Journal of Experimental Psychology: Applied
2013, Vol. 19, No. 2, 143–157

1076-898X/13/$12.00 DOI: 10.1037/a0033298

Verbal and Numerical Consumer Recommendations: Switching Between Recommendation Formats Leads to Preference Inconsistencies

Boris Maciejovsky
University of California, Riverside

David V. Budescu
Fordham University

Many Web sites provide consumers with product recommendations, which are typically presented by a sequence of verbal reviews and numerical ratings. In three experiments, we demonstrate that when participants switch between formats (e.g., from verbal to numerical), they are more prone to preference inconsistencies than when they aggregate the recommendations within the same format (e.g., verbal). When evaluating recommendations, participants rely primarily on central-location measures (e.g., mean) and less on other distribution characteristics (e.g., variance). We explain our findings within the theoretical framework of stimulus–response compatibility and we make practical recommendations for the design of recommendation systems and Web portals.

Keywords: consumer recommendations, numerical presentation, preference inconsistency, stimulus–response compatibility, verbal presentation

More and more consumers purchase products on the Internet. Unlike traditional brick and mortar stores, virtual marketplaces do not allow consumers to inspect directly the physical properties of the goods and commodities. This lack of immediate and direct experience increases the uncertainty about their quality. One way to reduce this uncertainty is by allowing consumers to share their experiences in the form of reviews and recommendations (Roselius, 1971). Indeed, many leading Internet marketplaces (e.g., Amazon, eBay) offer consumers the opportunity to post accounts of their experiences in the form of free-format short verbal essays and numerical ratings.

These two recommendation formats, numerical and verbal formats, are likely processed differently by consumers and might lead to different patterns of preferences. Thus, one can expect increased levels of preference inconsistency across the two modes. Preference consistency is an important indicator of decision quality (Lee, Bertini, & Ariely, 2010), as it is a core assumption of standard decision and economic theory (Mas-Collel, Whinston, & Green, 1995). In this article we examine how preference consistency varies as a function of the compatibility between the presentation and the processing of verbal and numerical recommendations.

Related Literature

We study how—and to what extent—people integrate consumer recommendations by drawing upon the literatures on compatibility effects and people's ability to extract statistical properties of distributions. Based on these insights, we formulated a simple mathematical model that allows us to investigate preference consistency as a function of recommendation format and its underlying statistical properties. We designed a series of experiments that tested the predictions of the model, allowing us to provide evidence-based recommendations to improve the design of recommendation systems and Web portals, facilitating consumer choice.

Our systematic investigation of compatibility effects for the aggregation of recommendations/advice has practical relevance, as the majority of Web sites that display recommendations to consumers differentiate between numerical and verbal recommendations. We provide the first systematic investigation of whether or not this practice has implications on consumers' decision quality.

We are doing this by eliciting two individual preference orderings over various products in our studies. One ordering is based on the assessment of individual products, whereas the other is based on pairwise comparisons among the same products. We evaluated decision quality by studying the consistency between the two orderings under compatible and incompatible formats and using different distributions of recommendations.

Compatibility Effects

Research in industrial psychology has shown that people respond faster and more accurately to visual displays of information, such as instrument panels, if the structure of the response apparatus is *compatible* with the spatial arrangement of the stimuli (Fitts & Deininger, 1954; Fitts & Seeger, 1953). For example, a stove is easiest to operate when the controls are arranged in a similar fashion to the actual burners. Compatibility effects were also observed for tasks in which the spatial dimension is not relevant (Simon & Rudell, 1967) and for more complex displays, involving various types of manual response modes (Yamaguchi & Proctor, 2006).

Boris Maciejovsky, School of Business Administration, University of California, Riverside; David V. Budescu, Department of Psychology, Fordham University.

Correspondence concerning this article should be addressed to David V. Budescu, Anne Anastasi Professor of Psychometrics and Quantitative Psychology, Department of Psychology, Fordham University, Dealy Hall Room 220, 441 East Fordham Road, Bronx, NY 10458-9993. E-mail: budescu@fordham.edu

Compatibility effects have also been invoked as predictors in multiattribute decision-making (e.g., Slovic & McPhillamy, 1974) and as explanations for preference reversals, that is, for inconsistencies in revealed preferences as a function of elicitation method (Cubitt, Munro & Starmer, 2004; Lichtenstein & Slovic, 1971; Rubaltelli, Dickert & Slovic, 2012; Tversky, Slovic & Kahneman, 1990). It has been suggested that the weight of a stimulus attribute is enhanced by its compatibility with a response mode (Tversky, Sattath & Slovic, 1988). For example, the amount of money one can gain or lose is more prominent in pricing tasks (that also involve monetary responses) than in choice tasks. This work has been extended to different attribute–task combinations (Nowlis & Simonson, 1997) and to gambles with different probability information. For instance, González-Vallejo and Wallsten (1992) showed that when probabilities were presented in verbal format (e.g., likely), the rate of preference reversals was significantly reduced compared with numerical values because of a reduction of risk aversion in the choice task.

In the typical preference reversal studies the stimuli are identical but the response methods differ (choices, matching or pricing) and the basic explanation is violation of procedural invariance (Tversky et al., 1990). However, compatibility has many facets and manifestations (see Shafir, 1995). In this study we add a new layer of complexity by combining two of these facets. Our subjects receive recommendations in a specific format (e.g., verbal or numerical), and process it in a format that is either compatible or incompatible with the presentation format when asked to rate the stimuli or choose among them. The new prediction of the current study is that decision makers (DMs) who encounter recommendations in a specific format (e.g., verbal descriptions), and process it in a similar (compatible) format, integrate the information quicker and better than DMs who are forced to process the same information in a different (incompatible) format by requiring them to respond in a different modality (e.g., numerical rating scales). Consequently, decisions made under incompatible formats display higher levels of inconsistency compared with cases where the two formats are compatible.

Extracting Statistical Properties

Research on visual perception indicates that DMs extract the statistical properties of stimuli seemingly effortlessly. For instance, they judge accurately the mean size of a group of circles and recall the range of visual displays (Ariely, 2001). However, estimates of variability are systematically too low (Kareev, Arnon & Horwitz-Zeliger, 2002). DMs were also shown to be more confident in their judgments for mean differences than differences in variability (Obrecht, Chapman & Gelman, 2007).

The efficiency of averaging as a process of information aggregation (e.g., Anderson, 1991; Budescu, 2006; Budescu & Yu, 2007; Busemeyer, 1991) extends to different domains, including size, motion, and even facial expressions (e.g., Alvarez & Oliva, 2008: Ariely, 2001; Haberman, Harp & Whitney, 2009; Peterson & Beach, 1967; Watamaniuk & Duchon, 1992), and is surprisingly accurate, a fact that continues to astonish laypeople and experts alike (Clemen, 2008; Larrick & Soll, 2006).

Based on these findings, we predict that when DMs evaluate multiple recommendations, they rely primarily on measures of central location (such as the mean, median or mode of the input values), and are less sensitive to other features of the distribution (e.g., its variance or range). We also generalize the well-known symbolic distance effect (Moyer & Bayer, 1976), which states that the time required to compare two items is inversely related to the distance between them on the target attribute. This leads to the prediction that DMs make more accurate and faster decisions for well-differentiated recommendations.

The Present Work

Previous research has shown how conflicting opinions are weighted (Budescu & Yu, 2006, 2007; Chevalier & Mayzlin, 2006; Gershoff, Mukherjee & Mukhopadhyay, 2003; West & Broniarczyk, 1998) and how DMs process and compare numerical and verbal expressions of uncertainty (Jaffe-Katz, Budescu & Wallsten, 1989; Wallsten, Budescu & Tsao, 1997).

We extend this line of work by studying (a) how well DMs aggregate recommendations across different modalities (numerical and verbal), (b) how the statistical properties of the recommendations (mean and variance) affect the aggregation process, and (c) how the aggregation process impacts preference consistency. Our results have practical implications for consumer decision-making and the design of Web portals.

We derive predictions for our experiments from a mathematical representation of the decision problem implemented in our studies. In Stage 1 of our studies, DMs evaluate products based on multiple recommendations. These recommendations are either numerical or verbal. Judgments based on these recommendations induce a preference ordering for each DM over the products. In Stage 2, DMs are shown pairs of the products from Stage 1 (some pairs use only numerical recommendations and others use only verbal ones) and are asked to aggregate the recommendations for each product and to choose their preferred product in each pair (by picking one of five possible summary options, which are either compatible or incompatible to the recommendation format). These choices result in a second preference ordering over the products. Preference consistency is investigated by comparing the orderings from Stage 1 and Stage 2.

In Study 1 the DMs are asked explicitly to quantify each product before making a choice in a mode that can be compatible or incompatible with the recommendation format. In Study 2 they are asked to make a choice first and are instructed to quantify only the chosen product verbally or numerically. Postchoice quantification can also be either compatible or incompatible with the recommendation format. For instance, if the recommendations are shown in verbal format, the DM knows whether the summary options will be described in verbal terms (compatible format) or in numerical ratings (incompatible format) before making the choice.

Thus, the DMs recognize that they are in a compatible or in an incompatible situation before making a choice in both studies. Although the priming to switch from one modality to another is more direct and explicit in the first study, and more subtle in the second, we do not distinguish between the two studies in the description of our model:

Notation

J = Number of judges (recommendations)

N_{ijx} = Numerical recommendation j of product x presented to rater i (Stage 1 and 2).

