Available online at www.sciencedirect.com
ScienceDirect
Journal of Consumer Psychology 19 (2009) 158–170
Journal of CONSUMER PSYCHOLOGY

Scale-dependent automatic shifts in brand evaluation standards

Claudiu V. Dimofte *, Johny K. Johansson

Department of Marketing and International Business, McDonough School of Business, Georgetown University, USA

Available online 10 March 2009

Abstract

Different schema-based expectations for competing brands can produce shifting evaluative standards in consumers' relative ratings of these brands. This shift in standards differentially affects objective (number-based) and subjective (word-based) rating scales. Several studies support the proposition that a brand rated as objectively inferior to another can be subjectively perceived as equivalent to—or even better than—the same brand. Such anomalous response inconsistency originates in consumers' recourse to different expectations for the competing brands and their automatic adjustment of expectations for the inferior brand when responding to subjective measures. The implicit relaxing of evaluative standards for the objectively dominated brand leads to an unconscious upward bias in its subjective ratings. The effect is moderated by consumer expertise and is asymmetric—it is not accompanied by an escalation of evaluative standards for objectively dominating brands.

Keywords: Expectations; Ratings; Automatic processing; Measurement; Consumer expertise

Imagine being in the market for a luxury family sedan and watching an advertisement for the new Hyundai Azera entrant in the segment. Among the series of attributes it mentions, you learn about its new 234 horsepower engine. You also know that Hyundai's engines on earlier models have generally put out under 220 hp (the typical cutoff for a powerful engine in a family sedan) and thus well below category competitors such as the Cadillac CTS. How would you rate this Azera's engine power if you had to use the following scale: extremely poor–very poor–poor–average–good–very good–extremely good? It is likely that the verbal descriptor that comes to mind, all things considered, is *very good*. On the other hand, Cadillac has featured motors of 250 hp or more for years and is now advertising 263 horses. Now how would you describe the engine performance of the CTS? Likely, *very good* as well. The fact that two brands scoring quite differently (234<263) on the objective, horse-power attribute display performance that is at the same time evaluated equally in subjective, engine power terms (*very good*) is an illustration of a scale-induced cognitive phenomenon with the potential to play an intriguing and important role in brand evaluations.

Scales that require respondents to rate marketing stimuli on numeric-based units (e.g., number of dots-per-inch, calories, waiting minutes, horsepower, etc.) are commonplace in consumer evaluations of products, services, or brands. So are word-based rating scale evaluations (e.g., for screen resolution, energy content, time savings, engine power, etc.) where the endpoints are anchored by adjectives such as *poor* or *excellent*. A considerable amount of literature has demonstrated many of the strengths but also many of the pitfalls inherent to these scales. For example, it is well established that a problem with word-based scales is the idiosyncratic meaning that different respondents give to the verbal descriptors of interval and end points (Jacoby, 1978). Thus, adjectives such as *fair* or *good* are not interpreted uniformly across all individuals. It is also known that slight changes in the verbal descriptors employed can make for considerable shifts in perceptions and ratings. For example, changing an anchor description from *very inexpensive* to *very cheap* is likely to induce a change in respondents' frame of reference and could thus affect the observed scores significantly (Gannon & Ostrom, 1996).

Respondents' idiosyncratic interpretations of verbal descriptors can generally be treated as random noise, a problem usually managed by employing larger samples. However, we argue that a common but unconscious reaction to many such rating tasks may sometimes influence consumers in a systematic manner, in effect biasing their responses. This reaction involves participants' use

* Corresponding author. McDonough School of Business, Georgetown University, Washington, DC 20057, USA.
E-mail address: dimofte@msb.edu (C.V. Dimofte).

1057-7408/$ - see front matter © 2009 Society for Consumer Psychology. Published by Elsevier Inc. All rights reserved.
doi:10.1016/j.jcps.2009.02.009

of specific expectations to guide their evaluation process, expectations which automatically shift the framing and thus the meaning of a scale. We argue that for consumers such cognitive adjustments are dependent on the type of scale employed.

The main thrust of our argument draws on recent research by psychologists on cognitive response measurement, wherein a distinction is made between *objective* ratings collected on interval scales (expressed in common-rule units such as dollars, seconds, etc.) and *subjective* ratings elicited via scales anchored by words (e.g., unsuccessful–successful, slow–fast, etc.). Briefly put, this line of research shows that individuals shift their subjective ratings upwards for a stimulus associated with low expectations, thus reducing, eliminating, or even reversing the gap observed on objective scales between this stimulus and one associated with higher expectations. The present work argues that these results are directly relevant to the context of brand evaluations, by contrasting word-based (i.e., subjective) rating scales with numerical unit-based (i.e., objective) scales.

The next section presents the background psychology research in more depth. We then draw on relevant marketing literature on branding, using the established finding that brand identity and image produce well-defined consumer expectations about performance and satisfaction. The subsequent section discusses this conceptual account further and hypothesizes specific relationships between consumer expectations and cognitive responses, while introducing consumer expertise as an important moderating variable. We subsequently present a series of experimental studies aiming to first establish the strength of the effect on ratings, next to explore and explain the nature of the underlying cognitive process in more depth (including establishing mediators and moderators), and finally to predict and adjust for its occurrence. The article ends with a general discussion of the theoretical and practical significance of the findings and suggestions for future research.

Background literature

The shifting standards effect

Biernat, Manis, and Nelson (1991) developed a schematic model of a "shifting standards effect" (hereinafter SSE) to explain the process by which prior, schema-based expectations can impact the word-based ratings elicited from respondents. Their SSE model proposes that recourse to commonly used stereotypes will change the meaning of subjective scales for respondents and thus shift their evaluative scores. This shift will not occur for objective scales.

Biernat et al. (1991) empirically tested the model in the context of judgments about the financial success of several profiled men and women. When actual dollar figures for annual income were used as scale units, the resulting target ratings reflected the judges' more objective knowledge (i.e., factual beliefs) confirming the stereotype that males earn more than females. However, when word-based scales were used (anchored by *financially very unsuccessful* and *financially very successful*), judges differentially adjusted the meanings of the anchors for the two genders, rating the same profiled women as more successful than the same men. The subjective scales effectively reversed the well-known income stereotype (Biernat et al., 1991). The SSE model argues therefore that in the latter case commonly used stereotypes shifted the evaluative judgment by changing the meaning of the scale for the evaluated stimuli. Whereas prior expectations suggest that a woman will earn significantly less than a man in terms of annual dollars, this discrepancy will not be evident if earnings are measured via subjective, financial success scales.

Why would individuals "cut slack" to objectively lower-performing targets? Although not addressed in the original SSE work, we believe that the answer lies with the inherently human trait of providing emphatic support for those trailing, a predisposition conceptually captured in the *support for the underdog effect* in cognitive psychology (Vandello, Goldschmied, & Richards, 2007). As a target with low expectations of success (and therefore at a competitive disadvantage), the underdog is perceived in a more favorable light, so that its performance is seen in such a way as to compensate for the disadvantage. In the process, respondents attribute greater effort on the part of the underdog, which in turn reinforces their liking thereof (Vandello et al., 2007; also see Cho & Schwarz, 2008). The use of subjective response scales apparently makes the effect more likely, as the use of words is more readily suited for the compensatory processing (i.e., it is easier to claim that an inferior brand is *good* when hard numbers would not allow for such favoritism).

It is well established that brand identity and image produce well defined consumer expectations of performance and satisfaction. Keller (1993) argues for example that brands guide the processing of product information by discriminating between brands that engender positive expectations and those associated with negative expectations. Thus, we are assessing the existence of the SSE in consumer response to marketing scales, while also addressing a mediator (expectations) and a moderator (consumer expertise) of the effect.

