The negative effect of a larger size of the rating scale on extreme response will be larger for a scale with emotional labels at the end points (compared to non-emotional labels).

3 Methodology

The conceptual model including the hypothesized effects is illustrated in Figure 1. To test the hypothesized effects, an online experiment was conducted. The empirical context of the study was movie ratings. The product choice is very appropriate to the product ratings since (a) movies are experience products which are vastly prone to subjectivity allowing for a greater variance in evaluations, (b) evaluating movies is a natural setting for users as movie rating websites belong to the most frequently visited websites worldwide (e.g. imdb) and (c) there is a within-industry variance in the way movie rating websites apply the rating systems. More precisely, participants were asked to watch three movie trail-

[2] A short 5-points scale offers 2 gradation points from the midpoint (1-2 or 4-5) whereas a larger 9-points scale offers 4 gradation points (1-4 or 6-9).

ers of not yet released movies[3] and rate them. I chose unreleased movies from different genres (i.e. comedy, adventure, animation) to account for unobservable variations in the ratings due to different levels of exposure to the movies. All three movies received average ratings (in retrospect) in popular rating websites (e,g., Imdb), so no quality differences exist among the movies.

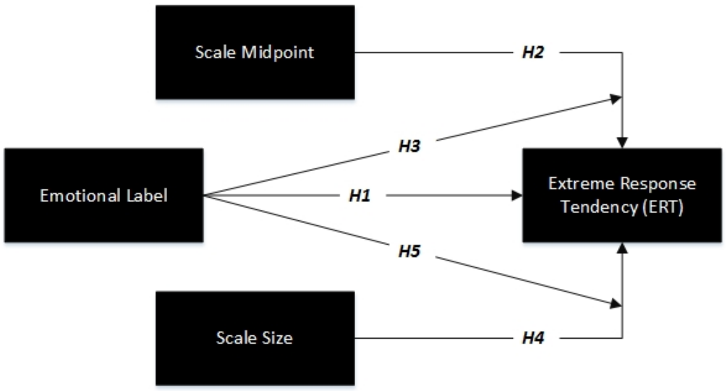

Figure 1. Conceptual Framework.

The rating scales were manipulated in a 2 (endpoint labels: Emotional vs. Non-Emotional) x 2 (midpoint label: Shown vs. Not Shown) x 2 (size: 5-point vs. 9-point scale) between-subjects design. Participants were randomly assigned to one of the eight conditions and were asked to rate all 3 movies (allowing us to control for carry over effects). The emotional endpoint labels were operationalized in a scale between "hate it" and "love it" compared to a non-emotional scale between "very poor" and "excellent". Such labels spun in a balanced way across a clearly favorable and unfavorable item and are widely used in real practice. The label of the midpoint was indicated as "neutral" and it was placed above the actual midpoint of the scale. No other intermediate points were labeled. The scale size was constructed based on either a five-point or a nine-point scale. Both scale sizes contained a midpoint and allowed for comparability of the extreme response tendency in the ratings. Regardless of the condition all scales were built in the same visual size to avoid the possible effect of the visual size on the extreme response (Dillman & Christian, 2002).

Raw Rating when size=5	1		2		3		4		5
Raw Rating when size=9	1	2	3	4	5	6	7	8	9
Adjusted Distance from midpoint	-2.0	-1.5	-1.0	-0.5	0	0.5	1.0	1.5	2.0
Extreme Response Tendency	2.0	1.5	1.0	0.5	0	0.5	1.0	1.5	2.0

Table 1. Original Ratings and Extreme Rating Tendency.

The dependent variable is the extreme response tendency in the ratings. Although many studies in the context of response styles measure extreme response in a rather absolute manner where only the choice of the endpoints is considered as extreme response (Steenkamp and Baumgartner, 2001;

[3] The movies used were: "Last Vegas", "Ender's Game" and "Free Birds". The study was conducted in October 2013 and the movies were about to be released the earliest in December 2013 or January 2014.