V_{ijx} = Verbal recommendation j of product x presented to rater i (Stage 1 and 2).

R_{iNx} = Rating (Stage 1 and Stage 2) of product x by rater i based on J numerical recommendations.

R_{iVx} = Rating (Stage 1 and Stage 2) of product x by rater i based on J verbal recommendations.

Stage 1

We assume that the DMs' ratings are the average of the numerical and verbal recommendations, perturbed by a random component that is (a) unbiased (i.e., has zero mean) and (b) captures various sources of errors (e.g., misperceptions of the recommendations, miscalculations, and misreporting) that cannot be disentangled (see Budescu, Rantilla, Yu, & Karelitz, 2003). Thus, for the numerical case,

$$R_{iNx} = \sum_{j=1}^{j} N_{ijx}/J + N(0, \sigma_{iN}^2). \quad (1)$$

In particular we assume that $\sigma_{iN}^2 \propto Var\,(N_{ij})$, that is, the variance of rater i is proportional to the variance of the numerical recommendations seen by him/her, so the lower (higher) the agreement between the recommendations the higher (lower) is the variance (e.g., Budescu et al., 2003). For the verbal recommendations the process is similar but it seems unnecessarily restrictive to assume that the variance will be the same. In a general model it makes sense to assume an additional parameter θ that reflects the vagueness of verbal communication:

$$R_{iVx} = \sum_{j=1}^{j} N_{ijx}/J + N(0, \sigma_{iN}^2) \text{ with } \sigma_{iV}^2 \propto Var(\theta V_{ij}) \text{ where } \theta \geq 1. \quad (2)$$

In our experiments we made a special effort to equate the two response modes (see details below), so we assume that the variance is the same, that is, $\theta = 1$.[1] Thus, we can think of the observed ratings of products *x, y, z,* and so forth, by a randomly selected judge, as being drawn from normal distributions respectively, where $\sigma_{iV}^2 \geq \sigma_{iN}^2$:

$$R_{iNx} \sim N\left[\left(\sum_{j=1}^{j} N_{ijx}/J\right), \sigma_{iN}^2\right] \quad (3)$$

$$R_{iV}x \sim N\left[\left(\sum_{j=1}^{j} V_{ijx}/J\right), \sigma_{iV}^2\right], \quad (4)$$

Stage 2

DMs aggregate the J (numerical or verbal) recommendations by the same process used in Stage 1. For example, in Study 1, we require the subjects to choose a summary measure for the 10 recommendations encountered for each product. Thus, when asked to choose one of the two products, the DM faces a typical Thurstonian choice process (Thurstone, 1927), and the choice can be modeled by the distribution of the difference between the two ratings (under the "standard" assumptions of case *V*, namely homogeneity of variance and zero covariance).

Compatible Cases

Numerical Recommendations and Numerical Responses

This is "the canonical case." The *i*th DM is presented with J numerical recommendations for product x and J numerical recommendations for product y, so the difference between the two aggregates (based of Eq. 3 and the case *V* assumptions) is distributed:

$$D_{iN(x-y)} = \left(R_{iNx} - R_{iNy}\right) \sim N\left[\left(\sum_{j=1}^{J} N_{ijx} - \sum_{j=1}^{J} N_{ijy}\right)/J, 2\sigma_{iN}^2\right] \quad (5)$$

If the difference is positive (negative), the DM picks product $x(y)$, and if it is 0 the DM chooses one of the products with equal probability. Everything else (e.g., the number of recommendations, the level of agreement between the recommendations) being equal, the choices are easier as the difference between the two aggregates increases. When the mean recommendations of the two products are similar, the proportion of choices (of either one) is close to 0.5; the more distinct the means of the two sets of recommendations are, the closer one of the two proportions is to 1 (with the other approaching 0).

In addition to this "distance effect" on the rate of choices two additional predictions follow. One relates to the decision time: the farther apart the two recommendations are and the easier the choice, the faster it will be completed.[2] This generalizes the well-known symbolic distance effect (Moyer & Bayer, 1976), stating that the time required for comparing two items is inversely related to the distance between them on the target attribute. Previous work on this effect has assumed and used unique and well-defined stimuli (e.g., two quantities), whereas here we consider the case where the DM needs to integrate J pieces of information by him/herself to obtain the two stimuli.

The other prediction relates to the inconsistency between the two stages that is the core of our work: Even when $R_{iNx} > R_{iNy}$ it is possible that product y will be chosen over product x, because of the variance and the overlap between the two distributions. This prediction is consistent with the "representational overlap view" in studies on the distance effect, suggesting that this effect is not restricted to numerical stimuli, but extends to other potentially spatially represented stimuli, like social status (Chiao, Bordeaux, & Ambady, 2005), the size of animals (Paivio, 1975), and geographical locations (Maki, 1981). In fact, the probability of such reversals is a monotonic function of the standardized difference between the two mean values, $(R_{iNx} - R_{iNy})/\sigma_{iN}$: The closer (more distant) the two distributions are, the more (less) likely we are to observe inconsistencies between the ratings and the choices.

[1] In Study 2, a group of 15 subjects rated the same (numerical and verbal) products twice so it was possible to estimate the within subject variance across the two replications. We did not find a significant difference between the variances under the two modalities (overall and for each of the individual products), confirming the assumption that $\theta = 1$.

[2] This is consistent with the model of Johnson and Busemeyer (2005), capturing the finding that in a preference reversal setting, individuals require a longer time to price a high-variance option as compared to a low-variance option.

Verbal Recommendations and Verbal Responses

This case mimics in all respects the canonical NN choice with obvious changes in notation. If the *i*th DM is presented with *J* verbal recommendations for products *x* and *y*, respectively, the difference between the two aggregates (based of Eq. 4 and the case *V* assumptions) is:

$$D_{iV(x-y)} = \left(R_{iVx} - R_{iVy}\right) \sim N\left[\left(\sum_{j=1}^{J} V_{ijx} - \sum_{j=1}^{J} V_{ijy}\right)/J, 2\sigma_{iV}^2\right] \tag{6}$$

Thus, if the mean verbal ratings and their variances match closely the means and variances from the numerical cases we should observe similar results in the two cases. Of course, if the variances in the verbal cases are higher (i.e., $\theta > 1$) we would expect higher rates of reversals, longer decision times, and so forth In our studies we made significant efforts to match the two cases, so we don't expect such differences (please see the pretest in Study 1 for details).

Incompatible Cases

Numerical recommendations and verbal responses and verbal recommendations and numerical responses. In both cases the DM is aware of the fact that the recommendation accompanying the final choice of a product requires switching to a different modality than the one used in the presentation of the recommendations. The key assumption invoked here is that this primes the DM to convert the mean ratings of the products, R_{iNx} or R_{iVx}, to the response modality before making the choice. This conversion preserves the gist of the representation,[3] as captured by the mean value, but introduces a new and distinct source of variance associated with the conversion process.

Empirical evidence suggests that there exists considerably more between-subjects variability in the interpretation of verbal (uncertainty) information as compared with numerical (uncertainty) information (e.g., Beyth-Marom, 1982; Jaffe-Katz et al., 1989; Wallsten et al., 1997), suggesting that intermodality comparisons can only be captured by a conversion or "translation" mechanism. Windschitl and Wells (1996) provide indirect evidence for the necessity of such a conversion mechanism by showing that numerical information leads to a more deliberate and rule-based reasoning relative to verbal information (which leads to a more associative and intuitive thinking). In their experiments, numerical measures of uncertainty failed to detect variation in subjects' responses that was picked up by verbal measures. Moreover, that variation was shown to be important for subjects' preferences.

To model these findings, we introduce a conversion factor β, which operates on the variance of the recommendations (assuming $\beta \geq 1$). Thus the observed ratings of products *x, y, z,* and so forth by a randomly selected judge in the incompatible settings is modeled by the following:

$$R_{iNx} \sim N\left[\left(\sum_{j=1}^{j} N_{ijx}/J\right), \beta\sigma_{iN}^2\right] \tag{7}$$

$$R_{iV}x \sim N\left[\left(\sum_{j=1}^{j} V_{ijx}/J\right), \beta\sigma_{iV}^2\right], \tag{8}$$

respectively, where $\sigma_{iV}^2 \geq \sigma_{iN}^2$ and $\beta \geq 1$. The rest of the considerations and calculations are similar to the compatible cases and all the predictions apply with some subtle differences: (1) the distance effect on choice is identical: the more distant the means of the two sets of recommendations are, the easier it is to choose the higher one; (2) the symbolic distance effect as reflected in decision times also holds but, because of the extra conversion component, and everything else being equal, decision times will be longer in the incompatible case (see Jaffe-Katz et al., 1989); (3) higher rates of reversals across the two stages. Recall that the probability of reversals is a monotonic function of the standardized difference between the two mean values, which now is affected by the conversion factor: $(R_{iNx} - R_{iNy})/\beta_{iN}$. Thus, the closer (more distant) the two distributions are, the more (less) likely we are to observe inconsistencies between the ratings and the choices. However, in all cases the inconsistency rates would be higher than in the compatible mode (verbal or numerical) because of the increase in variance due to the need to convert.

We test these predictions in three experimental studies.