There have been previous discussions in the literature on the differences among various types of scales (including between what we have termed objective and subjective scales). Rose, Miniard, Barone, Manning, and Till (1993) argue that the sensitivity of scales is affected by the correspondence between the encoding frame used during exposure and the cognitive frame imposed by the response measure. Viswanathan and Childers (1996) show that, relative to verbal information, numerical information is recognized and recalled more accurately and is generally more specific and less tainted by inherent meaning variations. According to their results, numerical units might be "more objective" than words. Similarly, when decisions are of critical importance (e.g., health risk assessments from physicians), research has shown that individuals prefer and trust numerical statements more than verbal information (Gurmankin, Baron, & Armstrong, 2004; Theil, 2002). Recently, Kruger and Vargas (2008) found that price comparisons (percentage differences) made on ratio scales produce varying subjective difference perceptions depending on the target of the comparison, another argument for the prevalence of scale-driven judgment biases in consumer evaluations.

Research hypotheses

Hypothesis 1: The SSE in brand ratings

Our starting point is the notion that when two brands evoke differing expectations of performance on some dimension, the observed differences in performance evaluations on that dimension may depend on the type of scale employed (i.e., numerical or word-based). According to the SSE, individuals implicitly use a lower (i.e., less demanding) standard when assessing a product that they expect to display lesser performance and for example judge that "for a Hyundai, this engine is quite powerful." The resulting discrepancy will lead their word-based, subjective judgments of the two brands to be closer than their more objective, numerical-unit ratings. Thus, under a scenario where consumers expect brand A to be superior to brand B, the following should occur:

H1. The use of subjective judgment standards (inherent to word-based measures) will automatically shift consumers' reported evaluations in accordance with prior brand expectations. Given the posited scenario, brand A will be judged as superior to brand B on numerical-unit scales, but this difference will be attenuated on word-based scales.

Low versus high expectations

As originally introduced in the literature, the SSE has been limited in focus to the lower end of individuals' expectations distributions. However, brands that engender highly positive expectations could also be more exigently scrutinized and unconsciously subjected to more stringent evaluative standards. As such, a consumer could feel that "for a Sony, this product is not as uniquely outstanding as I expected" and a bias against the brand may be observed on subjective (vs. objective) scales.

There are conceptual arguments both against and in favor of the potential symmetry of the SSE. On the one hand, it could be argued that the upper end of the expectations distribution will exhibit a ceiling effect. In other words, whereas consumers are quite good at knowing the minimal levels of performance that are acceptable (and thus adept at sensing surprise when they are surpassed), they are less aware of the truly highest possible levels thereof (and thus less able to feel let down). On the other hand, previous research has argued that high equity brands signal high quality and low perceived risk, therefore increasing consumer-expected utility (Erdem & Swait, 1998). If this is the case, then superior brands should be subject to more demanding expectations, such that consumers will shift their evaluative standards upwards for these brands. Thus, H1 could also be predicted from the perspective of the superior brand (i.e., brand A) as well, and the precise contribution of the brand A and B expectations to the observed SSE will have to be empirically disentangled.

The moderating role of consumer expertise

The managerial literature makes a clear distinction between the benefits that a brand or product provides and the product attributes that convey those benefits. A simple way to distinguish between the two involves the fact that attributes are essentially product features (most often concrete), whereas benefits result from these features and provide value to the consumer (thus virtually always abstract). In terms of the cognitive mode induced in the consumer's mind, the concrete-abstract distinction between attributes and benefits parallels our earlier juxtaposition of objective and subjective scales. Intuitively, objective scales are largely attribute-type scales, whereas subjective scales can be interpreted as largely benefit-type scales. When assessing the likely processor speed of an advertised computer on a 1.8–2.0–2.3–2.6–3.1 GHz scale, a consumer is essentially evaluating a concrete product attribute. However, when faced with the same task on a very slow–slow–average–fast–very fast scale, the consumer is more likely to settle into an abstract, benefit-oriented evaluation mode. As the SSE proposes that subjective scales have the potential to bias consumer response, it becomes relevant to search for individual difference variables that may make some consumers more susceptible to this bias by showing a processing preference for the benefit-oriented subjective scales.

Alba and Hutchinson (1987) proposed that individuals unfamiliar with a product category have yet to develop their domain knowledge. Accordingly, they are naturally inclined to rely on summary, stereotypical brand perceptions. This top-down processing is different from experts' bottom-up processing, which involves close scrutiny of product attributes in order to form evaluations. Along the same lines, Maheswaran and Sternthal (1990) evaluated the impact of information type on the processing and evaluation of product messages by novice and expert consumers. Novices processed the information in detail only when benefit information was presented, whereas experts engaged in detailed processing of information only when attributes were present in the message (Maheswaran & Sternthal, 1990; also see Monga & John, 2008).

Finally, work by Biswas and Sherrell (1993) found that the effect of brand image (i.e., a conceptual proxy for our brand stereotype construct) on internal price standards was greater for novices, who made more recourse to brand-driven expectations than experts.

These lines of research suggest that expertise is directly relevant to the SSE. As novices are likely to rely on brand stereotypes in their evaluations, their responses on objective scales should reflect this reliance and thus show for example that a Hyundai is less liked than a Cadillac. At the same time, their evaluation of brands via a benefit-focused approach suggests that on subjective scales they are likely to cut slack to Hyundai and prefer Cadillac less than on objective scales (if at all). Conversely, experts are less swayed by a single instance of exposure to expectations-disconfirming information and are likely to be more objectively accurate in their evaluations regardless of scale type. Accordingly, the SSE in brand evaluations should be most evident for novice consumers in the product category.

H2. Expertise moderates the SSE such that the evaluative shift on subjective relative to objective scales will be stronger for novice than for expert consumers.

The automatic nature of the SSE process

The next hypothesis involves the precise explanatory account and proposed automatic nature of the hypothesized shift. Tversky and Kahneman (1974) designed a now classic study wherein subjects span a wheel of fortune and were then required to estimate the percentage of African countries in the United Nations. Participants whose wheel spin had stopped on the number 65 gave a median estimate of 45%, whereas those whose wheel needle landed on number 10 gave a median estimate of 25%, an anchoring effect. The authors explained this occurrence in terms of subjects' anchoring of their judgments onto an external (albeit irrelevant) starting value and an insufficient subsequent adjustment away from it (Tversky & Kahneman, 1974). Unlike this external anchoring and adjustment, the SSE involves an automatic adjustment based on expectations, with no deliberate processing. If true, this automatic component of the SSE could be captured by implicit cognition measures. Thus, using the "A > B" scenario:

H3. Consumers will show weaker automatic associations with favorable attributes for brand B than for brand A. However, this difference will be less pronounced after responding to word-based rating scales than after responding to numerical-unit rating scales.

According to the SSE, we should be able to trace the shifting effect of brand expectations on consumer response. Accordingly, a pilot study first demonstrates that the standards shift occurs. Studies 1, 2a and 2b then address the effect's underlying processing mechanism in depth, demonstrating its automatic nature and evaluating the moderating role of consumer expertise and mediating role of expectations. Study 3 looks at the asymmetric nature of the SSE and finally study 4 predicts boundary conditions and successfully corrects for the effect.

Pilot study

The purpose of the pilot was to replicate the SSE in a marketing context that clearly makes use of individuals' expectations. A specific instance that entails consumer recourse to such brand expectations is the case of brand extensions in the marketplace. Substantial marketing literature has addressed this issue and found that consumer response to this type of brand management decision is generally constrained by what individuals perceive as categories where core brand benefits can credibly transfer (e.g., Kim & John, 2008). It is apparent that we have specific expectations as to what constitutes a likely brand extension, be that by Pepsi or Ford, and our evaluations of these new products are going to reflect these expectations. In evaluating hypothesis 1, the pilot study looks at possible differences between these evaluations, as guided by the type of response measurement employed.

Participants and design

The design was a mixed 2 (extension: peanuts or butter parent brand) × 2 (type of scale used to evaluate extension: numerical or word-based), with between-subject assignment to brand extension information and within-subject exposure to scale types (scale type exposure order was randomized). One hundred thirty eight undergraduate students participated in the experiment in return for credit in an introductory marketing class.

