Original Article

Optimizing stimuli order in marketing experiments: A comparison of four orders using six criteria

Received (in revised form): 17th July 2009

Bryan Lilly
is Professor of Marketing and Associate Dean at the College of Business at University of Wisconsin Oshkosh. His research interests include consumer decision making, research methods and the marketing of new products. He consults with companies in areas of product development and marketing strategy.

ABSTRACT For consumer experiments, if stimuli are balanced across subjects, what order of stimuli within subjects is optimal? In this study, four stimulus orders were tested within subjects, all balanced across subjects. Stimulus orders were tested in a similarity judgment task and were evaluated based on six criteria. The best overall performance was achieved from a non-balanced order, wherein subjects saw all stimuli first and judged similarities as desired. A commonly used within-subject balanced order performed relatively worse, and was no better than a stimulus order proposed over 100 years ago.

Journal of Targeting, Measurement and Analysis for Marketing (2009) **17,** 245–255. doi:10.1057/jt.2009.17;
published online 7 September 2009

Keywords: order effects; within subjects; randomization; evaluability; range effect; asymmetric dependency

INTRODUCTION

Within-subject experimentation is quite common in marketing, and has been used by practitioners and academics for decades. In a within-subject experiment, each subject provides multiple responses. Malcolm Gladwell[1] describes famous marketing taste test experiments of this nature that shaped major industry practices in product categories ranging from butter (from the 1940s) to soda (the Pepsi Challenge in the 1980s). His work also discusses changes in industry practice following experiments that investigated consumer reactions to product and package form, for example studies of bottle shapes for alcoholic beverages. Regarding academic research in marketing, over 30 per cent of recently published consumer studies involved experimentation, and the majority involved within-subject designs.

Order effects are often viewed as a concern in experimental research. Order effects are changes in subjects' evaluations of stimuli as a result of the position of a stimuli relative to other stimuli considered.[2] For example, if a consumer prefers one of three brands, and the preferred brand was the first brand considered, then the concern is whether the consumer preference is a result of the brand, or of the fact that it was the first brand considered. If order effects exist and are not considered, then stimulus order may be a confounding factor, violating internal validity and leading to misinterpretation of results. When discussing results from studies, a typical reporting practice is to explain possible order effects, and how the research design was

Correspondence: Bryan Lilly
University of Wisconsin Oshkosh, 800 Algoma Boulevard,
Oshkosh WI 54901, USA
E-mail: lilly@uwosh.edu

constructed to minimize these effects. Recent examples include a study of consumers' sensations that combine visual and tactile cues where order effects were explored as a preliminary step in data analysis,[3] a study of odd-even pricing in which subjects were presented with multiple orders in a randomized fashion to reduce potential effects,[4] and a study in which orders were balanced across subjects in an examination of reserve prices set by sellers in auctions.[5] From the examples above, one can see that marketers' concerns for order effects transcend industries and types of research.

The objective of this research is to provide recommendations for how researchers should present the order of stimuli to reduce order effects. Although this study examines order effects in a within-subject experimental context, order effects occur in surveys and between-subject experiments for similar reasons.[6] Thus, although the most immediate generalization of this study is to within-subject experimental research (which is fairly common), the study also aims at adding to the debate and knowledge of order effects in general.

WHY ORDER EFFECTS OCCUR AND HOW TO REDUCE THEM

Order effects may have multiple causes, and occur most readily when consumers make evaluations for products with which they have low (versus high) familiarity.[7] The tendency for consumers to be biased by order effects under low-familiarity conditions indicates that marketers should be especially thoughtful about order effects when gauging consumer reactions to new products, for which familiarity is quite low. One main reason that order effects exist is that consumers weigh information more highly if received early (primacy effect) or late (recency effect). An expectation of primacy-based order effects is consistent with Prospect Theory,[8] as information received first leads to the development of a reference point, by which information received later is compared. For example, primacy-engendered order effects were found in the study of product evaluations.[9] And in a study of advertising, primacy-based order effects where shown to become more pronounced, as processing entails more effortful elaboration of message content.[10]

Order effects also occur because some tasks may be more difficult to complete than others. When presenting subjects with multiple tasks in a row that are all fairly difficult, they may become fatigued and produce less accurate judgments. For example, in a study of humor in television advertising, subjects were given short breaks between tasks to lessen fatigue.[11] Task difficulty can also arise when evaluating pairs of brands, where some brand-pairs are more difficult to evaluate than others. This difficulty may occur when consumers try to compare brands that have different features. For example, in a study that examined how consumers weigh brand features when judging brand similarity, a lack of comparability led to instances of intransitivity.[12] Task difficulty can occur when subjects evaluate contrasts between stimuli that are sometimes high and other times low. For example, in a study of service quality, variation in stimuli order led to assimilation and contrasts effects.[13]