Study 1

We seek to establish the basic effect of stimulus–response compatibility on consistency of preferences, and to illustrate the salience of the central location of the distribution of advice. We obtained two individual preference orderings over a group of products—one based on graphical ratings of individual products, and the other based on pairwise comparisons among the same products, using compatible and incompatible modalities. As predicted by the model, we expect a higher rate of preference inconsistencies between the two when the stimulus presentation format (verbal or numerical) does not match the processing modality of the response format (verbal or numerical). For example, if DMs are presented with verbal recommendations, the rate of preference inconsistencies would be higher (lower) when they are asked to provide a numerical (verbal) evaluation which forces them to process these recommendations in a different modality. We also manipulated the means and standard deviations of the distribution of recommendations for each product. We predict the highest rates of preference inconsistencies when the two distributions have equal means and the salient differences between them are related to the two variances. Conversely, we predict the lowest rates of preference inconsistencies when the most salient differences between the two distributions are related to the two means.

[3] In the terminology of fuzzy-trace theory (see Reyna, 2012, for a recent survey), gist refers to the bottom-line meaning of information, which can be imprecise and "fuzzy." Verbatim information, in contrast, is literal, veridical, and detailed. When people make judgments, gist representations seem more important than verbatim representations, because individuals prefer to operate on fuzzier ordinal or categorical representations, and typically have difficulty assigning exact numerical values to entities, especially unfamiliar ones (Brainerd & Reyna, 2001; Reyna, 2008). Of particular importance for our setup is the fact that verbatim traces deteriorate more rapidly than gist traces. Thus, verbatim traces may be inaccessible when this information is required for subsequent decisions (Brainerd, Reyna, & Kneer, 1995).

Pretest

Seventy-one MIT students (41% females), aged 19 to 29 years ($M = 21.10$, $SD = 1.90$), were shown 50 short verbal statements, like "this product was all right" or "this product was amazing," and were instructed to assign each statement a numerical value of 1, 2, 3, 4, or 5, where a value of 1 indicated that the product was not at all recommended, and a value of 5 indicated that the product was highly recommended. We selected the 5 statements (one for each scale value) that yielded the highest rates of agreement across participants and used them in all subsequent studies. Table 1 lists the selected statements and their corresponding agreement rates.

Experimental Design and Procedure

Sixty-eight Princeton students (49% females), aged 18 to 24 years ($M = 20.20$, $SD = 1.50$), participated in the study. We used a 2 (stimulus format: verbal vs. numerical) × 2 (response format: verbal vs. numerical) mixed-factorial design. The first factor was manipulated between-subjects, the second within-subjects.

Participants were instructed to assume that all products were from the same category—household appliances or entertainment products—and had equal prices. In Stage 1 the participants rated the desirability of 12 products in isolation (see Figure 1 for a schematic screen shot of the task).[4] The products were shown in random order. For each product, participants saw a list of 10 recommendations, presented either as short verbal statements or as numerical ratings (in the form of "stars") on a 5-point scale. Thus, 6 products were displayed numerically and 6 products were displayed verbally. The verbal statements matched the 5-point numerical scale (according to Table 1). We manipulated the mean (low, medium, or high), the variability (low or high), and the presentation format (verbal or numerical) of the recommendations (see Table 2 for details) across products. Thus, each product represents a unique combination of 3 (means) × 2 (variances) × 2 (presentation formats).

After a short (about 10-min) filler task that involved the completion of unrelated consumer surveys, participants were shown pairs of the products used in Stage 1 in random order with the corresponding recommendations (see Figure 2a, for a screen shot). The task was to evaluate the two products. The evaluation was done both verbally and numerically, generating two blocks of ($6^*5/2 =$) 15 pairs of products. Thus, each participant evaluated 30 pairs of products. The order of the two blocks was counterbalanced. In this stage we manipulated compatibility: The response format in one block matched the format of the stimulus presentation (compatible condition), whereas in the other block it did not match (incompatible condition).

In the verbal response condition, participants were shown five verbal statements, and were asked to select a single global evaluation of the product's recommendations by clicking the most appropriate statement. The options ranged from "I am very unhappy with the product" to "I am very satisfied with the product" (see Figure 2b). In the numerical response condition, participants were shown five numerical scale values (1 to 5 stars), and were asked to select the most appropriate option (higher number of stars indicated a stronger recommendation). After evaluating the two products, participants were asked to pick their preferred product (see Figure 2c) and to indicate their confidence of having picked the superior one on a continuous rating scale (see Figure 2d).[5]

We provided incentives for consistent preferences across the two stages of the experiment through a lottery procedure. Financial incentives induce persistent diligence (which is particularly important for more mundane and repeated tasks, such as ours) and reduce the variability of responses, increasing replicability (Camerer & Hogarth, 1999). The probability of winning a $20 voucher for purchases on Amazon.com increased linearly in participants' preference consistency across both stages of the experiment.[6] Participants were told that the closer their choices among product pairs in Stage 2 replicated their initial preference orderings of Stage 1, the higher was their probability of winning the prize.

Table 1
Labels Used in the Verbal and Numerical Information Format

Scale value	Verbal format	Numerical format	Agreement rate in pilot study
1	I am very unhappy with the product	★☆☆☆☆	88.73
2	The product was quite bad	★★☆☆☆	85.92
3	The product was okay	★★★☆☆	90.14
4	The product was quite good	★★★★☆	84.51
5	I was very satisfied with the product	★★★★★	90.14
			$M = 87.89$

Results

The desirability ratings in Stage 1 allowed us to derive an individual preference ordering over the 12 products. Because the ratings were done graphically, they do not favor either processing (response) mode (verbal or numerical) and serve as the "neutral standard." The pairwise choices and the summary evaluations among product pairs in Stage 2 induced a second preference ordering.

Does incompatibility between the stimulus and response format increase the rate of preference inconsistencies? We computed for each participant the difference between the rate of preference inconsistencies in the incompatible and the compatible formats condition, based on choices among products (see Figure 2c).[7] An ANOVA of this measure as a function of the stimulus format (verbal or numerical) indicates that the number of preference inconsistencies is significantly higher with incompatible cases ($M_{NC} = 4.27/15 = 28.5\%$ vs. $M_C = 3.56/15 = 23.7\%$, and

[4] We used a sliding scale that discriminated between 650 levels to avoid ties in responses (but these levels were not shown to the participants; see Figure 1).

[5] The confidence ratings did not interact significantly with choices and evaluations, so we do not report these results.

[6] We counted the number of violations of the initial preference orderings across the two information formats, resulting in a number between 0 (no violations) and 30 (all violations). We used this number to derive the probability of winning the $20 voucher: (30 − # of violations)/30. Thus, the probability of winning the prize is 1 in case of 0 violations, and 0 in case of 30 violations.

[7] There was a 98% agreement between these choices and the evaluations (Figure 2b), so we only report these results.

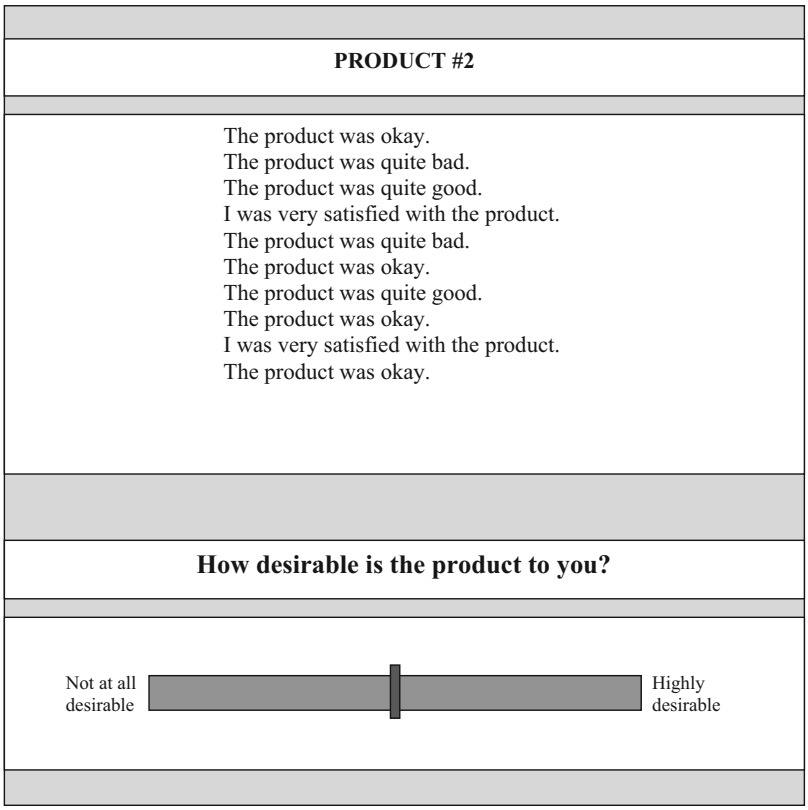

Figure 1. Schematic screen-shot of the desirability ratings task (Stage 1, Study 1, 2, and 3). This product is shown in verbal format and has a high mean (3.40) and a low standard deviation (1.07).

$F(1, 64) = 4.74$, $p < .05$, $\eta^2 = .11$), irrespective of the information format (i.e., there is no significant difference between the two stimulus presentation modes).