Method

The pilot thus involved two brands (Planters peanuts and Land O'Lakes butter) that were purportedly extending into the peanut butter category. A pretest confirmed that the two brands were equally familiar to our sample population ($M_{\text{Planters}}=4.19$, $M_{\text{Land O' Lakes}}=4.03$, $t(138)=1.66$, *ns*) and that Planters engenders expectations of a more relevant/appropriate parent brand relative to Land O'Lakes for a successful peanut butter brand extension (on a single item ranging from "definitely Land O'Lakes" = 1 to "definitely Planters" = 7, $M=4.87$, $SD=1.24$). Expertise in the product category was uniformly low (on a 7-point scale, $M=2.63$, $SD=.74$), allowing us to isolate the effect of interest in a largely novice sample population.

Participants were exposed to a one-page Industry News report in the *Progressive Grocer* magazine announcing the test marketing efforts of a new brand of peanut butter. The reports differed only in terms of the parent brand and this extension (Planters or Land O'Lakes peanut butter) was described overall as comparable with specialty, natural brands available at Whole Foods stores and superior to known brands such as Jif or Skippy.

After exposure to the product review (Planters or Land O'Lakes), participants provided their product ratings on both a numerical scale ("Please give this new product an overall evaluation between 1 and 10 points") and a word-based scale ("Please give this new product an overall evaluation on the following scale") anchored at "Would not like at all/Would like a lot." The order of the questions was counterbalanced and several filler questions about the review were always placed in between. Finally, participants were debriefed, thanked, and dismissed.

Results

Data show that the use of word-based judgment measures obscures consumers' numerical-unit evaluations. Given prior expectations of brand extension fit, a relevant brand (Planters) was judged in numerical units (a score ranging from 1 to 10) as more likely to succeed than a comparable but less relevant brand (Land O'Lakes): in a paired comparison test, $M_{\text{Planters}}=7.90$, $M_{\text{Land O' Lakes}}=7.26$, $t(137)=2.00$, $p<.05$ (in standardized *t*-scores, $M_{\text{Planters}}=51.64$, $M_{\text{Land O' Lakes}}=48.29$). However, when judged in word-based units, the two brand extensions emerged as equally viable in the same consumers' minds: $M_{\text{Planters}}=5.24$, $M_{\text{Land O' Lakes}}=5.41$, $t(137)=-1.05$, *ns* (in standardized *t*-scores, $M_{\text{Planters}}=49.13$, $M_{\text{Land O' Lakes}}=50.91$). Importantly, a contrast of participants' standardized scores across the two types of scales found no difference for Planters' evaluation but a significant difference in the case of Land

O'Lakes: M_{obj}=48.28, M_{subj}=50.91, $t(67)$=2.02, $p<.05$. As hypothesized, the SSE appears driven by the relaxing of evaluative standards and subsequent upward shift for the low-expectations brand, with no change for the high-expectations brand. H1 was thus supported.

Discussion

Previous research into proposed brand extensions has analyzed consumer evaluations of issues such as fit, relatedness, or congruity (e.g., Campbell & Goodstein, 2001). The current approach employs a new framework to show that in some cases consumer ratings of brand extensions may be influenced by another factor (i.e., scale type). Importantly, the pilot suggests that the effect is strong enough as to emerge in a within-subject design, such that individuals appear to alter their own brand evaluations depending on the type of scale employed.

Study 1

Having established the effect of interest, study 1 is designed to address the proposed role of consumer expertise in the SSE. The study also addresses some issues the pilot left unresolved. First, in study 1 we use interval scaled numerical units (e.g., pixels and seconds) as opposed to simple score points to capture objective judgments and thus make for more directly comparable objective/subjective evaluations. Second, the scales in the pilot involved a single item and thus showed that holistic evaluations are subject to the SSE. Study 1 also employs matched objective and subjective items referring to specific product attributes, to see whether they are subject to the same scale-based discrepancy as overall evaluations.

Participants and design

Two hundred thirteen undergraduate students participated in the experiment in return for credit in an introductory marketing class. Participants read a product review purportedly describing either Sony or RCA and had to subsequently provide associated ratings on either numerical-unit or word-based scales. The design was a 2 (brand: Sony or RCA)×2 (scale type: numerical or word-based)×2 (expertise: low or high) between subjects factorial.

Method

The stimuli employed Sony and RCA branded DVD players pretested to elicit different expectations (i.e., higher for the former: M_{Sony}=83.40 on a 100-point performance scale, and lower for the latter: M_{RCA}=64.02, $t(140)$=13.01, $p<.001$). Prior to the study, participants self-reported their level of knowledge in the DVD product category on a 6-point scale ranging from "know nothing" to "know everything." After separating respondents who answered 1–3 from those in the 4–6 group, the sample included 132 novices and 81 experts.

In the main experiment, participants were randomly exposed to a one-page *Consumer Reports* editorial review of a particular new RCA or Sony DVD player model (only brand name differed). The product description involved various attributes pretested to be of reasonably high interest to our sample population (e.g., MP3 playing-ability, picture adjustment and progressive scanning options, etc.). The tone of the review was relatively neutral and presented the new player as a competitive entrant in the DVD segment; while the message highlighted some of the product's qualities, it did not suggest it was exceptional in any way.

After exposure to the product review, participants in the numerical-unit response condition provided their product ratings on one holistic numerical scale ("Please give this product an overall evaluation score between 1 and 7 points") as well as 3 other objective items requiring them to estimate the player's picture resolution (in dots per inch—dpi—between 750 and 1500), its number of audio/video inputs and outputs (between 1 and 7 connections), and the buffer length of its progressive scan (between 1 and 15 s; lower length describes better performance).[1] Word-based response participants were queried across a word-based holistic item anchored at "Disliked a lot/Liked a lot," as well as 3 matched attribute items requiring them to estimate the DVD player's picture resolution (very poor/very good), number of audio/video outputs (very few/very numerous), and progressive buffer length (very short/very long).

Finally, we were also interested in addressing the claim that the adjustment process in the SSE involves unconscious, as opposed to conscious resources. To this end, we gave participants in the word-based conditions identical feedback on their response patterns, according to which they had supposedly shown a bias in favor of the evaluated brand. We measured participants' level of agreement with this fictitious assessment on a 7-point scale anchored at "Completely Disagree/Completely Agree." Our hypothesis was that participants would universally disagree, suggesting that the adjustment is likely unconscious.

Results

Numerical-unit evaluations were in line with consumers' prior expectations: an ANOVA on respondents' overall product evaluation with brand, scale type, and consumer expertise as factors uncovered a main effect of brand, such that Sony was perceived to have an overall better product than RCA, despite exposure to the very same review: M_{Sony}=5.77, M_{RCA}=5.44, $F(1, 212)$=7.96, $p<.01$. There was also an unexpected but only marginally significant effect of scale type (M_{obj}=5.45, M_{subj}=5.74, $F(1, 212)$=3.16, $p<.08$), suggesting that the subjective scale generally induced more favorable responses. Further analyses revealed that this effect was mainly driven by novice consumers' significant upwards shift in subjective relative to objective ratings of RCA (M_{obj}=5.29, M_{subj}=5.76, $t(74)$=2.45, $p<.02$). Whereas novices perceived the RCA player to be significantly less performing than the Sony on the

[1] The ranges and means of these objective scales were pretested with a comparable sample (N = 38: M_{res} = 1161dpi, M_{con} = 4.2, M_{buf} = 7.3s).

objective scale ($M_{Sony}=5.75$, $M_{RCA}=5.29$, $t(60)=2.06$, $p<.05$), this gap was eliminated on the subjective scale ($M_{Sony}=5.81$, $M_{RCA}=5.76$, $t(68)=-.23$, *ns*).

For both the holistic and the 3 attribute items, experts did not show any significant effects and only slightly favored Sony each time and for both types of scales. Novice participants however perceived Sony to have a better picture resolution than RCA when measured numerically in dots-per-inch ($M_{Sony}=1285.42$, $M_{RCA}=1162.66$ dpi, $t(60)=2.54$, $p<.02$) but not on the word-based scale ($M_{Sony}=6.03$, $M_{RCA}=5.74$, $t(68)=1.68$, *ns*); more audio/video connections when measured objectively ($M_{Sony}=5.67$, $M_{RCA}=4.68$ in/outputs, $t(60)=2.13$, $p<.04$) but not subjectively ($M_{Sony}=5.00$, $M_{RCA}=4.87$, $t(68)=.64$, *ns*); and a shorter progressive scanner buffer measured in s ($M_{Sony}=7.67$, $M_{RCA}=9.47$ s, $t(60)=-2.76$, $p<.01$) but not in words ($M_{Sony}=5.06$, $M_{RCA}=4.87$, $t(68)=.65$, *ns*). H1 was thus supported.