Perhaps the most common recommendation with regard to reducing order effects is to balance the order across subjects. In a balanced order, stimuli are presented to subjects so that any given stimulus is presented before other stimuli just as frequently as it is presented after other stimuli. Great importance is placed on balance, as evidenced by the development of various techniques used to achieve this. For example Latin Square and Graeco-Latin experimental designs allow marketers balance stimuli order across multiple treatments, and is named after an ancient puzzle in which letters are arranged to balance (the Latin Square concept is employed with the now-popular Sudoku puzzle, using numbers). Another common practice with stimuli order is to randomize the order. The logic behind randomizing is that randomization helps achieve balance. Therefore, the prescription to randomize stimuli further supports the notion that balance is viewed as being important. Recent examples of balancing and randomization used to address order effects can be found in studies on event sponsorship,[14] pricing,[15] product development,[16] advertising,[17] consumer decisions on retail store travel,[18] consumer reactions to product claims[19] and consumer reactions to attribute-framing information.[20]

A balanced order is often beneficial but not completely satisfactory. Specifically, three order effects occur even in the presence of balance, and may even be prompted by a balanced order. Therefore, balancing may solve some problems associated with order effects, may not solve some problems, and may create some problems. One order effect insufficiently addressed by balancing may be caused by range differences, meaning that the similarity across stimuli range from small to large for a specific feature. Psychologists have long recognized and extensively discussed the 'range effect' as being insufficiently addressed through balancing.[21,22] Depending on the context to which the research results are to be generalized, the most valid order might entail first presenting stimuli that are very similar or stimuli that are very different, whereas a balanced order accomplishes neither.

A second order effect insufficiently addressed by balancing is asymmetric dependency. This occurs when a consumer's response to one stimulus biases a response to a second stimulus, but the response to the second stimulus does not bias the response to the first. Asymmetric dependency has been studied in social psychology contexts and has been linked to anchoring and causal mechanisms.[23] As an example from social psychology that could be analogous to a marketing situation, consider marital and life happiness. People often view marital happiness as a determinant of life happiness. Thus, in a study that involved assessments of marital and life happiness, valid assessments were obtained when participants considered life happiness first; their subsequent assessments of marital happiness were unaffected by their prior reporting of life happiness. However, when participants considered marital happiness first, their resulting assessments of life happiness were greatly affected.[24] As an analogy in a marketing context, consumers who travel may be happy with their accommodations and their trip overall, and happiness with accommodations may be a determinant of overall trip happiness. Accordingly, from the above we would expect valid assessments to occur by asking consumers about overall trip happiness first, whereas we would expect a biased assessment to occur if we asked consumers about accommodation happiness first. Where asymmetric dependency exists, balancing order entails multiple orders in which some orders create bias not counterbalanced by a bias from other orders.

A third order effect insufficiently addressed by balancing is fatigue, meaning weariness that occurs after exertion. Studies have shown that a balanced order may be more cognitively demanding to subjects than other designs, leading to fatigue and higher response times, and correspondingly less judgment accuracy.[25]

STIMULI ORDERS: WHICH ORDERS TO TEST AND HOW TO TEST THEM

A main motivation for this study was a realization that within-subject experiments could entail a non-balanced stimulus order presented to each subject, but one that is balanced across subjects. Perhaps the benefits of balancing may be sufficiently achieved by designing some level of balance either within *or* across subjects. If so, then designing studies to achieve balance within *and* across may not be needed, and may even be detrimental. From the discussion above, a stimulus order that entails balance across subjects and some other non-balanced sequence within subjects might be preferred.

To test different stimuli orders, a task was selected that has subjects provide paired judgments, indicating the degree to which two items are similar. Stimuli are balanced across subjects, and then different stimuli orders are tested within subjects. The similarity judgments are then evaluated with multidimensional scaling (MDS) analysis, to evaluate the impact of balanced versus non-balance orders.

Four reasons motivated the decision to test stimuli order using similarity judgments and MDS analysis. First and most importantly, consumer judgments of similarity are the building blocks for differentiation, which is important to marketers in practice. For example, in new product research, commercially available software packages exist that let marketers use subject judgments of similarity to quantitatively predict new product market shares and break-even sales amounts. Biased responses received from consumers could encourage marketers to develop new products that have insufficient potential, or to forgo new products that have blockbuster potential. Examining similarity judgments in this study therefore has great immediate usefulness. Second, stimuli pairs can be constructed to possess objectively measurable degrees of similarity. Judgments arising from different stimuli orders can then be assessed in terms of how well they match objective similarities. This type of objective validation would be impossible to achieve with affective judgments, which are also commonly used in practice. Third, several stimuli order recommendations have been proposed for similarity judgments, providing useful orders to compare, and these proposed stimuli orders have not been tested against each other empirically. Fourth, psychological theory about comparison judgments is well developed,[26] and the use of comparison data in developing perceptual maps is also well rooted in the literature.[27]

Stimulus orders tested

Four stimuli orders were tested in this study: one per group of subjects. The stimuli orders reflect four different properties. The properties and corresponding stimuli orders are described below, and hereafter are referred to as the balancing group, stationary group, flexible focus group and evaluability group.