Which features of the recommendations drive preference inconsistencies? We considered separately four different types of product pairs (see Table 3 for a complete enumeration of all cases): (1) pairs of products with **d**ifferent **m**eans but **e**qual **s**tandard deviations (e.g., $M = 3.40$, $SD = 1.90$ and $M = 2.60$, $SD = 1.90$), which we label DMES; (2) pairs of products with **e**qual **m**eans but **d**ifferent **s**tandard deviations (e.g., $M = 3.40$, $SD = 1.90$ and $M = 3.40$, $SD = 1.07$), labeled EMDS; (3) pairs of products where the one with the **h**igher **m**ean has also a **h**igher **s**tandard deviation (e.g., $M = 3.40$, $SD = 1.90$ and $M = 3.00$ and $SD = 1.05$), labeled HMHS; and (4) pairs of products where the one that has the **h**igher **m**ean has a **l**ower **s**tandard deviation (e.g., $M = 3.40$, $SD = 1.07$ and $M = 3.00$, $SD = 1.94$), which we label HMLS.

We ran a 3-way ($2 \times 2 \times 4$) mixed ANOVA on the rate of preference inconsistencies with the between-subjects factor stimulus format (verbal or numerical) and two within-subjects factors—compatibility of response format (yes or no) and product combination (DMES, EMDS, HMHS and HMLS). The mean rates of preference inconsistencies are displayed in Table 4. We observed a significant main effect for stimulus format, $F(1, 66) = 5.39$, $p < .05$, $\eta^2 = .02$, indicating that the proportion of inconsistencies was slightly higher in the verbal (30%) than in the numerical format (24%). We also observed a significant main effect for product combination, $F(3, 64) = 37.40$, $p < .05$, $\eta^2 = .64$. Bonferroni post hoc tests that control the experimenter-wise error rate indicate that the rate of preference inconsistencies was significantly higher for EMDS than for the other combinations ($p < .05$).

To shed further light on the determinants of the preference inconsistencies, we regressed the number of inconsistencies for each pair of products on the following explanatory variables: The perceived similarity between the members of the pair, measured by the absolute mean differences in the ratings of the two products in Stage 1; the two means (1 = *the two products have the same mean*, 0 = *otherwise*); the two standard deviations (1 = *the two products have the same standard deviation*, 0 = *otherwise*); compatibility (1 = *stimulus and response format match*, 0 = *otherwise*); stimulus mode (1 = *numerical*, 0 = *verbal*); and response mode (1 = *numerical*, 0 = *verbal*). Table 5 presents correlation coefficients,

Table 2
Means (Standard Deviations) of the 10 Verbal and Numerical Recommendations Presented (Study 1 and Study 3)

	High standard deviation	Low standard deviation
High mean	3.40 (1.90)	3.40 (1.07)
Medium mean	3.00 (1.94)	3.00 (1.05)
Low mean	2.60 (1.90)	2.60 (1.07)

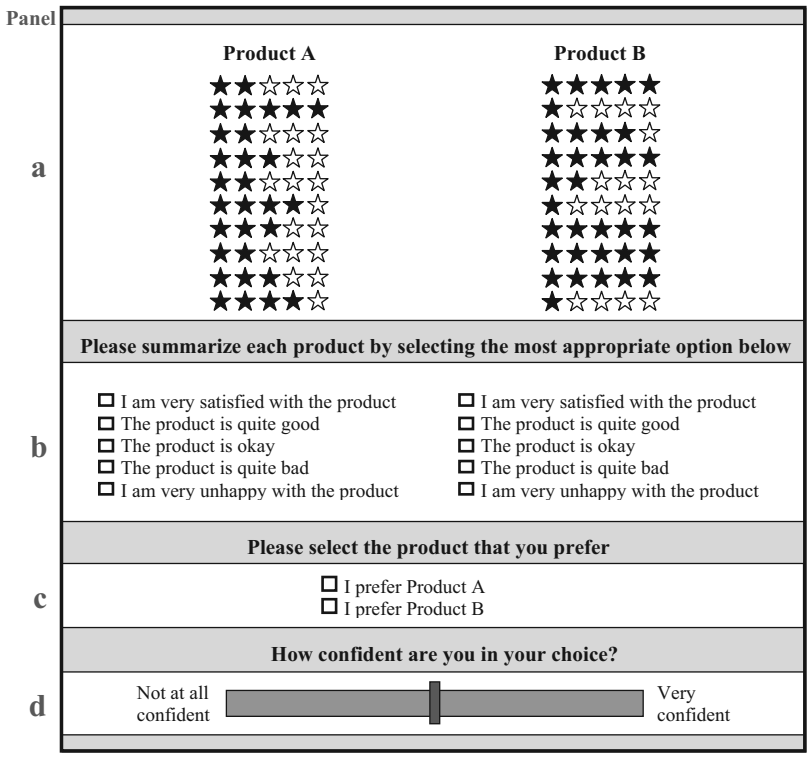

Figure 2. Schematic screen-shot of the pairwise choice task (Stage 2, Study 1 and 3). Products A and B are presented in numerical format. Product A has a medium mean (3.00) and a low standard deviation (1.05), whereas product B has a high mean (3.40) and a high standard deviation (1.90). Panel *a* shows the stimulus presentation, *b* the response task, *c* the choice task, and *d* the confidence task. For this particular example, the stimulus presentation (numerical) was incompatible with the response task (verbal).

standardized regression coefficients, and measures of global dominance (Azen & Budescu, 2003; Budescu, 1993) for all predictors. The results indicate a systematic and highly predictable pattern, Adjusted $R^2 = .74$, $F(5, 54) = 35.16$, $p < .05$. Inconsistencies are more likely when products are perceived as very similar, when they have identical means, when the stimulus format is incompatible with the response format, and when products are represented numerically. A dominance analysis (last column of Table 5), designed to identify the predictors' contributions, indicates that most variance is associated with products having identical (or different) means, followed by-products that are perceived to be similar to each other. Compatibility explains more variance than the stimulus and response modes combined.

Preference inconsistencies across response modes. To study the nature and source of the inconsistencies, we classified each participant according to his or her modal preference for low or high variability in all tasks. When the recommendations were presented verbally, a slight majority of participants favored high variance products (17/32 = 53% in Stage 1 and 19/32 = 59% in Stage 2), but when the recommendations were presented numerically most participants favored low variance products (23/36 = 64% in Stage 1 and 22/36 = 61% in Stage 2). The inconsistencies can be due to the differences between the two tasks (e.g., Tversky et al., 1988) and/or the underlying mental processes involved in these decisions. For instance, fuzzy trace theory (see Reyna, 2012, for a recent survey) suggests that in many contexts DMs radically simplify the decision relevant information by truncating and converting it (Reyna & Brainerd, 1991). According to this theory, when DMs face multiple options to choose from, they often rearrange and reformulate the problem mentally to facilitate judgments and decisions. In the numerical task, product comparisons are particularly easy, as the distributions of the numerical ratings (displayed as stars) can be compared visually. One can also argue that extracting an overall "average" feels easier in distributions with less variance. Prior evidence suggests that if something feels "easy," this information enters the decision-making process (e.g., Oppenheimer, 2008; Schwarz, 1998), and boosts choice shares for the easy-to-compare option (e.g., in the attraction effect/asymmetric dominance effect; Huber & Puto, 1983; Simonson, 1989). In our study, lower-variance numerical products are "easier" to summarize and might therefore be preferred over higher variance options. In contrast, products described verbally are not susceptive to such a "visual aggregation" strategy. Here, the information needs to be extracted and evaluated in a more piecemeal fashion, with no clear preference for low- or high-variance distributions.

The rate of inconsistencies across the two tasks was 40% when the formats were compatible and 57% when they were incompatible. Thus, the net effect of format incompatibility was 17% (or about half the size of the standard preference-reversal baseline). The net effect of incompatibility was considerably higher for the numerical stimuli (28%) than for the verbal stimuli (6%).

Table 3
Classification of the 15 Product Pairs Into 4 Classes (Study 1 and Study 3)

Class of pairs			
DMES (6 pairs)	EMDS (3 pairs)	HMHS (3 pairs)	HMLS (3 pairs)
3.40 (1.07)	3.40 (1.07)	3.40 (1.90)	3.40 (1.07)
3.00 (1.05)	3.40 (1.90)	3.00 (1.05)	3.00 (1.94)
3.40 (1.07)	3.00 (1.05)	3.00 (1.94)	3.00 (1.05)
2.60 (1.07)	3.00 (1.94)	2.60 (1.07)	2.60 (1.90)
3.00 (1.05)	2.60 (1.07)	3.40 (1.90)	3.40 (1.07)
2.60 (1.07)	2.60 (1.90)	2.60 (1.07)	2.60 (1.90)
3.40 (1.90)			
3.00 (1.94)			
3.40 (1.90)			
2.60 (1.90)			
3.00 (1.94)			
2.60 (1.90)			

Note. The entries in the table identify product pairs by showing their means and standard deviations (in parenthesis). *DMES* denotes product pairs with different means, but equal standard deviations; *EMDS* denotes product pairs with equal means, but different standard deviations; *HMHS* denotes product pairs where one has the higher mean and a higher standard deviation; and *HMLS* denotes product pairs where one has the higher mean and a lower standard deviation.

Discussion

The results support our predictions. When DMs were asked to integrate recommendations and summarize them in an incompatible format, forcing a different mode of information processing (e.g., verbal inputs and numerical evaluations), preferences were significantly less consistent with the original ratings compared to compatible integration (e.g., verbal to verbal). Preferences were least consistent for product pairs with identical means but different standard deviations, suggesting that DMs' choices were most difficult when they could not rely on measures of central locations to differentiate between them.