Weijters et al., 2010), I apply a more relaxed measurement as an indication of extreme response tendency. In this study, I measure extreme response tendency as the absolute distance from the midpoint of the scale (De Langhe et al., 2011; Van Herk et al., 2004). The extreme response tendency (ERT) ranges from 0 (midpoint) to 2 (endpoint). The distance from the midpoint was adjusted for a 9-points scale (divided by 2), to allow comparability (see Table 1 for an illustration).

Next to the main task, participants were asked a series of additional questions that helped eliminate alternative explanations of the effects. I asked participants how often do they watch movies (in TV, cinema, DVD, streaming, illegal download) and created a variable with the average of the top-3 scores (assuming that some of the options were substitute). I also asked respondents how knowledgeable do they consider themselves compared to an average movie viewer. Moreover, participants indicated whether they had seen or been aware of a list of 15 popular movies. Studies have shown that familiarity and involvement with the topic increases extreme response behavior (Arce-Ferrer, 2006). Next, I asked respondents to rank in order of preference 5 sets of 3 movies each. As movies in every set differed in terms of genre, I constructed a weight for the genres of the three movies in the main task. Also, I controlled for whether respondents were aware of the movie or had watched the movie trailer before taking the survey. I further controlled for a general extreme response style of the participants by using items from Greenleaf (1992). Participants were asked about the extent of agreement with five unrelated to the task (and to each other) statements on a 7-point scale (to be symmetric to the scale size of the main task regardless the allotted condition). Based on their answers I measured the absolute distance from the midpoint (to measure extreme response) as well as an acquiescence index (in terms of positiveness of the answers). I measured demographic characteristics, such as gender, age, education and country of origin. The latter was used based on the premise that cultural characteristics may influence response behavior of individuals (Baumgartner and Steenkamp, 2001; De Jong et al., 2008). To further investigate that assumption I collected the respective national scores in cultural differences in two relevant cultural dimensions: individualism (based on the assumption that users from high individualistic countries would deviate more from the average scores; Hong and Li, 2014) and uncertainty avoidance (users from high uncertainty avoidance cultures would be more reluctant to be extreme in their responses) (Harzing, 2006).

Participants were recruited via social networks (movie related groups), as well as through the student network of a large European university. To avoid processing ratings for movie trailers that were not actually watched, respondents who spend less than 30 seconds on a movie trailer were removed (unless they indicated that they had seen the trailer before; in that case I relaxed the imposed threshold).

4 Analysis and Results

4.1 Manipulation Checks

In this study the labels “Excellent” and “Very poor” were considered non-emotional and the labels “Love it” and “Hate it” emotional. To control whether participants derived the same meaning from the labels, they were asked whether they perceived the displayed labels as being emotional or not. The manipulation check was successful, as emotional labels were perceived significantly more emotional than the non-emotional labels ($M_{emotional}$=0.80 vs. $M_{non\text{-}emotional}$=0.57, F=12.65, p<0.01). Additionally, the labels were evaluated in terms of extremeness and strength. In addition, participants had to rate the endpoint labels (positive and negative) in terms of perceived strength and extremeness to make sure the scales were symmetric around the neutral midpoint. Both emotional and non-emotional labels used did not differ in terms of strength ($M_{excellent}$=72.6 vs. M_{love_it}=71.2, p>0.1 and M_{very_poor}=52.4 vs. M_{hate_it}=59.3, p>0.1). However, there is a significant difference in label extremeness ($M_{excellent}$=65.9 vs. M_{love_it}=56.3, p<0.05 and M_{very_poor}=47.1 vs. M_{hate_it}=57.9, p<0.05). These differences indicate that the perceived strength and extremeness of the labels need to be controlled for. Also, no significant differences in terms of user characteristics were found across the conditions.

4.2 Results

The total sample was 203 participants. The average age of the participants was 34 years, 45.3% were males, at least 59% were graduates and originating from 28 different countries (majority from Europe). Participants did not differ in terms of demographic characteristics across the 8 conditions. Ratings (converted in a 5-ponts scale) were normally distributed around an average of 3.45 and the average extreme response tendency was 0.90 out of a maximum of 2, in line with the typical distribution of ratings in experimental studies (Hu et al., 2009). All categories in the rating scales were chosen multiple times, indicating that on average participants considered all the range of available responses. On average, participants had an above average involvement in the movie category (M=4.33), considered themselves moderately knowledgeable (M=3.1) and had watched on average 10 out of the 15 movies from the control task. On average they spent 118 seconds per movie trailer (a trailer lasted 150 seconds, yet the time spent of those who had not seen the trailer before was 165 seconds). Since the hypothesized effects (rating scale condition) were invariant across the 3 movies participants rated, in this balanced panel dataset a random effects model with robust standard errors was preferred over a fixed effects model (as the latter would absorb these effects in the intercept).