Participants' level of agreement with the fictitious assessment that they had favored the reviewed product more than it deserved was independent of the type of scale: $M_{obj}=4.63$, $M_{subj}=4.49$, $F(1, 212)=.73$, *ns*. Consistent with an unconscious processing mechanism, the evaluative standard shift did not seem to be explicitly salient to the novice participants in the subjective scale RCA condition. In fact, they were the least agreeing with the statement out of all the experimental cells. Interestingly, however, a main effect of brand was observed on this item, such that in general participants agreed more with being unjustly biased toward Sony rather than RCA ($M_{Sony}=4.79$, $M_{RCA}=4.37$, $F(1, 212)=6.38$, $p<.02$), in apparent acknowledgment that much of the message processing relied on prior expectations.[2]

Finally, we asked respondents how happy they were with their current DVD player. No significant differences emerged for experts, who were uniformly happy with their home player (86% reported being happy, 14% reported being unhappy), regardless of the brand or evaluative scale they were exposed to. However, although expressing the same overall feelings (84% happy, 16% unhappy), novice consumers showed a different pattern by cell. A binary logistic regression with satisfaction as the variable predicted by brand, scale type, and their interaction term uncovered a main effect of scale ($B=-4.20$, $p<.04$) as well as a significant interaction ($B=2.63$, $p<.04$), both driven by the responses in the subjective scale RCA evaluation condition. These participants were the least happy with their current DVD player (30% were unhappy), suggesting that the contrast provided by the evaluated player was more favorable than expected—indirect evidence for the SSE.

Discussion

Study 1 complements the pilot, replicating the SSE in a different consumer context and showing the impact of different brand expectations on word-based scales. First, the fact that identical *Consumer Reports* reviews for the two brands lead to different numerical scores suggests that brand expectations for the two manufacturers are objectively different, for both expert and novice consumers. Sony is generally perceived to be superior to RCA, and this fact significantly impacts the effectiveness of the review. Second and more interestingly, this difference is reduced (in our case obliterated) for novices on word-based scales. These respondents seem to have adjusted for their expectations, by shifting their standards of evaluation for the inferior brand. This shift is, we argue, precisely the mechanism described by the SSE model: an implicit correction for prior expectations.

Study 1 also showed that the proposed SSE account involves a likely unconscious processing mechanism, as participants were uniformly unwilling to consciously acknowledge the shift in evaluative standards when informed about it post hoc. Studies 2a and 2b will present more evidence for the automatic nature of the effect. That said, recent research by Ofir and Simonson (2007) finds that bringing prior expectations to consumers' awareness *prior* to their providing satisfaction ratings leads to a focus on the negatives of the brand (especially in cases of low prior expectations) and subsequent cognitive corrections that are likely to nullify the SSE. Study 4 will directly test a related way to control for the shifting effect on word-based scales.

Study 2a

Having established the SSE in a branding context and the moderating role of consumer expertise therein (i.e., hypotheses 1 and 2), the next experiment was designed to more closely assess the cognitive mechanism proposed to underlie the effect: the expectation-driven shift in inferior brand ratings for novices on subjective scales. We also expanded the range of consumer decisions under investigation to financial products, in particular evaluations of earnings-per-share for specific technology stocks.

Participants and design

One hundred and fifty-one undergraduate students participated in the experiment in return for credit in an introductory marketing class. Participants read a financial news report describing the supposedly imminent introduction to the marketplace of an innovative web browser by either Yahoo! or Google and the potential impact of this launch on the respective firm's financials (details on the company's previous quarter financials were provided). Respondents had to subsequently provide estimates of the likely earning-per-share announcement by the respective stock in the upcoming quarter on either numerical-unit or word-based scales. The design was a 2 (brand: Google or Yahoo!) × 2 (scale type: numerical or word-based) × 2 (expertise: low or high, measured) between subjects factorial.

Method

The stimuli employed Google and Yahoo! financial stocks pretested to elicit different performance expectations (i.e., higher for the former: $M_{Google}=5.87$, and lower for the latter: $M_{Yahoo!}=$

[2] When using the brand owned as a covariate, results did not change.

4.42, $t(150)=12.33$, $p<.001$). Prior to the study, participants self-reported their level of knowledge in the financial products category on a 6-point scale ranging from "know nothing" to "know everything." After separating respondents who answered 1–3 from those in the 4–6 group, the sample included 98 novices and 53 experts. As an additional check, we also looked at participants' (all business school students) study major—coded as *Finance* or *Other*. As expected, the two measures were highly associated ($\chi^2=28.56$, $p<.001$).

After exposure to the stock review, participants in the numerical-unit response condition provided their estimates on one numerical scale ("Please estimate the earnings-per-share number that Google/Yahoo! is likely to report at the end of the financial quarter") between $.01 and $5.00 (participants were allowed to provide both dollars and cents in their answers). Word-based response participants were similarly queried across a single word-based item anchored at "Very low/Very high shareholder profitability."

To evaluate the role of stock-related expectations, participants were asked in the end to provide the extent to which they felt that the reviewed stock had met or not their personal prior expectations (1=did not meet at all, 7=completely surpassed). Participants also had to self-gauge the amount of influence that this [mis]match with expectations had had on their EPS estimates (1=did not influence me at all, 7=completely influenced me).

Results

The SSE emerged as in previous studies. Experts favored Google over Yahoo! overall ($t(51)=3.86$, $p<.001$) and this main effect was not qualified by scale type. For novices, numerical-unit evaluations also confirmed pretests, such that Google was judged in dollar amounts ranging from $0 to $5 as likely to present higher earnings-per-share than Yahoo!: in a paired comparison test, $M_{Google}=3.51$, $M_{Yahoo!}=2.32$, $t(48)=3.28$, $p<.01$ (in standardized *t*-scores, $M_{Google}=53.80$, $M_{Yahoo!}=45.42$). However, when judged in word-based units, the two stocks were perceived as equally profitable for their shareholders by our novice respondents: $M_{Google}=5.04$, $M_{Yahoo!}=4.80$, $t(46)<1$, *ns* (in standardized *t*-scores, $M_{Google}=51.83$, $M_{Yahoo!}=49.98$). The contrast of participants' standardized scores across the two types of scales found no difference for Google's stock evaluations but a significant difference in the case of Yahoo!: $M_{obj}=45.42$, $M_{subj}=49.98$, $t(50)=2.10$, $p<.04$.

To address the role of expectations, an ANOVA on participants' rating of the extent to which they felt that the reviewed stock had met or not their prior expectations uncovered a main effect of expertise: $M_{novice}=4.01$, $M_{expert}=4.34$, $F(1, 150)=3.78$, $p<.05$. Overall, experts were relatively more positively surprised by the reviewed stock, likely due to their higher ability to accurately judge the somewhat complex review information. In line with our hypothesis, the analysis also uncovered a significant interaction of expertise and scale type: $F(1, 150)=6.50$, $p<.01$. Critical to our account, planned contrasts revealed that—relative to their Yahoo! ratings on the objective scale—novices' responses on the subjective Yahoo! rating scale showed more surprise: $M_{obj}=3.63$, $M_{subj}=4.40$, $t(50)=-2.77$, $p<.01$. Notably however, this enhanced mismatch with expectations was not acknowledged by the same respondents as having influenced their stock rating ($M_{obj}=3.89$, $M_{subj}=3.84$, $t(50)<1$, *ns*).