Study 2

The majority of inconsistencies in Study 1 occurred for products with identical means but different standard deviations. It is possible that these "inconsistencies" are spurious, because most DMs might be indifferent between products that have similar (in some cases identical) means. Many inconsistent preferences occurred for products that were rated similarly in the initial graphical rating task. One possible interpretation is that this task resulted in errors in evaluating these products and "correction" of these errors in the pairwise choice task qualified as an inconsistency. This finding may merely reflect the fact that the pairwise choice task is superior, in terms of evaluating options consistently, to the graphical rating task. Similarly, because consistency was evaluated across two independent stages, it might be that one of the two formats provides better memory cues than the other, and this difference between modalities was misinterpreted as an intermodality inconsistency.

Study 2 was designed to rule out these alternative accounts by using only products that have similar standard deviations, but different means, ensuring that decision quality is more directly linked to preference consistency. We also put the two evaluation methods—graphical rating scales and the pairwise comparisons—on equal footing by randomizing their order. In the new study we also measure decision times. This allows us to investigate whether one of these methods is superior in facilitating consistent evaluations and to test the symbolic distance hypothesis.

Experimental Design and Procedure

One hundred sixty-five Imperial College London students (41% females), aged 18 to 31 ($M = 22.04$, $SD = 2.41$), participated in the study. Participants saw products, which were characterized by 10 independent recommendations each. We used five different products with equally spaced means and with (almost) identical standard deviations. The means (*SD*s) of the five distributions were 2.2 (1.03), 2.6 (1.07), 3.0 (1.05), 3.4 (1.07), and 3.8 (1.03). Participants were instructed to study the products carefully before recommending them to a close friend. This task involved (a) choosing the one product that they would recommend their friend to buy and (b) select the most appropriate recommendation (out of five possible recommendations) that best captures its strength (see Figure 3 for a screen-shot).

We ran, essentially, two yoked experiments. The majority of the subjects ($n = 120$) took part in the main study that used both elicitation methods (ratings and pairwise choices). We used a three-way mixed design with two between-subjects factors and one within-subjects factor. The between-subjects factors were the order of the tasks (rating followed by pairwise comparison, or pairwise comparison followed by ratings) and the presentation format of the recommendations in the pairwise comparison condition (verbal or numerical). The within-subjects factor was the response format in the pairwise comparison condition (verbal or numerical). The two stages were separated by 10 minutes of unrelated filler tasks.

The second subexperiment used the same five distributions of recommendations. A sample of $n = 45$ subjects used the same

Table 4
Rate (in %) of Preference Inconsistencies (and Standard Errors) as a Function of Stimulus and Response Compatibility (Study 1)

Format		Product pairs				
Stimulus	Response	DMES	EMDS	HMHS	HMLS	Overall
Numerical	Numerical	22.0 (2.7)	32.4 (3.5)	22.7 (2.7)	19.5 (2.9)	24.2 (2.9)
Numerical	Verbal	21.6 (2.5)	32.7 (3.5)	20.6 (2.9)	24.2 (2.8)	24.8 (2.9)
Verbal	Verbal	25.8 (2.8)	39.0 (3.6)	24.2 (3.4)	25.4 (3.5)	28.6 (3.3)
Verbal	Numerical	30.1 (2.6)	40.0 (3.6)	30.2 (3)	29.0 (3)	32.3 (3.0)
Overall		24.9 (2.6)	36.0 (3.5)	24.4 (3.0)	24.5 (3.0)	27.5 (3.1)

Table 5
Regression Analysis on the Frequency of Preference Inconsistencies (Study 1)

Predictor	Zero-order correlations	Standardized coefficients	t (df = 59)	General dominance
Perceived similarity	−.70	−.37	−4.39*	.27
Equal means	.79	.51	5.69*	.36
Equal standard deviations	−.42	−.12	−1.69	.08
Compatibility	−.19	−.19	−2.84*	.03
Stimulus numerical	.12	.16	2.46*	.02
Response numerical	−.07	−.07	−1.10	.01
Model's R^2 (adjusted R^2)				.77 (.74)

* $p < .05$.

method (ratings or pairwise choices) on both stages, allowing us to measure the within task consistency and stability. Table 6 summarizes the experimental conditions and sample sizes. Incentives for consistent responses were provided as in Study 1.

Results

Comparing ratings with pairwise comparisons. Consider first the 4 groups in which participants performed both ratings and pairwise comparisons, allowing us to evaluate preference consistency. We performed a four-way mixed ANOVA with two between-subjects factors (1) recommendation format (numerical/verbal) and (2) order (pairwise comparison-rating/rating-pairwise comparison) and two within-subjects factors (3) compatibility between presentation and response formats (yes/no) and (4) distance between products in a pair (1,2,3,4). To understand the last factor, recall that we had five different products, whose average recommendations were equally spaced (with average recommendations of 2.2, 2.6, 3.0, 3.4, and 3.8). Thus, we have 4 pair-comparisons that involve products that are 1 step apart (e.g., products with

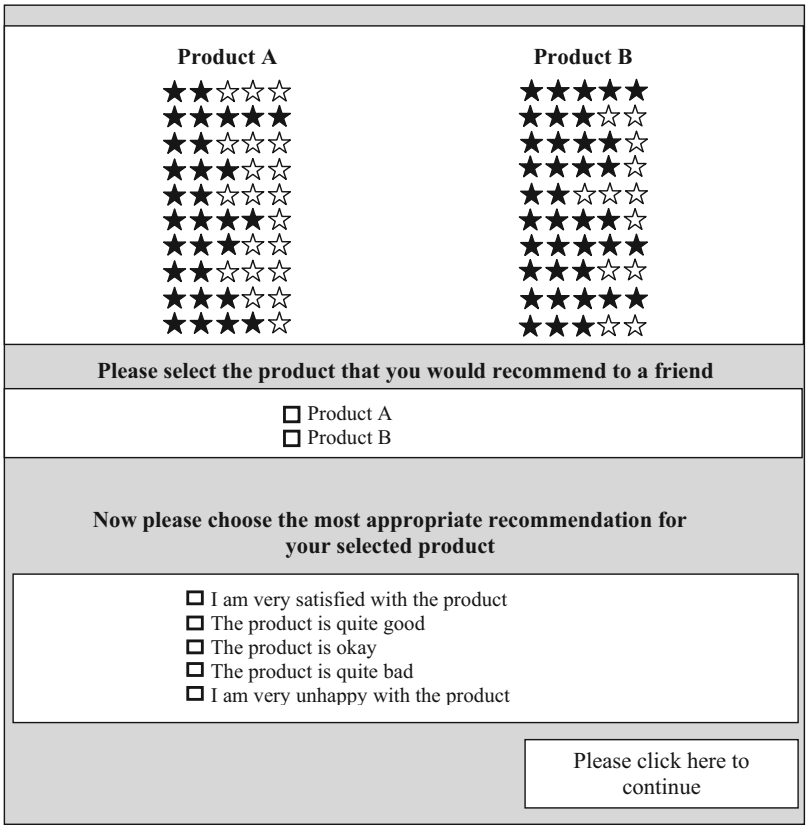

Figure 3. Schematic screen-shot of the pairwise choice task (Study 2). Products A and B are presented in numerical format. Product A has a mean of 3.0 (standard deviation of 1.05), whereas product B has a mean of 3.8 (standard deviation of 1.03). For this particular example, the stimulus presentation (numerical) was incompatible with the response task (verbal).

Table 6
Summary of the Experimental Conditions (Study 2)

Stage 1	Stage 2	Presentation mode	Goal of the experimental condition	Sample size
R	R	N and V	Evaluate reliability/stability of ratings: All 10 products (5N and 5V) presented in random order in both stages	15
PWC	PWC	N	Evaluate reliability/stability of choices: All 10 pairs of products presented twice (response N and V) in random order in both stages	15
PWC	PWC	V	Evaluate reliability/stability of choices: All 10 pairs of products presented twice (response N and V) in random order in both stages	15
R	PWC	N	All 10 products (5N and 5V) presented in random order in Stage 1: All 10 pairs of products presented twice (response N and V) in random order in Stage 2	30
R	PWC	V	All 10 products (5N and 5V) presented in random order in Stage 1: All 10 pairs of products presented twice (response N and V) in random order in Stage 2	30
PWC	R	N	All 10 pairs of products presented twice (response N and V) in random order in Stage 1: All 10 products (5N and 5V) presented in random order in Stage 2	30
PWC	R	V	All 10 pairs of products presented twice (response N and V) in random order in Stage 1: All 10 products (5N and 5V) presented in random order in Stage 2	30
Total		165		

Note. R = rating; PWC = pairwise choice; N = numerical; and V = verbal.

means 2.2 and 2.6 or products with means 3.0 and 3.4), 3 pair-comparisons that involve products that are separated by 2 steps (e.g., products with means 2.2 and 3.0 or products with means 3.0 and 3.8), 2 pair-comparisons that involve products 3 steps apart (e.g., products with means 2.2 and 3.4), and one pair-comparison that involves a distance of 4 (products with means 2.2 and 3.8).

The 4-way ANOVA yields insignificant effects for recommendation format and order of tasks. However, compatibility, $F(1, 116) = 5.39$, $p < .05$, $\eta^2 = 0.04$, and distance between products pairs, $F(1, 114) = 37.04$, $p < .05$, $\eta^2 = 0.49$, were statistically significant. There are more inconsistencies for incompatible pairs ($M = 27.6\%$ $SD = 17.7$ compared with $M = 24.1\%$ $SD = 18.0$). A trend analysis using orthogonal polynomials indicates that the rate of inconsistencies increases linearly as a function of distance (only the linear contrast was significant): $F(1, 116) = 98.10$, $p < .05$, $\eta^2 = 0.46$. Figure 4 shows the average rate of preference inconsistencies as a function of distance and compatibility.