Dependent variable:	Extreme Response Tendency (*regression*)		Extreme Response Style[(2)] (*logit model*)	
	β	*S.E.*	*β*	*S.E.*
Intercept	1.50**	0.41	2.76	1.77
Emotional Label	0.04	0.11	-0.52	0.42
Midpoint label	-0.25**	0.08	-0.98**	0.41
Emotional Label * Midpoint label	0.26*	0.11	-1.03*	0.53
Size	0.08	0.08	-0.26	0.40
Emotional Label * Size	-0.24*	0.12	-1.44**	0.57
Lagged Extreme Response	0.13*	0.06	0.35	0.35
Label Strength & Extremeness	*Included*		*Included*	
Movie fixed effects	*Included*		*Included*	
ERS generic scale [(3)]	0.09*	0.06	0.23	0.29
Positiveness of ERS	0.02	0.05	0.17	0.21
Control Variables [(4)]	*Included*		*Included*	
N	203		203	
R^2-Between	0.35		Log-Likelihood: -235.5	
R^2-Overall	0.12			

Note. (1) * *p*<.05, ** *p*<.01; (2) alternative measure of DV, see Additional Findings; (3) measured according to the respective dependent variable; (4) Control variables: Age, gender, Education, Country of origin (dummy), Seen/aware of trailer, perceived movie knowledge, , Movie Involvement, Movie preferences, time spent.

Table 2. Results from random effects models.

The results of the random effects regression model on extreme response tendency confirmed some of the expected hypotheses (Table 2). Regarding the direct effects of the manipulations, only the labeling of the midpoint showed a significant direct negative effect on extreme response[4]. For rating scales where a “neutral” label was displayed above the midpoint, the average distance from the midpoint was

[4] Though I find a marginally larger extreme response tendency for emotional labels ($M_{emotional}$=0.95 vs. $M_{non-emotional}$=0.85, F=3.87, p<0.01).

significantly smaller than rating scales with only endpoint labels. Therefore, H2 was confirmed. However, the effect of midpoint label seems to only hold for non-emotional labels, as a positive interaction effect between emotional labels and midpoint labels on extreme response was found. Therefore H3 is confirmed. Regarding the size of the scale, though I cannot confirm a direct effect (H4), I found a negative interaction between emotional labels and size on extreme response tendency. A larger scale with emotional labels reduces the extreme response tendency of raters, thereby confirming H5. Additionally, there is an internal anchoring (carry over) effect in the ratings (in line with De Jong et al., 2012), as a significant effect of the extreme response tendency in previous ratings on current rating is found. I further found a marginally significant positive effect of ERS (based on Greenleaf 1992), capturing a general predisposition in extreme response style. Movie specific fixed effects indicated that the presence of systematic differences in extreme response tendency across movies (thereby capturing general quality related differences in the ratings). Further, none of participants' involvement-related control variable as well as their perceived strength and extremeness of the labels did not yield any significant results. None of the demographic characteristics had a significant effect on extreme response tendency.

4.3 Additional Findings and Robustness Checks

To deepen the understanding in the effects of the treatments in extreme response of movie ratings, I conducted some further analyses to rule out some alternative explanations. First, I included in the model the interaction between midpoint label and size and, whereas the direct effect of midpoint label disappears, I found a negative significant effect of the interaction effect, showing that a midpoint label decreases extreme response when size of the scale is large. The rest of the results remained unaffected. I further included a three-way interaction, which yielded no significant results. Moreover, I applied an additional operationalization of extreme response for participants where extreme response was measured in a more absolute way (1 if endpoint is chosen, 0 for all other intermediate points) (Weijters et al., 2010). I conducted a random effects logit model, with the same set of independent variables and found that the results were comparable to the original model of continuous extreme response tendency (Table 2). Finally, I tested the effects of the variations in the rating scales on the actual rating given (normalized across sizes) as well as the acquiescence of the ratings (1 if rating above midpoint) and found that size increases the acquiescence of the ratings, yet this effect is attenuated when the labels at the endpoints of the scale are emotional.