To further pursue this mismatch with expectations for novices evaluating Yahoo! stock, a mediation analysis (Baron & Kenny, 1986) was pursued. First, we used regression to establish that the rating scale was significantly related to earnings estimates: $B=4.56$ ($SE=2.18$), $t=2.10$, $p<.04$. Second, a linear regression confirmed that the rating scale predicted the mismatch with expectations: $B=.77$ ($SE=.27$), $t=2.77$, $p<.01$. Third, this mismatch was a significant predictor of earnings: $B=2.70$ ($SE=1.00$), $t=2.69$, $p<.01$. Finally, in order to examine whether mismatch with expectations mediates the effect of rating scale on earnings estimates we simultaneously entered both into the regression. Expectations remained predictive ($B=2.21$, $t=2.06$, $p<.05$), whereas the effect of rating scale was rendered insignificant ($B=2.86$, $t=1.26$, *ns*). A Sobel test of the mediating effect was significant, $z=1.96$, $p<.05$ indicating that expectations mediate the effect of rating scale on earnings estimates for novices evaluating Yahoo! stock. Importantly, similar mediating effects were not observed for novices evaluating Google or for experts.

Discussion

Study 2a replicates the SSE in a novel context and finds renewed support for the moderating role of consumer expertise. Further, it demonstrates that novices experience more of a mismatch with expectations (in favor of the inferior brand) after rating it on subjective scales, in line with the proposed explanatory mechanism of the SSE. Importantly, consumers who report having been more positively surprised also claim that this surprise did not play a major role in their subjective rating. Although this could arguably suggest a social desirability motive (i.e., fear of appearing a less-than-sophisticated consumer by overweighting the surprise factor), we note that the expectations mismatch item found similar levels of disagreement with this statement ($M=3.86$) across cells, suggesting that it is simply not a salient or consciously considered factor in judgments. Furthermore, our entire sample showed relatively low levels of social desirability responding on Greenwald and Satow's (1970) scale. Faced with increasing evidence for the unconscious nature of the SSE, we next attempt to directly measure it.

Study 2b

Study 2b was designed to address hypothesis 3 and directly assess the automatic processing component behind the SSE. The hypothesis posits that consumers responding to numerical-unit scales would show implicit responses more consistent with their explicit expectations than consumers responding to word-based rating scales. In other words, consumers' implicit attitudes would closely parallel their explicit expectations in the former case, but not in the latter, where less consistency will

be present due to their automatic, internal shifting of evaluative standards. To isolate the effect of interest, we employed a product category of universally low expertise in our sample population.

The Implicit Association Test (IAT—Greenwald, McGhee, & Schwartz, 1998) was employed to directly test the automatic account. In the IAT, a subject responds to a series of items that are to be classified into four categories—typically, two representing a concept discrimination such as *music* versus *noise* and two representing an attribute discrimination such as *pleasant* versus *unpleasant* valence. Subjects are asked to respond rapidly with a right-hand key press to items representing one concept or one attribute (e.g., *noise* or *pleasant*), and with a left-hand key press to items from the remaining two categories (e.g., *music* or *unpleasant*). Subjects then perform a second task in which the key assignment for one of the pairs is switched (such that *music* and *pleasant* share a response key, likewise *noise* and *unpleasant*). The IAT produces measures derived from latencies of responses to these two tasks. These measures are interpreted in terms of association strengths by assuming that subjects respond more rapidly when the concept and attribute mapped onto the same response key are strongly associated (e.g., *music* and *pleasant*) than when they are weakly associated (e.g., *noise* and *pleasant*). In our case, the IAT shows the degree to which one brand is implicitly favored over the other. Our hypothesis is that the strength of this automatic preference for the superior brand will be attenuated for respondents primed by word-based scales.

Participants and design

The design was a simple 2 (brand: Michelin or Firestone) × 2 (scale type: numerical-unit or word-based) between subjects factorial. One hundred and ten undergraduate students participated in the experiment in return for credit in an introductory marketing class. As established in previous research, Michelin is perceived to be a high safety brand (cf. Lane, 2000), whereas Firestone is expected to provide less safety (a pretest showed that out of 9 tire brands, the mean safety rankings of the two brands were significantly different: $M_{\text{Michelin}}=3.00$, $M_{\text{Firestone}}=4.96$, Wilcoxon Test $Z=-5.00$, $p<.001$).

Method

Participants were exposed to a one-page editorial material describing the efforts of a specific tire manufacturer to develop and introduce safer new products. Message content was identical across the two brands. After exposure to this material, participants provided their brand ratings on either a numerical scale ("Please give these tires an overall safety score between 1 and 7 points") or a word-based scale. The latter was anchored at extremely poor/extremely good safety performance. Note that although these anchors seem more extreme than in previous studies, they are appropriate in magnitude for the attribute under investigation (i.e., it is difficult to overstate the danger that a tire scoring 1 on a 7-point scale poses).

Finally, participants performed an IAT assessing their implicit associations between the two focal brands (represented by brand logos) and attributes describing the concepts of *safe* (e.g., secure, protected, etc.) or *unsafe* (e.g., vulnerable, exposed, etc.).

Results

Numerical-unit evaluations confirmed pretested expectations of Michelin as a safer tire brand than Firestone: $M_{\text{Michelin}}=5.73$, $M_{\text{Firestone}}=4.52$, $t(41)=3.39$, $p<.01$. In terms of word-based evaluations, the discrepancy between the two brands' evaluations was reduced as predicted by the SSE and H1: $M_{\text{Michelin}}=5.81$, $M_{\text{Firestone}}=5.43$, $t(67)=1.66$, *ns*. As predicted, an ANOVA on respondents' log-transformed IAT response times found a main effect of brand, such that overall Firestone was less strongly associated with *safe* than Michelin ($F(1, 102)=7.74$, $p<.01$). The interaction of brand and scale type was only marginally significant ($F(1, 102)=3.45$, $p<.06$). Planned contrasts on participants' IAT results in the Firestone conditions found that they more strongly associated Firestone and *unsafe* implicitly after being exposed to the numerical-unit scale than after exposure to the word-based scale ($M_{\text{Firestone-Numerical}}=.003$, $M_{\text{Firestone-Word}}=-.06$, $t(52)=1.82$, $p<.07$).[3] No differences emerged for participants in the Michelin conditions ($M_{\text{Michelin-Numerical}}=.03$, $M_{\text{Michelin-Word}}=.05$, $t(47)<1$, *ns*). These results provide support for H3.

Discussion

The hypothesized automatic nature of the SSE-based evaluative standard shift was directly captured in the IAT results. The automatic activation of stereotypical perceptions produced different expectations of safety performance for two familiar tire brands. Accordingly, Michelin was relatively more strongly associated implicitly with *safe* than Firestone, but this difference was eliminated after exposure to the word-based scale. As respondents consciously deny relaxing their evaluative standards for the weaker brand, the IAT results suggest that the effect is unconscious.

Study 3

Having established the processing mechanism underlying the SSE, the next study is designed to take a closer look at the emerging evidence for the asymmetric nature of the effect. The previous experiments have employed product information that was moderately incongruent with the inferior brand, allowing for its associated low expectations to be surpassed. At the same time, an argument could be made that the description was largely positive and not discrepant enough for the high expectations associated with the superior brand to be let down. The issue goes back to the point made earlier: whereas

[3] Based on the setup of the specific IAT task, higher log-transformed values suggest a stronger automatic association of Michelin and *safe*, whereas lower values describe a stronger Firestone-*safe* association.

better-than-expected performance is relatively easy to notice, it is somewhat tougher to realize that a brand is not doing as well as it could (in particular for novices). Calibrating a product review at the exact middle point of the performance interval anchored by the two brands is admittedly a difficult task, and this study manipulates the levels of discrepancy to directly evaluate the asymmetric nature of the SSE.

Participants and design

The design was a 2 (brand: Acer or Sony)×2 (scale type: numerical-unit or word-based)×2 (expertise: low or high)×2 (performance level: low or high) between subjects factorial. Two hundred and ten undergraduate students participated in the experiment in return for credit in an introductory marketing class. A pretest confirmed that Sony notebooks are associated with higher expectations than Acer notebooks (out of 12 brands, the mean performance rankings were $M_{Sony}=2.26$, $M_{Acer}=9.66$, Wilcoxon Test $Z=-11.59$, $p<.001$) and allowed us to capture consumer perceptions of the average notebook in the market across several attributes (screen resolution: $M_{res}=1280\times800$; number of ports: $M_{port}=4.04$; hard drive capacity: $M_{drive}=54.8$ GB; processor speed: $M_{proc}=1.74$ GHz).