After making their choices, participants were asked to summarize the chosen product on a 5-point scale in order to recommend it to their friends. We computed the absolute deviation of the summary judgment from the actual mean recommendation of the product and used the mean (across all pairs) of this measure in a 3-way ANOVA with one within-subject repeated factor (1) compatibility (yes/no) and two between-subjects factors, (2) order (pairwise comparison - rating/rating - pairwise comparison), and (3) recommendation format (numerical/verbal). The order and recommendation format were not significant; however, compatibility, $F(1, 116) = 59.44$, $p < .05$, $\eta^2 = 0.34$, was, indicating that the mean deviation was larger for incompatible product pairs ($M = 0.86$; $SD = 0.23$) than for compatible ones ($M = 0.66$, $SD = 0.15$).

The compatibility hypothesis predicts that decision times would be slower for pairs of products, which are processed in a modality that is incompatible with the presentation format, and the symbolic distance effect predicts that decision times would be related to the distances between the members of the product pairs. The results of a 4-way mixed ANOVA on decision times in the pairwise comparison task with two repeated factors (1) compatibility (yes/no) and (2) distance within pair (1, 2, 3, or 4 steps apart) and two between-subjects factors (3) recommendation format (numerical/verbal) and (4) order (pairwise comparison - rating/rating - pairwise comparison) yield significant effects for compatibility, $F(1, 116) = 5.38$, $p < .05$, $\eta^2 = 0.04$, distance, $F(3, 114) = 21.96$, $p < .05$, $\eta^2 = 0.37$, recommendation format, $F(1, 116) = 6.11$, $p < .05$, $\eta^2 = 0.05$, and the interaction between distance and compatibility, $F(3, 114) = 5.64$, $p < .05$, $\eta^2 = 0.13$.

As expected, decision times were slower for incompatible product pairs ($M = 12.63$; $SD = 2.61$) than for compatible ones ($M = 11.96$; $SD = 2.55$), and they were faster for pairs of products that were presented numerically ($M = 11.85$; $SD = 1.99$) than for pairs presented verbally ($M = 12.75$; $SD = 2.00$). The relationship between decision time and distance has significant linear and quadratic components. Figure 5 shows the average decision times as a function of distance and compatibility.

Within-method comparisons. The average rate of inconsistencies in the repeated ratings condition was 2.93/10 = 29.3% ($SD = 14$). The average rate of inconsistencies for the repeated pairwise comparisons condition with verbal recommendations was 2.87/10 = 28.7% ($SD = 15$) and for the repeated pairwise com-

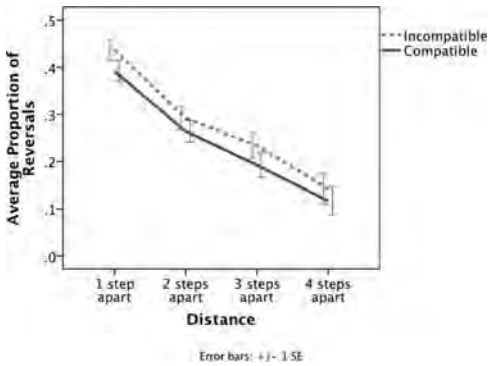

Figure 4. Average rate of preference inconsistency as a function of distance and compatibility (Study 2).

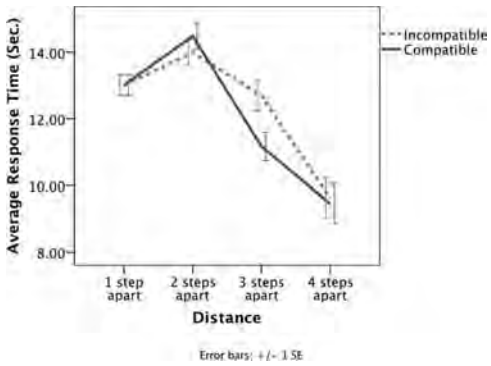

Figure 5. Average decision time (in seconds) as a function of distance and compatibility (Study 2).

parisons condition with numerical recommendations it was 2.8/10 = 28% (*SD* = 8). These values are consistent with similar measures reported in the literature that attest to the "noisiness" of individual choices (e.g., Loomes, Pinto-Prades, Abellan-Perpinan, & Rodriguez-Miguez, 2010; Por & Budescu, 2012). Most importantly, however, the results of a one-way ANOVA indicate that the three rates of inconsistencies do not differ significantly, $F(2, 44) = 0.04$, $p > .05$ across method. Thus, our findings cannot be attributed to different levels of reliability under the two elicitation methods.

Discussion

Study 2 confirmed the compatibility and distance effects while ruling out some methodological artifacts. We showed that the compatibility effects identified in the first study are neither qualified by the evaluation methods used (graphical ratings vs. pairwise comparisons) nor by the order of the two tasks. The results also indicate that the inconsistencies do not reflect random choices in cases of indifference, and cannot be attributed to the differential reliability/stability of the two tasks.

Remarkably, we showed that incompatibility led to participants' evaluations of products that are more distant from the original recommendations, which provides unequivocal proof that inconsistency causes deterioration in the quality of participants' decisions. Finally, this study shows that processing the information in a different modality is not necessary for the incompatibility effect to emerge—mere anticipation of a different modality is sufficient! We speculate that knowing that one is about to perform a task in a different modality evokes modality-specific expectations about what that task entails. For instance, knowing that after seeing verbal recommendations, one is about to evaluate these recommendations numerically, leads one to have a specific numerical scale in mind.

Study 3a

Having established the persistence and robustness of the compatibility effect, we investigate next whether DMs can anticipate its effects by asking for their predictions for various stimulus–response combinations. If participants recognize the importance of compatibility between stimulus and response format, they should choose the response format that matches the stimulus format in a majority of cases. If they don't, we expect to observe a preference for precise (numerical) information.

Experimental Design and Procedure

Fifty-one Harvard and MIT students (35% females), aged 18 to 24 years ($M = 20.45$, $SD = 1.55$), participated in a two-group between-subjects design, varying the stimulus format (verbal vs. numerical). Participants saw a product pair, described either by 10 verbal or 10 numerical recommendations each, and were asked to indicate whether they believed consumers would summarize the information more accurately verbally or numerically.

Results and Discussion

Most participants predicted that consumers would be able to summarize the information more accurately when using the numerical information *regardless* of the stimulus presentation, $\chi^2(1) = 14.23$, $p < .05$. Twenty-one of the 25 (84%) participants who saw the numerical stimulus presentation selected the compatible (numerical) response format, whereas only eight of the 26 (31%) participants who saw the verbal stimulus presentation selected the compatible (verbal) response format, suggesting that DMs do not anticipate compatibility effects.

Next we investigate whether this observation holds only when DMs predict the choices of others, or whether it also extends to one's own decisions. To this end we return to the experimental paradigm of Study 1.

Study 3b

This is a partial replication of Study 1, and uses the same distributions of recommendations. The unique feature of this study is that we allow participants to choose their own response format (verbal, numerical, or either mode) for every product pair. This study also allows us to retest the corollary of the compatibility hypothesis and the prediction of our model that DMs who process the stimulus information in one format (e.g., verbal), and then provide a summary evaluation in a different format (e.g., numerical), require a longer time to evaluate the product because of effort and time associated with cross-modality translation (see Jaffe-Katz et al., 1989).

Experimental Design and Procedure

Sixty-two Princeton students (34% females), aged 18 to 25 years ($M = 20.5$, $SD = 1.5$), participated in the study. The participants first rated the desirability of the 3 (means) × 2 (variances) × 2 (format: verbal and numerical) = 12 products used in Study 1 in isolation. After a filler task, they were shown 15 pairs of products, presented in verbal format, and 15 pairs of products, presented in numerical format. The 30 pairs were presented in random order. For each pair, participants could select one of three response formats for the product evaluation—verbal, numerical, or random (in which case the program selected the verbal or numerical format randomly). After providing their evaluations, participants were asked to pick their preferred product and to rate the confidence in their choice. Incentives for consistent preferences were provided as in Studies 1 and 2.

Results

Are participants sensitive to format incompatibility? Most participants preferred the numerical response format (67.6%), followed by the verbal format (17.6%), and the random determination of the response format (14.8%). Clearly, our participants did not anticipate the detrimental effects of incompatibility between stimulus and response format.

Table 7 displays the distribution of the preferred response format as a function of the presentation mode. The average (within subject) proportion of the numerical response format for products that were presented verbally was 0.56 as compared to 0.79 for products presented numerically, $t(61) = -4.90$, $p < .05$. Excluding all instances where participants opted for a random choice of the response mode, the corresponding proportions are 0.67 and 0.92, respectively, $t(56) = -4.58$, $p < .05$. Thus, the intensity of the general preference for numerical responses is stronger for compatible (numerical) cases.

Do participants process compatible and incompatible cases alike? Evaluations were considerably faster for compatible stimulus–response formats ($M = 23.19$ seconds, $SD = 6.00$) than for incompatible formats ($M = 29.57$ seconds, $SD = 8.62$). This difference is significant, $t(21) = 3.21$, $p < .05$.[8]

Discussion

Participants are unaware of the detrimental effects of stimulus–response incompatibility. We found a general preference for responding numerically, replicating the results of Study 3a and previous findings in the domain of probability communications (Brun & Teigen, 1988; Olson & Budescu, 1997; Wallsten, Budescu, Zwick & Kemp, 1993). One reason for this preference is the relative ease of comparing numerical values, such as prices, compared to verbal information, such as brand names (Viswanathan & Narayanan, 1994). Preference consistency was impaired by the lack of compatibility only when participants opted to respond numerically to product pairs that were presented verbally.