To ensure that the results of the model were not observed due to some confounding factors, I ran some additional robustness checks. First, I controlled for the acquiescence of the rating and found that positive ratings tend to be more extreme than negative ones, however the main effects remain unaltered. Accounting for a non-linearity of age, I found a U-shape effect on extreme response tendency (De Jong et al., 2012). However, when imposing stricter restrictions for the age, results remain consistent. I also relaxed or restricted further the time threshold (time spent watching a trailer) but results were robust. To control for possible endogenous effects on general response style and extreme response tendency, I first modelled movie ratings (ERT) as a function of the average ERS (based on Greenleaf 1992). The residuals of this model characterize the extreme response tendency that cannot be explained by the endogenous ERS. Therefore, I included these residuals (along with the general ERS) as independent variables in the model but the results remained consistent to the original model. Results are stable when I control for cultural differences by using individualism and uncertainty avoidance national scores of participants (Harzing, 2006). Finally, I examined the role of the grading system (e.g. 5, 10, or 20 scale) in education of different countries (and its fit to the respective scale size), as participants may accordingly use and value differently the scales, but results remained robust and no grading related effect was found.

5 Conclusion

5.1 Discussion of the findings

This study shows that product evaluations are subject to the way information is conveyed in the rating scales where products are being evaluated. I focus on extreme response tendency, an extensively observed bias in the context of response behavior (De Jong et al., 2008; Weijters et al., 2010). Extreme response is defined as the tendency to opt for the extreme points of a rating scale and is implemented as the absolute distance from the midpoint. Whereas response theories are mostly empirically tested in survey response, I find indication of such a response bias in product ratings as well. More precisely, I enrich the insights in the field by focusing on the labels (emotional vs. non-emotional) in the endpoints of the rating scale, the labeling of the midpoint and the size of the rating scale. I use experience goods (movies), whose objective quality index cannot be extracted based on product descriptors. Thus, potential consumers have to rely on user-generated ratings to support their choices. In fact, movie-rating websites (e.g. IMDB) belong to the most visited websites worldwide and popular movies receive more than 1 million ratings. As a result, rating a movie is a natural setting to apply in this study.

I designed a 2^3 between-subject experiment where participants were asked to rate 3 unreleased (at the time of the study) movies. One of the main findings is that using a neutral label above the midpoint decreases the extreme response tendency. Users are anchored in the labeled points. Therefore, priming the midpoint makes that option more salient to users (Dillman and Christian, 2012). As a result, that would make users readjust their ratings and anchor them towards the midpoint, compared to when only endpoints of the scale are labeled. Further, studies showed that respondents choose the midpoint more if it is labeled as ambivalent compared to neutral (Klopfer and Madden, 1980; Schaeffer and Presser, 2003). Though the effect of the inclusion of a midpoint has been shown in past studies, I use a rather conservative approach where I manipulate only the presence of a label above that point and using a moderate label. Next, I focus on the effect of endpoint labels in the rating scale. More precisely, I focus on the emotionality of the labels, based on the presumption that the intensity of the rating scale labels as well as the familiarity with the labels strongly influence responses. Such evidence in the literature has been contradictory (Weijters et al., 2013) and that is depicted in the results, where I find no direct effect of emotional labels on extreme response tendency (hence not supporting H1). However, I provide an additional explanation for the lack of a strong direct effect as I find a negative interaction effect between emotional labels and midpoint label. Though a labeled midpoint attracts raters, such an effect is attenuated when the labels at the endpoints are emotional, since they serve additional contrasting anchors and drive ratings away from the midpoint. Additionally, emotional labels repel users from the extreme points (hence decreasing extreme response tendency) when the size of the scale is larger. Users move in steps across the gradation points of the scale. As emotional labels shift ratings away from the midpoint, a step away from that point is still closer (in ratio to the total distance they can cover) in a larger scale compared to a small scale. Therefore, emotional labels seem to attract users but their responses are less susceptible to extreme response tendencies when the rating scale is extended.