Method

Participants were randomly exposed to a supposed one-page Cnet.com editorial review of a particular new Acer or Sony notebook model pretested to describe either high or low performance. The product description involved various attributes pretested to be of reasonably high interest to our sample population (e.g., screen size and resolution, USB ports, battery life, etc.). The tone of the review was more favorable for the high review condition and overall presented the new model as a new entrant in the personal notebook segment; while the messages highlighted some of the product's qualities but not others, neither version explicitly stated it was exceptional or poor in any way. At the bottom of the review, the specifics of the tested model were provided, under the assumption that these will be guiding the response of experts, whereas the more qualitative body of the review will be relied on by novice consumers (cf. Maheswaran & Sternthal, 1990).[4]

After exposure to the product review, participants in the numerical-unit response condition provided their product ratings on one holistic scale ("Please give this product an overall evaluation score between 1 and 5 points") as well as 4 other objective items requiring them to estimate the notebook's screen resolution (800×600–1024×768–1280×800–1440×900–1680×1050 pixels), number of ports (1–2–4–6–8 ports), drive capacity (5–20–40–80–100 GB), and processor speed (0.7–1.1–1.5–2.0–2.4 GHz). Word-based response participants were queried across a word-based holistic item anchored at "Very Poor/Very Good," as well as 4 matched 5-point attribute items requiring them to estimate screen resolution (very poor/good), number of ports (very few/numerous), hard drive size (very limited/extensive), and processor speed (very slow/fast). We also recorded participants' response latencies for the holistic evaluation items.

As in study 1, we again gave participants in the word-based conditions identical feedback on their response patterns according to which they had supposedly shown a bias in favor of the evaluated brand. We measured participants' level of agreement with this fictitious assessment.

Results

Self-reported expertise in the product category (6-point scale) was dichotomized to describe a novice group (answers of 1–3, $N=126$) and an expert group (answers of 4–6, $N=84$). As a manipulation check, we analyzed two items that asked respondents to provide the part of the review that they had paid most attention to when reading and had been most influenced by (where review body=benefits and specs=attributes). In line with Maheswaran and Sternthal (1990), 77% of the novices mentioned the body of the message as the focal point of their attention and 54% of the experts referred to the notebook specs ($\chi^2=20.62$, $p<.001$), while similarly 72% of the novices reported being mostly influenced by the body of the review and 60% of the experts by the specs ($\chi^2=16.85$, $p<.001$).

A manipulation check of review levels found that, as expected, the low and high reviews were perceived as significantly different: $M_{high}=4.30$, $M_{low}=3.86$, $t(208)=4.01$, $p<.001$.

An ANOVA on participants' holistic evaluations with brand, scale type, review level, and expertise as factors uncovered a main effect of brand ($F(1, 209)=11.81$, $p<.001$), a main effect of review level ($F(1, 209)=17.25$, $p<.001$), and a significant interaction of brand and expertise ($F(1, 209)=4.89$, $p<.03$). To facilitate comparison with the previous studies and to directly evaluate the proposed asymmetric effect of the SSE across novices and experts, further analyses were pursued for each of the review levels.

When looking at the effects for the high review conditions, they largely replicate the findings from previous studies. An ANOVA on participants' holistic evaluations with brand, scale type, and expertise as factors uncovered a main effect of brand ($F(1, 106)=4.60$, $p<.04$). Planned contrasts on novice consumers' scores in the objective scale condition uncovered the expected stereotypical perception of Sony as superior to Acer ($M_{Sony}=4.67$, $M_{Acer}=3.82$, $t(24)=2.19$, $p<.04$), but this difference went away on subjective scales ($M_{Sony}=4.42$, $M_{Acer}=4.26$, $t(36)<1$, *ns*). Similar results emerged for each of the 4 specific attributes of the notebook (see Table 1 for specific means). Interestingly, these novices' level of agreement with the fictitious statement that they had unduly favored Acer did not vary with the type of scale ($M_{obj}=4.45$, $M_{subj}=4.42$, $t(28)<1$, *ns*), suggesting that they were unaware of the shift occurring on the subjective scale. No significant effects occurred for experts in the high review conditions, as they slightly favored Sony for both types of scales.

[4] We felt specific attributes had to be listed, lest experts disbelieve and discount the low review altogether.

Table 1
Means and standard errors for main dependent variables in study 3

	Numerical-unit scale rating						Word-based scale rating					
	Overall		Hard drive		Processor		Overall		Hard drive		Processor	
	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Low review—novices												
Acer	3.27[a]	.19	3.00[b]	.36	3.09[c]	.31	3.62[d]	.22	3.44[e]	.18	3.50[f]	.18
Sony	4.15[a]	.29	4.00[b]	.30	4.08[c]	.26	4.27[d]	.13	3.91[e]	.13	4.05[f]	.11
Low review—experts												
Acer	3.80	.20	3.10	.40	3.40	.22	3.70	.30	3.20	.25	3.50	.17
Sony	3.83	.12	3.17	.24	3.50	.23	3.80	.13	3.30	.30	3.60	.27
High review—novices												
Acer	3.82[g]	.29	3.18[h]	.26	3.36[i]	.20	4.26	.15	3.68	.13	3.95	.14
Sony	4.67[g]	.25	4.00[h]	.24	4.07[i]	.23	4.42	.19	3.84	.14	3.79	.13
High review—experts												
Acer	4.10	.28	3.60	.27	3.80	.25	4.20	.29	3.50	.17	3.90	.23
Sony	4.30	.21	4.00	.21	4.10	.23	4.42	.19	3.75	.25	4.00	.17

All same superscripts represent contrasts significant at $p<.05$ or lower.

In the low review conditions, an ANOVA on participants' overall evaluations with brand, scale type, and expertise as factors uncovered a main effect of brand ($F(1, 104)=7.60$, $p<.01$) and a significant interaction of brand and expertise ($F(1, 104)=5.36$, $p<.03$). Planned contrasts on novice consumers' scores in the objective scale condition uncovered the same stereotypical perception of Sony as superior to Acer ($M_{Sony}=4.15$, $M_{Acer}=3.27$, $t(22)=2.39$, $p<.03$), but at lower levels than for the high review. The effect remained significant on subjective scales ($M_{Sony}=4.27$, $M_{Acer}=3.62$, $t(36)=2.64$, $p<.02$), suggesting the low review was in line with low prior Acer expectations. Similar results emerged for each of the 4 specific attributes of the notebook (see Table 1 for specific means). Once again, no significant effects occurred for experts, who perceived the two brands to be comparable for both types of scales, suggesting that they indeed were less swayed by brand stereotypes and correctly processed the attribute-level information.

Finally, a main effect of scale type was also observed on participants' response latency (in ms) for their overall product evaluation, such that responses on the numeric-unit scale were slower than those on the word-based scale, regardless of consumer type, brand, or review level: $M_{obj}=4932.28$, $M_{subj}=3216.55$, $F(1, 209)=55.77$, $p<.001$. First, this suggests that dealing with numeric units is a naturally more complex task. Second, it shows that the shift induced in novices' cognitive response (i.e., the "for an Acer…" qualification) does not involve extra processing time and thus represents another argument for the automatic nature of the SSE.

Discussion

Study 3 results offer renewed support for the SSE among novice consumers, but do not find evidence of a symmetric effect for experts. Even a low review that is discrepant with experts' expectations does not produce immediate derogation of the dominant brand. Although in terms of both types of scales experts evaluate the dominant brand relatively low after a low review, their open thoughts suggest that they often encode the reviewed model as an outlier in the dominant brand's line of products (there is little expressed disbelief—after all, the low-level attributes are printed right after the review).