We showed that stimulus–response compatibility led to faster decisions (see similar results for congruent and noncongruent judgments in Jaffe-Katz et al., 1989), confirming that DMs require extra time to translate the information from one format to another.

General Discussion

The purpose of this article was to strengthen the link between basic research on preferences and consumer decision-making based on recommendations, a practice that has become increasingly common in the electronic marketplace. We focused on the compatibility hypothesis, which states that congruence between stimulus presentation and response modes facilitates information processing. This implies that information would be aggregated faster and communicated more accurately in cases in which stimulus and response formats match, compared with cases of incompatible formats, which force DMs to switch and convert across modes.

We showed that incompatibility between stimulus and response formats increases the rate of preference inconsistencies. Preferences are driven, almost exclusively, by the central values of the distribution of recommendations and are almost insensitive to their variances. The second study replicated and extended these findings while rejecting various artifactual (mostly methodological) interpretations. Although the study used a much more subtle manipulation of stimulus–response (in)congruence than the first one, we replicated the key compatibility effect and documented a systematic distance effect. Both effects were also confirmed by the measured decision times. DMs make faster decisions for compatible rather than incompatible situations, confirming that it is easier to process information when the response mode is compatible with the format of the input information and there is no need to convert from one mode to the other. Also, compatibility effects are neither driven by differential features of the evaluation methods used (graphical ratings and pairwise comparisons were shown to be equally reliable) nor by the order in which these methods were used.

We found that incongruence induces evaluations that are more distant from the original recommendations. In other words, more information is lost in the communication when the DMs are forced to switch modalities. This provides unequivocal proof that, at least in this context, the inconsistency induced by incompatibility causes deterioration in the quality of participants' decisions. Interestingly, we showed that processing the information in a different modality is not necessary for the incompatibility effect to emerge—mere anticipation of a different modality is sufficient!

In the last study we showed that DMs do not anticipate the detrimental effects of stimulus–response incompatibility. This highlights the importance of optimal default setting for recommendation systems. It is well established that DMs tend to reverse their preferences across elicitation modes, and that the rate of inconsistencies varies as a function of certain features of the stimuli (e.g., Tversky et al., 1990). Previous research has documented the salience and primacy of measures of central locations. Obrecht et al. (2007) showed that in pairwise comparisons DMs rely more heavily on mean differences, using sample sizes as a secondary criterion, but attach little weight to the standard deviations. The DMs in our studies also showed the highest levels of inconsistencies for products with identical means but different variances and documented a clear monotonic relationship between the distance between mean recommendations and the rate of inconsistencies. We were able to determine that the rate of inconsistencies across the two tasks (ratings and pairwise choices) was 40% when the formats were compatible and 57% when they were incompatible. Thus, the net effect of format incompatibility is about 17%. We believe that this sizable effect calls for additional study.

Table 7
Distribution (in %) of Preferred Response Formats as a Function of the Stimulus Presentation Format (Study 3b)

	Preferred response format			
Stimulus presentation format	Numerical	Random	Verbal	Total
Numerical	78.8	14.1	7.1	930
Verbal	56.5	15.5	28.1	930
Total	67.6	14.8	17.6	1,860

[8] This analysis is based only on those participants who selected to answer, at least, once in each format for every presentation mode.

All these results are consistent with a simple model that combines two well established modes of information processing in the context of individual decision making—aggregation through averaging (e.g., Budescu, 2006, 2007; Clemen, 2008; Larrick & Soll, 2006) and classic stochastic choice processes (e.g., Bock & Jones, 1968; Luce, 1959; Thurstone, 1927). The averaging model is invoked to aggregate the multiple recommendations into single summary values (one for each product) that preserved the gist of the distribution (Reyna, 2012). These summaries are either translated directly into overt responses (in the rating task) or serve as the basic input for the choice task, which we assumed follows the Thurstonian model. The standard stochastic choice model predicts the distance effect and with a modest modification—higher variance attributable to the cross modality conversion (e.g., Jaffe-Katz et al., 1989)—it also anticipates the increase in inconsistencies in the incompatible choices.

Implications and Applications

Our finding that compatibility between presentation format and mode of processing matters begs the question what is the most prevalent information format that consumers encounter, and how they process it. The presentation format may vary systematically across sources. When consumers rely on the recommendations of friends, the information format is predominately verbal, because these recommendations are most likely communicated directly, and informally, and because this is most people's preferred mode of sharing (Wallsten et al., 1993). However, if consumers consult systematic surveys (e.g., www.consumerreports.org), the dominant information format seems to involve numerical ratings, and most Internet portals mix and combine the two formats.

It also seems reasonable to assume that the information format is a function of the products (see Dhar & Wertenbroch, 2000). One could imagine that for products with high hedonic appeal consumers search for confirmatory cues to justify their purchase intentions. These cues are more likely to be verbal because words are more vague, ambiguous, and elastic than numbers (Budescu, Weinberg & Wallsten, 1988; Piercey, 2009; Schweitzer & Hsee, 2002; Windschitl & Weber, 1999). On the other hand, for products with high utilitarian appeal consumers may prefer numerical information, which is more precise and more likely to help identify the best product. Here consumers might be less emotionally involved with a specific, implicitly favored, product. And, of course, these two factors may interact in the sense that one is likely to rely more on verbal advice from friends for products with high hedonic appeal (e.g., entertainment), and gravitate toward formal, data driven numerical ratings for utilitarian products (e.g., appliances).

Our findings have implications for the design of recommendation systems and Web portals (see also Ho & Quinn, 2008) and suggest that the best architecture—one that could help consumers make better, that is, more consistent, choices—would tailor the presentation mode of recommendations (numerical or verbal) to the expected recommendation format of the consumers. Olson and Budescu (1997) and, more recently, Du, Budescu, Shelley and Omer (2011) have shown, in different contexts, that DMs' preferences for information format are driven, at least, in part by their expectations about the nature of the target events. For example, contrary to the commonly held view that people always prefer precise information, Du et al. (2011) found that people prefer, and value more, financial forecasts involving some level of imprecision (e.g., relatively narrow ranges) over precise forecasts. The former were judged to be more informative and credible and induced higher confidence, reflecting the subjects' perception that overly precise (quarterly or annual) financial forecasts are unlikely to be accurate. Du et al. (2011) argued that DMs favor formats that are consistent with their perceptions of, and expectations regarding, the target events.

We speculate that a similar principle applies here and that users would make the best (most consistent) choices based on information that fits the nature of the product or service. For instance, if a consumer is looking to buy an appliance, precise information (i.e., numerical information in the form of ratings) might be more useful. If, on the other hand, a consumer is looking for advice regarding a book or a movie, then contextually "richer" information (i.e., verbal recommendations) might be advantageous.

References

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science, 19,* 392–398. doi:10.1111/j.1467-9280.2008.02098.x

Anderson, N. H. (1991). *Information integration theory*. Hillsdale, NJ: Erlbaum.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12,* 157–162. doi:10.1111/1467-9280.00327

Azen, R., & Budescu, D. V. (2003). Dominance analysis: A method for comparing predictors in multiple regression. *Psychological Methods, 8,* 129–148. doi:10.1037/1082-989X.8.2.129

Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting, 1,* 257–269. doi:10.1002/for.3980010305

Bock, R. D., & Jones, L. V. (1968). *The measurement and prediction of judgment and choice*. San Francisco, CA: Holden–Day.

Brainerd, C. J., & Reyna, V. F. (2001). Fuzzy-trace theory: Dual processes in memory, reasoning, and cognitive neuroscience. *Advances in Child Development and Behavior, 28,* 41–100.

Brainerd, C. J., Reyna, V. F., & Kneer, R. (1995). False-recognition reversal: When is similarity distinctive? *Journal of Memory and Language, 34,* 157–185. doi:10.1006/jmla.1995.1008

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context dependent, or both? *Organizational Behavior and Human Decision Processes, 41,* 390–404. doi:10.1016/0749-5978(88)90036-2

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin, 114,* 542–551. doi:10.1037/0033-2909.114.3.542

Budescu, D. V. (2006). Confidence in aggregation of opinions from multiple sources. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 327–352). Cambridge, UK: Cambridge University Press.

Budescu, D. V., Rantilla, A. K., Yu, H., & Karelitz, T. M. (2003). The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes, 90,* 178–194. doi:10.1006/jmla.1995.1008

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 281–294. doi:10.1037/0096-1523.14.2.281

Budescu, D. V., & Yu, H.-T. (2006). To Bayes not to Bayes? A comparison of two classes of models of information aggregation. *Decision Analysis, 3,* 145–162. doi:10.1287/deca.1060.0074

Budescu, D. V., & Yu, H.-T. (2007). Aggregation of opinions based on correlated cues and advisors. *Journal of Behavioral Decision Making, 20,* 153–177. doi:10.1002/bdm.547

Busemeyer, J. R., (1991). Intuitive statistical estimation. In N. H. Anderson (Ed.), *Contributions to information integration theory: Vol. 1. Cognition* (pp. 187–215). Hillsdale, NJ: Erlbaum.