5.2 Academic and practical contributions

This study makes some important contributions to IS literature. There is a nascent body of research in IS dealing with the prevalent importance of user generated ratings and word of mouth (Forman et al., 2014; Yin et al., 2014) as well as self-selection biases occurring in rating systems (Hu et al., 2009; Li and Hitt, 2008). However, the role of response sets in product rating systems has been neglected. Research on response biases has been widespread, yet focused in the fields of survey design (van Vaerenbergh & Thomas, 2013) and marketing research (Baumgartner and Steenkamp, 2001; Weijters et al., 2010). Product ratings differ from survey responses in that they are less prone to social desirability and acquiescence biases, and they refer to predefined evaluations of product experiences. In addition, a contribution is made to user generated ratings research by investigating the emotionality of the end-point labels of a rating scale. Previous studies analyzed rating behavior and its consequences to further ratings or product performance, focusing on the numerical part of the ratings (rather than the labels these numbers correspond to). I find that the use of emotional labels attracts users to the extreme points in conjunction with the explicit labeling of a neutral midpoint, yet such an effect is reversed when the size of the rating scale is extended. Such effects and interactions provide new insights in understanding the role of response sets in users' products ratings. Finally, I operationalize extreme response in a continuous way (unlike binary methods in previous studies), which allows us to get more insightful conclusions about how consumers use the range of available rating scores.

Besides the academic implications, this study provides practical insights that can be used in the application and use of current rating systems. Knowing that the use of the studied variations in the rating scales might decrease extreme response tendency, such treatments might help correct in a natural way the purchasing and under-reporting biases that lead to a J-shaped distribution of product ratings (Hu et al., 2009). Given the self-selection of the raters that can be accounted for (Li and Hitt, 2008), this study provides an additional way to depolarize ratings. Such a correction becomes essential if I take into account that many companies use consumer ratings in order to create further recommendations. Also, many companies use models that predict the performance of their products (in terms of sales) based on user generated ratings, and they accordingly plan their marketing budget allocation. The quality of such recommendations and predictions can be seriously distorted if the effects of rating scale variations are not accounted for. Finally, though companies would want to inflate the ratings of their products, they should also aim at deflating excessively positive ratings, as there is evidence that users discount them as untrustworthy (Ludwig et al., 2013).

5.3 Limitations and further research

Despite the above contributions, some limitations of the study provide opportunities for future research. First, the use of an experimental approach might render the provided ratings less sensitive to under-reporting bias. The provided ratings are normally distributed compared to the observed J-shaped distribution of actual ratings, in line with the findings of Hu et al. (2009). The next step would be to examine if extreme response tendency follows the same mechanism in actual ratings from websites that differ in terms of rating scales. For example, movies are rated on 10-points (IMDB) or 5-points scales (Rottentomatoes), with non-emotional (Movielens) or emotional labels (Netflix). Other variations in a rating scale can also be explored. Recent studies investigated the use of emoticons, colors or the place of the labels on response biases (De Langhe et al., 2011; Tourangeau et al., 2013). Fully labeled rating scales can also be implemented, investigating whether they become more insightful or add to users' cognitive load. Also, future studies can vary the levels of emotional intensity as well as the balance between negatively and positively valenced labels and test the actual meaning respondents derive from various labels. Further, for the vast majority of participants, the study was not in their native language. Since individuals tend to become more extreme in their responses when answering in their second language (De Langhe et al., 2011), it would be interesting to replicate the study to native speakers. Finally, though the setting of movie ratings is very natural for users due to the characteristics of the product category but also the popularity of movie rating websites, future research could investigate the generalizability of the findings. Product categories such as hotels or restaurants (where product quality can be improved based on the ratings) may be a very fertile field to test the effects of variations in the rating scales.