It could be argued that for numerical-unit condition experts (who presumably focused on the bottom-of-review specs) evaluations were in a sense a simple memory task. However, note that the effect was identical for the few experts who acknowledged focusing on and being influenced by the body of the message. This suggests that when exposed to qualitative, benefit-focused reviews such as in the previous studies, experts tend to disregard the information and revert back to their own attribute-level personal knowledge, thus avoiding the SSE.

Novices exposed to the high review replicated the SSE as in the previous experiments, even though attribute information was available and could have been used in judgments. This suggests that these consumers either ignore attribute-level information (although roughly 30% of them said they were influenced by it) or—more likely although not acknowledged—are simply not informed enough to interpret it correctly. By contrast, novices exposed to a low review for the dominated brand do not have their expectations surpassed and thus respond according to prevalent stereotypes on both types of scales (thus favoring the dominating brand each time).

Study 4

The previous experiments have addressed the SSE as a particular cognitive phenomenon directly relevant to marketers (via consumers' brand expectations). Study 4 is designed to underscore the importance of the measurement choices made by marketers (practitioners and academics alike), by demonstrating how the SSE can be predicted and controlled. To try to forestall the automatic shift, one might proactively bring this implicit cognitive bias to consumers' attention (Morwitz & Fitzsimons,

2004). One way to operationalize this is by framing word-based scale items as conditional statements (e.g., "Given your prior knowledge of the brand and your expectations, please rate the performance of this brand from poor to outstanding"). Such manipulation should make the subject explicitly aware of prior expectations and thus less likely to automatically shift evaluative standards. The scores on these word-based scales would then be consistent with the brand ratings captured via numerical scales.

Note that an alternative prediction in this case would be that the SSE replicates even stronger when brought into consciousness. To the extent that people "cut slack" to inferior brands without awareness, shouldn't they be even more (or at least equally) inclined to do so when made aware of it? Two arguments go against this intuition: first, studies 1, 2a, and 2b have shown that, in general, consumers do not [like to] acknowledge the unconscious impact of the SSE when presented as a hypothetical explanation of their behavior. Second, when made aware of this potential bias, consumers are essentially faced with a choice: relying on their strong prior expectations (suggesting that they do not "cut slack") or accepting the novel incoming information (suggesting that they do). We believe that, all things considered, consumers are unlikely to update their beliefs as quickly as following a single instance of expectation-disconfirming information. Accordingly, they will revert to their extant knowledge and will respond to subjective scales just as they would to objective ones, in effect canceling the SSE.

Participants and design

One hundred thirty five undergraduate students participated in the experiment in return for credit in an introductory marketing class. To focus on the effect of interest, this study only involved novices in the product category (selected in a pretest 3 weeks before the main experiment) and was limited to a single overall evaluation. All participants read a product review purportedly describing either a Sony or RCA DVD player and had to subsequently provide associated ratings on either numerical-unit or word-based scales. This time, two types of word-based scales were used: direct scale items (such as those employed in the previous studies) or the same items framed as conditional statements. The design was thus a mixed 2 (brand: Sony or RCA, between subjects) × 2 (scale type: numerical and word-based, within subjects) × 2 (word-based scale type: direct or conditional, between subjects).

Method

Participants were randomly exposed to a one-page *Consumer Reports* editorial review of a particular new RCA or Sony DVD player model as in study 1. After exposure to the product review, participants provided their product ratings on one numerical item ("Please give this product an overall evaluation score between 1 and 7") and a word-based scale. The latter was comprised of four 7-point items framed as either direct questions (e.g., "How would you rate the performance of this Sony model?") or conditional questions (e.g., "Given your prior knowledge of Sony, how would you rate the performance of this model?"). The word-based scale items were anchored at extremely poor/extremely good performance, far behind/far ahead of market trends, no/all desirable features, and would definitely avoid/consider, respectively ($\alpha_{RCA\text{-}d}=.87$, $\alpha_{Sony\text{-}d}=.78$; $\alpha_{RCA\text{-}c}=.91$, $\alpha_{Sony\text{-}c}=.90$).

Results

The ANOVA associated with our mixed design (scale type as within-subject and brand and word-based scale variety as between-subjects factors) found the predicted three-way interaction on product ratings: $F(1, 131)=10.82$, $p<.001$. In particular, numerical evaluations highlighted the expectation that Sony was a superior brand to RCA: $M_{Sony}=5.89$, $M_{RCA}=5.55$, $t(133)=2.47$, $p<.02$. In terms of word-based evaluations, a two-way ANOVA with brand and word-based scale variety as factors found a main effect for each (for brand, $F(1, 134)=11.10$, $p<.001$; for word-based scale variety, $F(1, 134)=3.93$, $p<.05$). Planned contrasts revealed no difference between the evaluations of Sony and RCA on word-based scales framed as direct questions ($M_{Sony}=5.79$, $M_{RCA}=5.51$, $t(69)=1.64$, $p=.11$) but a significant difference ($M_{Sony}=5.69$, $M_{RCA}=5.06$, $t(62)=2.90$, $p<.01$) on word-based scales framed as conditional questions (replicating the objective scale results and eliminating the SSE). Evoking the brand expectations prior to the ratings thus canceled the SSE-based shift.

Discussion

Study 4 suggests that once consumers are asked to consciously retrieve their brand expectations, they will avoid the SSE. In effect, this instruction is a de facto discounting cue that cautions consumers about being overly enthusiastic toward a brand with a history (captured via expectations) of relatively low performance. The numerical-unit scores are thus more likely to be replicated, without the attenuation of differences that non-conditional word-based ratings induce. Calling expectations to individuals' attention can be done by phrasing the rating item as prior knowledge-based. The effect of the conditional statement is thus to forestall an automatic process, and instead activate a conscious counter-adjustment mechanism.

General discussion

At the core of consumer marketing lies the intent to differentiate the company or its brands from similar offerings in the marketplace on one or more evaluative dimensions. Our results suggest that a common but unconscious consumer cognitive response to brand information may significantly impact the measured differentiation. Whereas more objective, numerical-unit scores would indicate a sustainable differentiation between two brands along a particular attribute, more subjective, word-based responses may in fact show the very same brands to be virtually indistinguishable. This anomalous inconsistency was shown to originate in consumers' use of different evaluative standards for high and low expectation brands. In word-based judgments, evaluative standards are

automatically more relaxed for the brand associated with lower expectations, allowing it to match its competitor's advantage. A pilot study established the effect in the context of expectations related to brand extension fit. Study 1 evaluated the moderating role of consumer expertise and confirmed that novices are more likely to exhibit the judgment biases induced by the SSE. Studies 2a and 2b uncovered the mediating role of brand expectations and the automatic nature of the shift by actually measuring the automatic adjustment of standards (i.e., comparative anchors) for word-based scales via the IAT. Study 3 found that the upward shift in evaluations for brands engendering low expectations is not accompanied by downward shifts for high-expectation brands. Finally, study 4 eliminated the SSE by bringing this cognitive heuristic into individuals' awareness.

The effect was stronger for novice consumers in a product category (in our studies, DVD players, peanut butter, tires, financial stocks, and notebook computers) and largely absent for experts. As proposed by Maheswaran and Sternthal (1990), experts look for attribute information (see Study 3) and are able to make objective judgments regardless of brand or scale type. When such information is not provided, they tend to discount qualitative, benefit-focused information and instead revert to own knowledge-based attribute information.

We also presented results that suggested a way to prevent novice consumers from making these automatic adjustments. By framing the word-based scales explicitly against respondents' expectations, we showed that it is possible to prevent the unconscious shift and to obtain the same ratings as for the numerical scales. In a sense, this manipulation and objective scales in general appear to induce what cognitive psychology terms a confirmation bias, whereas subjective scales avoid this bias and instead induce another (i.e., the SSE).