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor production framework. *Journal of Risk and Uncertainty, 19,* 7–42. doi:10.1023/A:1007850605129

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research, 43,* 345–354. doi:10.1509/jmkr.43.3.345

Chiao, J. Y., Bordeau, A. R., & Ambady, N. (2004). Mental representations of social status. *Cognition, 93,* B49–B57. doi:10.1016/j.cognition.2003.07.008

Clemen, R. T. (2008). Comment on Cooke's classic method. *Reliability and System Safety, 93,* 760–765. doi:10.1016/j.ress.2008.02.003

Cubitt, R. P., Muro, A., & Starmer, C. (2004). Testing explanations of preference reversal. *Economic Journal, 114,* 709–726. doi:10.1111/j.1468-0297.2004.00238.x

Dhar, R., & Wertenbroch, K. (2000). Consumer choice between hedonic and utilitarian goods. *Journal of Marketing Research, 37,* 60–71. doi:10.1509/jmkr.37.1.60.18718

Du, N., Budescu, D. V., Shelley, M., & Omer, T. C. (2011). The appeal of vague financial forecasts. *Organizational Behavior and Human Decision Processes, 114,* 179–189. doi:10.1016/j.obhdp.2010.10.005

Fitts, P. M., & Deininger, R. L. (1954). S-R compatibility: Correspondence among paired elements within stimulus and response codes. *Journal of Experimental Psychology, 48,* 483–492. doi:10.1037/h0054967

Fitts, P. M., & Seeger, C. M. (1953). S-R compatibility: Spatial characteristics of stimulus and response codes. *Journal of Experimental Psychology, 46,* 199–210. doi:10.1037/h0062827

Gershoff, A. D., Mukherjee, A., & Mukhopadhyay, A. (2003). Consumer acceptance of online agent advice: Extremity and positivity effects. *Journal of Consumer Psychology, 13,* 161–170.

González-Vallejo, C., & Wallsten, T. (1992). Effects of probability mode on preference reversals. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 855–864. doi:10.1037/0278-7393.18.4.855

Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expressions over time. *Journal of Vision, 9,* 1–13. doi:10.1167/9.11.1

Ho, D. E., & Quinn, K. M. (2008). Improving the presentation and interpretation of online ratings data with model-based figures. *The American Statistician, 62,* 279–288. doi:10.1198/000313008X366145

Huber, J., & Puto, C. (1983). Market boundaries and product choice: Illustrating attraction and substitution effects. *Journal of Consumer Research, 10,* 31–44. doi:10.1086/208943

Jaffe-Katz, A., Budescu, D. V., & Wallsten, T. S. (1989). Timed magnitude comparisons of numerical and nonnumerical expressions of uncertainty. *Memory & Cognition, 17,* 249–264. doi:10.3758/BF03198463

Johnson, J. G., & Busemeyer, J. R. (2005). A dynamic, stochastic, computational model of preference reversal phenomena. *Psychological Review, 112,* 841–861. doi:10.1037/0033-295X.112.4.841

Kareev, Y., Arnon, S., & Horwitz-Zeliger, R. (2002). On the misperception of variability. *Journal of Experimental Psychology: General, 131,* 287–297. doi:10.1037/0096-3445.131.2.287

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science, 52,* 111–127. doi:10.1287/mnsc.1050.0459

Lee, L., Bertini, M., & Ariely, D. (2010). *Price muddles thinking: The effect of price on preference consistency*. *Working Paper,* Columbia University.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preferences between bids and choices in gambling decisions. *Journal of Experimental Psychology, 89,* 46–55. doi:10.1037/h0031207

Loomes, G., Pinto-Prades, J. L., Abellan-Perpinan, J. M., & Rodriguez-Miguez, E. (2010). *Modelling noise and imprecision in individual decisions. Working Papers 10.03, Universidad Pablo de Olavide, Department of Economics*. Retrieved June 20, 2010 from http://ideas.repec.org/p/pab/wpaper/10.03.html

Luce, R. D. (1959). *Individual choice behavior.* Oxford, Wiley.

Maki, R. (1981). Categorization and distance effects with spatial linear orders. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 15–32. doi:10.1037/0278-7393.7.1.15

Mas-Collel, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic theory*. New York, NY: Oxford University Press.

Moyer, R. S., & Bayer, R. H. (1976). Mental comparison and the symbolic distance effect. *Cognitive Psychology, 8,* 228–246. doi:10.1016/0010-0285(76)90025-6

Nowlis, S. M., & Simonson, I. (1997). Attribute-task compatibility as a determinant of consumer preference reversals. *Journal of Marketing Research, 34,* 205–218. doi:10.2307/3151859

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2007). Intuitive t tests: Lay use of statistical information. *Psychonomic Bulletin & Review, 14,* 1147–1152. doi:10.3758/BF03193104

Olson, M., & Budescu, D. V. (1997). Patterns of preferences for numerical and verbal probabilities. *Journal of Behavioral Decision Making, 10,* 117–131. doi:10.1002/(SICI)1099-0771(199706)10:2<117::AID-BDM251>3.0.CO;2-7

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences, 12,* 237–241. doi:10.1016/j.tics.2008.02.014

Paivio, A. (1975). Perceptual comparisons through the mind's eye. *Memory & Cognition, 3,* 635–647. doi:10.3758/BF03198229

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68,* 29–46. doi:10.1037/h0024722

Piercey, M. D. (2009). Motivated reasoning and verbal vs. numerical probability assessment: Evidence from an accounting context. *Organizational Behavior and Human Decision Processes, 108,* 330–341.

Por, H -H., & Budescu, D. V. (2012). Revisiting the gain–loss separability assumption in Prospect Theory. *Journal of Behavioral Decision Making*. doi:10.1002/bdm.1765

Reyna, V. F. (2008). A theory of medical decision making and health: Fuzzy trace theory. *Medical Decision Making, 28,* 850–865. doi:10.1177/0272989X08327066

Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision Making, 7,* 332–359.

Reyna, V. F., & Brainerd, C. J. (1991). Fuzzy-trace theory and framing effects in choice: Gist extraction, truncation, and conversion. *Journal of Behavioral Decision Making, 4,* 249–262. doi:10.1002/bdm.3960040403

Roselius, T. (1971). Consumer rankings of risk reduction methods. *Journal of Marketing, 35,* 56–61. doi:10.2307/1250565

Rubaltelli, E., Dickert, S., & Slovic, P. (2012). Response mode, compatibility, and dual-processes in the evaluation of simple gambles: An eye-tracking investigation. *Judgment and Decision Making, 7,* 427–440.

Schwarz, N. (1998). Accessible content and accessibility experiences: The interplay of declarative and experiential information in judgment. *Personality and Social Psychology Review, 2,* 87–99. doi:10.1207/s15327957pspr0202_2

Schweitzer, M. E., & Hsee, C. K. (2002). Stretching the truth: Elastic justification and motivated communication of uncertain information. *Journal of Risk and Uncertainty, 25,* 185–201. doi:10.1023/A:1020647814263

Shafir, E. (1995). Compatibility in cognition and decision. In J. R. Busemeyer, R. Hastie, and D. L. Medin (Eds.), *Decision making from the*

perspective of cognitive psychology (The psychology of learning and motivation, Vol. 32, pp. 247–274). New York, NY: Academic Press.

Simon, J. R., & Rudell, A. P. (1967). Audiotry S-R compatibility: The effect of irrelevant cue on information processing. *Journal of Applied Psychology, 51,* 300–304. doi:10.1037/h0020586

Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research, 16,* 158–174. doi:10.1086/209205

Slovic, P., & McPhillamy, D. (1974). Dimensional commensurability and cue utilization in comparative judgment. *Organizational Behavior & Human Performance, 11,* 172–194. doi:10.1016/0030-5073(74)90013-0

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34,* 273–286. doi:10.1037/h0070288

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review, 95,* 371–384. doi:10.1037/0033-295X.95.3.371

Tversky, A., Slovic, P., & Kahneman, D. (1990). The causes of preference reversals. *American Economic Review, 80,* 204–217.

Viswanathan, M., & Narayanan, S. (1994). Comparative judgments of numerical and verbal attribute labels. *Journal of Consumer Psychology, 3,* 79–101. doi:10.1016/S1057-7408(08)80029-0

Wallsten, T. S., Budescu, D. V., & Tsao, C. J. (1997). Combining linguistic probabilities. *Psychologische Beiträge, 39,* 27–55.

Wallsten, T. S., Budescu, D. V., Zwick, R., & Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society, 31,* 135–138.

Watamaniuk, S. N. J., & Duchon, A. (1992). The human visual-system averages speed information. *Vision Research, 32,* 931–941. doi:10.1016/0042-6989(92)90036-I

West, P. M., & Broniarczyk, S. M. (1998). Integrating multiple opinions: The role of aspiration level on consumer response to critic consensus. *Journal of Consumer Research, 25,* 38–51. doi:10.1086/209525

Windschitl, P. D., & Weber, E. U. (1999). The interpretation of "likely" depends on the context, but "70%" is 70% - Right? The influence of associative processes on perceived certainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1514–1533. doi:10.1037/0278-7393.25.6.1514

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2,* 343–364. doi:10.1037/1076-898X.2.4.343

Yamaguchi, M., & Proctor, R. W. (2006). Stimulus–response compatibility with pure and mixed mappings in a flight task environment. *Journal of Experimental Psychology: Applied, 12,* 207–222. doi:10.1037/1076-898X.12.4.207

Received August 16, 2011
Revision received February 11, 2013
Accepted March 21, 2013 ■