References

Albaum, G., C. Roster, J. H. Yu and R. D. Rogers (2007). “Simple rating scale formats: Exploring extreme response.” *International Journal of Market Research* 49, 633–650

Arce-Ferrer, A. J. (2006). “An Investigation Into the Factors Influencing Extreme-Response Style Improving Meaning of Translated and Culturally Adapted Rating Scales.” *Educational and Psychological Measurement* 66(3), 374-392.

Arana, J. E. and C. J. Leon (2008). “Do emotions matter? Coherent preferences under anchoring and emotional effects.” *Ecological Economics* 66(4), 700-711.

Baumgartner, H. and J.B.E.M. Steenkamp (2001). “Response Styles in Marketing Research: A cross national investigation.” *Journal of Marketing Research* 38(4), 143-156.

Chevalier, J. A. and D. Mayzlin (2006). “The Effect of Word of Mouth on Sales: Online Book Reviews.” *Journal of Marketing Research* 43(3), 345–354.

Clemons, E. K., G. G. Gao and L. M. Hitt (2006). “When online reviews meet hyperdifferentiation: A study of the craft beer industry.” *Journal of Management Information Systems* 23(2), 149-171.

Cox, E. P. (1980). “The Optimal Number of Response Alternatives.” *Journal of Marketing Research* 17(4), 407-422.

Dawes, J. (2008). “Do Data characteristics Change According to the Number of Scale Points used? An Experiment using 5-point, 7-point and 10-point scales.” *International Journal of Market Research* 51(1), 1-20.

De Jong, G. M., J.B.E.M. Steenkamp, J.P. Fox and H. Baumgartner (2008). “Using Item Response Theory to Measure Extreme Response Style in Marketing Research: a Global Investigation.” *Journal of Marketing Research* 45(1), 104-115.

De Jong, M. G., D. R. Lehmann and O. Netzer (2012). “State-Dependence Effects in Surveys.” *Marketing Science* 31(5), 838-854.

De Langhe, B., S. Puntoni, D. Fernandes, S.M.J. van Osselaer (2011). “The Anchor Contraction Effect in International Marketing Research.” *Journal of Marketing Research* 48(2), 366-380.

Dellarocas, C. (2003). “The digitization of word of mouth: Promise and challenges of online feedback mechanisms.” *Management Science* 49(10), 1407-1424.

Dellarocas, C., X. M. Zhang and N. F. Awad (2007). “Exploring the value of online product reviews in forecasting sales: The case of motion pictures.” *Journal of Interactive Marketing* 21(4), 23-45.

Dillman, D. A. and L. Christian (2002). *The influence of Words, Symbols, Numbers and Graphics on Answers to Self-administered Questionnaires: Results from 18 Experimental Comparisons*. URL: http://sesrc.wsu.edu/dillman/papers/2002/theinfluencewords.pdf (visited on 11/23/2014).

Dolnicar, S. and B. Grun (2013). “Translating Between Survey Formats.” *Journal of Business Research* 66(9), 1298-1306.

Duan, W., B. Gu and A. B. Whinston (2008a). “The dynamics of online word-of-mouth and product sales—An empirical investigation of the movie industry.” *Journal of Retailing* 84(2), 233-242.

Duan, W., B. Gu and A. B. Whinston (2008b). “Do online reviews matter?—An empirical investigation of panel data.” *Decision Support Systems* 45(4), 1007-1016.

Floyd, K., R. Freling, S. Alhoqail, H. Y. Cho and T. Freling (2014). “How Online Product Reviews Affect Retail Sales: A Meta-analysis.” *Journal of Retailing* 90(2), 217-232.

Forman, C., A. Ghose and B. Wiesenfeld (2008). “Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets.” *Information Systems Research* 19(3), 291-313.

Friedman, H. H. and T. Amoo (1999). “Rating the Rating Scales.” *The Journal of Marketing Management* 9(3), 114-123.

Garland, R. (1991). “The mid-point on a rating scale: is it desirable?” *Marketing Bulletin* 2(1), 66-70.

Greenleaf, E. A. (1992). “Measuring extreme response style.” *Public Opinion Quarterly* 56(3), 328-351.

Godes, D. and D. Mayzlin (2004). “Using online conversations to study word-of-mouth communication.” *Marketing Science* 23(4), 545-560.