Beyond uncovering the moderating role of expertise, the present research extends the work of Biernat et al. (1991) by demonstrating that—at least in a brand evaluation context—the SSE is asymmetric, and consumers (including experts) do not appear to place brands that engender high expectations under more stringent evaluations. This suggests that when faced with expectation-incongruent information, consumers exhibit a contrast effect away from their low performance stereotype for inferior brands, whereas they show an assimilation effect to their high performance stereotype for superior brands. In a sense, consumers appear predisposed to "cut slack" to reviewed brands in each case. To get at the cognitive nature of the effect, study 2a showed that subjective scales prompt novices to perceive a more positive mismatch with expectations that mediates their brand ratings and study 3 found that when these novices own a competing product they claim to be less happy with it. Although subjects will not consciously acknowledge the effect, its automatic nature is captured in study 2a via the IAT and in study 3 by finding lower latencies for responses to subjective scales.

Implications

The present work raises questions about the validity of inferences made when evaluating consumer attitudes toward well-known brands (i.e., those engendering strong expectations) along word-based scales. Product positioning maps are typically developed on the basis of word-based responses on salient attributes for an evoked set of competing brands. The resulting maps might look quite different if the marketer employed numerical-unit rating scales.

In the area of brand extensions, our results suggest that the observed levels of fit or congruity might vary depending on the measurement used. Expectations associated with the parent brand are typically seen to place limits on extensions to new product categories (e.g., Campbell & Goodstein, 2001). Research by Lane (2000) appears to contradict this work on brand extensions by showing relatively positive consumer response to incongruous extensions after repeated ad exposures. However, the Lane (2000) design makes use of word-based scales, which may have in fact artificially bolstered consumer reactions to these incongruous extensions.

The bias in word-based ratings may also explain some anomalies in published product ratings. For example, *Consumer Reports* automobile owner satisfaction ratings (collected via word-based scales anchored at *extremely dissatisfied* and *extremely satisfied*) do not appear to track well with reliability ratings measured objectively via items such as repair occasions/costs. For example, the 2006 Ford Mustang is both one of the most satisfying and one of the least reliable sports cars (Consumer Reports and JD Power online). According to our results, the practice of summarizing numerical data via word-based descriptions may be misleading at times.

Limitations and future research

Our findings are limited to a single instance of exposure to expectation-relevant information. Perhaps employing repeated processing of such counter-schematic information might allow for more conscious shifts (similar to after-purchase product trial as in Stayman, Alden, & Smith, 1992) and this line of research could provide new insights into the issue of brand equity formation.

It should also be stated that we have not investigated which ratings are more important in determining actual consumer behavior. Thus, we cannot say that numerical ratings are more predictive than word-based ratings of what consumers might do. However, it seems reasonable to argue that the numerical ratings give a more representative picture of how the brands compare. We can assume that, over time, brand expectations will closely align with actual facts (in terms of product quality, performance, etc.), which are more clearly reflected in numerical-unit ratings. By contrast, subjective, word-based ratings involve a better-than-expected element that is likely a short term phenomenon.

An immediate extension of the present research would therefore address consumers' behavioral response after having engaged in the attitudinal shift produced by the SSE. Thus, the shift may well induce a consumer to make what might be called a Type II error of commission: taking a chance on glowing reviews of a presumed inferior brand only to find out that it was as bad as expected. More objective ratings, of course, might also

induce mistakes, call them Type I errors of omission: missing out on that well-reviewed but presumed inferior brand which may in fact be much better than expected. Thus reliance on objective ratings goes naturally with the more risk-averse stance of avoiding Type I errors, whereas reliance on subjective judgments might induce the greater risk-taking chance of committing Type II errors.

In the end, while it may be soothing for a weak competitor to be evaluated in positive terms, one suspects that high praise and positive subjective ratings are no more than empty words. Consumers may well act on their perceptions rather than objective facts, but one would presume that over time the true product qualities and related brand expectations become closely aligned and based in reality. Then, the meaningful words associated with subjective ratings will be in fact meaningless.

Acknowledgments

This research was supported in part by fellowships to the first author from Georgetown University. We are grateful for the helpful comments provided by Ron Goodstein, Chris Janiszewski, Maureen Morrin, and participants in the summer research seminar series at the McDonough School of Business.

References

Alba, J. W., & Hutchinson, W. J. (1987). Dimensions of consumer expertise. *Journal of Consumer Research*, *13*(4), 411−455.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, *51* (6), 1173−1182.

Biernat, M., Manis, M., & Nelson, T. E. (1991). Stereotypes and standards of judgment. *Journal of Personality and Social Psychology*, *60*(4), 485−499.

Biswas, A., & Sherrell, D. L. (1993). The influence of product knowledge and brand name on internal price standards and confidence. *Psychology & Marketing*, *10*(1), 31−47.

Campbell, M. C., & Goodstein, R. C. (2001). The moderating effect of perceived risk on consumers' evaluations of product incongruity: Preference for the norm. *Journal of Consumer Research*, *28*(3), 439−449.

Cho, H., & Schwarz, N. (2008). Of great art and untalented artists: Effort information and the flexible construction of judgmental heuristics. *Journal of Consumer Psychology*, *18*(3), 205−211.

Erdem, T., & Swait, J. (1998). Brand equity as a signaling phenomenon. *Journal of Consumer Psychology*, *7*(2), 131−157.

Gannon, K. M., & Ostrom, T. M. (1996). How meaning is given to rating scales: The effects of response language on category activation. *Journal of Experimental Social Psychology*, *32*(4), 337−360.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464−1480.

Greenwald, H. J., & Satow, Y. (1970). A short social desirability scale. *Psychological Reports*, *27*(1), 131−135.

Gurmankin, A. D., Baron, J., & Armstrong, K. (2004). The effect of numerical statements of risk on trust and comfort with hypothetical physician risk communication. *Medical Decision Making*, *24*, 265−271.

Jacoby, J. (1978). Consumer research: How valid and useful are all our consumer behavior research findings? A state of the art review. *Journal of Marketing*, *42*(3), 87−96.

Keller, K. L. (1993). Conceptualizing, measuring and managing customer-based brand equity. *Journal of Marketing*, *57*(1), 1−22.

Kim, H., & John, D. R. (2008). Consumer response to brand extensions: Construal level as a moderator of the importance of perceived fit. *Journal of Consumer Psychology*, *18*(2), 116−126.

Kruger, J., & Vargas, P. (2008). Consumer confusion of percent differences. *Journal of Consumer Psychology*, *18*(1), 49−61.

Lane, V. R. (2000). The impact of ad repetition and ad content on consumer perceptions of incongruent extensions. *Journal of Marketing*, *64*(1), 80−91.

Maheswaran, D., & Sternthal, B. (1990). The effects of knowledge, motivation, and type of message on ad processing and product judgments. *Journal of Consumer Research*, *17*(1), 6−73.

Monga, A. B., & John, D. R. (2008). When does negative brand publicity hurt? The moderating influence of analytic versus holistic thinking. *Journal of Consumer Psychology*, *18*(4), 320−332.

Morwitz, V. G., & Fitzsimons, G. J. (2004). The mere measurement effect: Why does measuring intentions change actual behavior? *Journal of Consumer Psychology*, *14*(2), 64−74.

Ofir, C., & Simonson, I. (2007). The effect of stating expectations on customer satisfaction and shopping experience. *Journal of Marketing Research*, *44* (1), 164−174.

Rose, R. L., Miniard, P. W., Barone, M. J., Manning, K. C., & Till, B. D. (1993). When persuasion goes undetected: The case of comparative advertising. *Journal of Marketing Research*, *30*(3), 315−340.

Stayman, D. M., Alden, D. L., & Smith, K. H. (1992). Some Effects of Schematic Processing on Consumer Expectations and Disconfirmation Judgments. *Journal of Consumer Research*, *19*(2), 240−255.

Theil, M. (2002). The role of translations of verbal into numerical probability expressions in risk management: A meta-analysis. *Journal of Risk Management*, *5*(2), 177−186.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124−1130.

Vandello, J. A., Goldschmied, N. P., & Richards, D. A. R. (2007). The appeal of the underdog. *Personality and Social Psychology Bulletin*, *33*(12), 1603−1616.

Viswanathan, M., & Childers, T. L. (1996). Processing of numerical and verbal product information. *Journal of Consumer Psychology*, *5*(4), 359−385.