Hartley, J. and L. R. Betts (2010). “Four Layouts and a Finding: the Effects of Changes in the Order of the Verbal and Numerical Values on Likert-type Scales.” *International Journal of Social Research Methodology* 13(1), 17-27.

Harzing, A. W. (2006). “Response Styles in Cross-national Survey Research A 26-country Study.” *International Journal of Cross Cultural Management* 6(2), 243-266.

Hong, Y. and C. Li (2014). “The 'Assertive' Consumers: Effects of Culture in Online Word of Mouth.” Available at SSRN: http://ssrn.com/abstract=2428407

Hu, N., J. Zhang and P. A. Pavlou (2009). “Overcoming the J-shaped distribution of product reviews.” *Communications of the ACM* 52(10), 144-147.

Hui, C. H. and H. C. Triandis (1989). “Effects of culture and response format on extreme response style.” *Journal of Cross-Cultural Psychology* 20(3), 296-309.

Jacoby, J. and M. S. Matell (1971). “Three-Point Likert Scales Are Good Enough.” *Journal of Marketing Research* 8(4), 495-500.

Kalton, G., J. Roberts and D. Holt (1980). “The effects of offering a middle response option with opinion questions.” *Journal of the Royal Statistical Society* 29(1), 65-78.

Kieruj, N. D. and G. B. D. Moors (2010). “Variations in response style behavior by response scale format in attitude research.” *International Journal of Public Opinion Research* 22(3), 320-342.

Kieruj, N. D. and G. B. D. Moors (2013). “Response style behavior: question format dependent or personal style?” *Quality & Quantity* 47(1), 193-211.

Klopfer, F. J., and T. M. Madden (1980). “The Middlemost Choice on Attitude Items Ambivalence, Neutrality, or Uncertainty?” *Personality and Social Psychology Bulletin* 6(1), 97-101.

Knowles, E. S. and C. A Condon (1999). “Why people say" yes": A dual-process theory of acquiescence.” *Journal of Personality and Social Psychology* 77(2), 379.

Krosnick, J. A. and L. R. Fabrigar (1997). “Designing rating scales for effective measurement in surveys.” In: *Survey measurement and process quality.* Ed. by L. E. Lyberg, P. Biemer, M. Collins, E. D. de Leeuw, C. Dippo, N. Schwarz and D. Trewin, Hoboken, NJ: Wiley. p. 14–164.

Lau-Gesk, L. and J. Meyers-Levy (2009). “Emotional Persuasion: when the valence versus the resource demands of emotions influence consumers’ attitude.” *Journal of Consumer Research* 36(4), 585-599.

Lehmann, D. R. and J. Hulbert (1972). “Are Three-Point Scales Always Good Enough?” *Journal of Marketing Research* 9(4), 444-446.

Lench, H. C., S. A. Flores and S. W. Bench (2011). “Discrete Emotions Predict Changes in Cognition, Judgement, Experience, Behaviour and Physiology: a Meta Analysis of Experimental Emotion Elicitations.” *Psychological Bulletin* 137(5), 834-855.

Li, X., & Hitt, L. M. (2008). Self-selection and information role of online product reviews. Information Systems Research, 19(4), 456-474.

Lissitz, R. W. and S. B. Green (1975). “Effect of the Number of Scale Points on Reliability: A Monte Carlo Approach.” *Journal of Applied Psychology* 60(1), 10-13.

Ludwig, S., K. de Ruyter, M. Friedman, E. C. Brüggen, M. Wetzels and G. Pfann (2013). “More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates.” *Journal of Marketing* 77(1), 87-103.

Moe, W. and D. A. Schweidel (2011). “Online Product Opinions: Incidence, Evaluations and Evolution.” *Marketing Science* 31(3), 372-386.

Mudambi, S. M. and D. Schuff (2010). “What makes a helpful online review? A study of customer reviews on Amazon.com.” *MIS Quarterly* 34(1), 185-200.

Myers, J. H. and W. G. Warner (1968). “Semantic Properties of Selected Evaluative Adjectives.” *Journal of Marketing Research* 5(4), 409-412.

Nichols, A. L. and J. K. Maner (2008). “The good-subject effect: investigating participant demand characteristics.” *Journal of General Psychology* 135(2), 151-165

Nowlis, S. M., B. E. Kahn and R. Dhar (2002). "Coping with ambivalence: The effect of removing a neutral option on consumer attitude and preference judgments." *Journal of Consumer Research* 29(3), 319-334.

O'Muircheartaigh, C., G. Gaskell and B. D. Wright (1995). "Weighing Anchors: Verbal and Numerical Labels for Response Scales." *Journal of Official Statistics* 11(3), 295-307.

Ostrom, T. M. (1966). "Perspective as an Intervening Construct in the Judgment of Attitude Statements," *Journal of Personality and Social Psychology* 3(2), 135–44.

Paulhus, D. L. (1991). "Measurement and control of response bias." In: *Measures of personality and social psychological attitudes.* Ed. by J. Robinson, P. Shaver and L. Wrightsman. New York. p. 1-59.

Preston, C. C. and A. M. Colman (2000). "Optimal Number of Response Categories in Rating Scales: Reliability, Discriminating Power, and Respondent Preferences." *Acta Psychologica* 104(1), 1-15.

Schaeffer, N. C. and S. Presser (2003). "The science of asking questions." *Annual Review of Sociology* 29, 65-88.

Schwarz, N., H. Bless, F. Strack, G. Klumpp, H. Rittenauer-Schatka and A. Simons (1991). "Ease of retrieval as information: Another look at the availability heuristic." *Journal of Personality and Social Psychology* 61(2), 195.

Sridhar, S. and R. Srinivasan (2012). "Social influence effects in online product ratings." *Journal of Marketing* 76(5), 70-88.

Tellis, G. J. and D. Chandrasekaran (2010). "Extent and impact of response biases in cross-national survey research." *International Journal of Research in Marketing* 27(4), 329-341.

Tourangeau, R., M. P. Couper and F. Conrad (2007). "Color, Labels and Interpretive Heuristics for Response Scales." *Public Opinion Quarterly* 71(1), 91-112.

Tourangeau, R., M. P. Couper, M. P. and F. Conrad (2013). "Up Means Good: The Effect of Screen Position on Evaluative Ratings in Web Surveys." *Public Opinion Quarterly*, 77(S1), 69-88.

Tversky, A. and D. Kahneman (1974). "Judgement under Uncertainty: Heuristics and Biases." *Science* 185(4157), 1124-1131.

Upshaw, H. S. (1965), "The Effect of Variable Perspectives on Judgments of Opinion Statements for Thurstone Scales: Equal-Appearing Intervals," *Journal of Personality and Social Psychology* 2 (1), 60–69.

Weijters, B., E. Cabooter and N. Schillewaert (2010). "The Effect of Rating Scale Format on Response Styles: the Number of Response Categories and Response Category Labels." *International Journal of Research in Marketing* 27(3), 236-247.

Weijters, B., Schillewaert, N., Geuens, M. (2008). "Assessing Response Styles Across Modes of Data Collection." *Journal of the Academy of Marketing Science* 36(3), 409-422.

Weijters, B., M. Geuens and H. Baumgartner (2013). "The effect of familiarity with the response category labels on item response to Likert scales." *Journal of Consumer Research* 40(2), 368-381.

Wildt, A. R. and M. B. Mazis (1978). "Determinants of Scale response: Label versus Position." *Journal of Marketing Research* 15(2), 261-267.

Van Vaerenbergh, Y. and T. D. Thomas (2013). "Response styles in survey research: A literature review of antecedents, consequences, and remedies." *International Journal of Public Opinion Research* 25(2), 195-217.

Voss, K. E., E. R. Spangenberg and B. Grohmann (2003). "Measuring the Hedonic and Utilitarian Dimensions of Consumer Attitude." *Journal of Marketing Research* 40(3), 310-320.

Yin, D., S. D. Bond and H. Zhang (2014). "Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews." *MIS Quarter*ly 38(2), 539-560.

Zhu, F. and X. M. Zhang (2010). "Impact of Online Consumer Reviews on Sales: the Moderating Role of Product and Consumer Characteristics." *Journal of Marketing* 74(2), 133-148.