The balancing property is characteristic in the first stimuli order tested, and seems to be widely accepted as a 'best practice'. For similarity judgments, balance stimuli orders have been developed by Ross[28] for sets that range from five objects to seventeen objects. For example, if a stimulus set comprises six brands, then each brand is judged five times: once with each other brand. For these six brands, 15 brand-pairs would be judged. Ross's order for six stimuli, referred to as 1–6, is (1934, p. 380) 1–2, 6–4, 5–1, 3–2, 5–6, 1–3, 2–4, 6–1, 4–3, 5–2, 1–4, 3–5, 2–6, 4–5, 3–6. Ross's prescribed orders continue to be used, for example in a study of honesty perceptions.[29]

The stationary property is characteristic in the second stimuli order tested, and involves taking one brand, making all possible comparisons between this brand and other brands, and then moving to the next brand. This stimulus order has been attributed to Kulpe, and was critiqued by Ross. For a set of six brands, numbered 1, 2, 3, 4, 5 and 6, the ordering prescribed by Kulpe was 1–2, 1–3, 1–4, 1–5, 1–6, 2–3, 2–4, 2–5, 2–6, 3–4, 3–5, 3–6, 4–5, 4–6, 5–6. Thus, attention is fixed or stationary for one brand that is completely judged, and then attention is stationary for a next brand that is completely judged, and so on. Kulpe's method allows subjects to maintain focus on a brand, and may enable them to be more cognitively adept in judging each brand.

The flexible-focus property is characteristic in the third stimuli order tested, and was proposed more recently.[30] With this approach, all brands are available to consider simultaneously. Subjects may refocus their attention as desired, in whatever way best allows them to complete their judgments. Thus, for a set of six brands, subjects complete 15 similarity judgments, and the order of these judgments is not predetermined. Two benefits are associated with this stimuli presentation technique. First, some judgments may be made more accurately if deferred until other judgments are made, which is possible with this technique. For example, researchers have found that simultaneous consideration of all stimuli may reduce or eliminate contrast effects.[13] Second, this technique implicitly encourages subjects to revise initial judgments. That is, because all brand-pairs remain available for review, subjects may be likely to 'correct' a judgment based on a subsequent judgment, and thus submitted judgments reflect revisions. Of course, this potential to judge and then later re-judge can be infused with all stimuli orders, which is addressed as part of the empirical testing described below (ultimately all stimuli orders tested included a second phase in which subjects re-judged initial scores, so that the issue of score revisions could be teased out empirically).

The evaluability property was explored by Hsee,[31] and is characteristic in the fourth stimuli order tested. Hsee explored the notion that overall evaluations may be affected by the ability to detect variation in the attribute. Evaluability refers to a reduced ability to compare stimuli as a result of not recognizing feature differences (a type of range effect discussed earlier). Hsee's initial inquiry involved music dictionaries, and subjects were at first unaware of how many musical terms were defined in each dictionary. Hsee's work involved only two objects, presented either separately or together, but this comparability issue is easily extended to stimulus order involving more than two stimuli. In testing six stimuli, variation in an attribute could be constrained so that initial comparisons would only show object pairs in which a focal attribute does not vary. To test whether the bias resulting from evaluability extends to stimulus ordering in this manner, a stimulus order was developed that maximally delayed the ability to assess variation on one dimension.

Criteria used to compare stimulus orders

The four stimulus orders described above were compared in six ways: scale, configuration, distances, time, revisions and qualitative feedback. The first three comparisons involve comparing subject data to objective MDS analysis results. Specifically, the six stimuli tested were constructed to differ from each other across three features in objectively quantifiable amounts, allowing the generation of a three-dimensional map of the stimuli based on objective differences. A three-dimensional perceptual map was also generated based on subject judgments for each stimuli order tested. Mapped solutions were then evaluated, comparing the map based on objective stimuli differences to maps based on perceived stimuli differences.

Scale. Scale refers to the overall size of the map across all dimensions. That is, the objectively produced map may appear to have the same configuration of stimuli as the map produced by subjective judgments, and yet the two maps may have different overall sizes, similar to how a model drawn in centimeters would differ from a model drawn in meters. MDS output provides a measure of scale discrepancy that can be expressed in standard deviations. If the scale discrepancy is zero, then the two maps have the same overall size. A stimuli order that leads to a lower scale discrepancy is preferred to an order that leads to a higher scale discrepancy.

Configuration. Each stimulus has location coordinates on three dimensions. Each map based on subjective judgments was scaled to have the same overall size as the map based on objective attributes differences. Euclidean distances between the locations of each stimulus on the subjective versus objective map were then measured. For any one stimulus, a distance of zero would indicate that a stimulus was identically located on both maps. The configuration measure is the sum of all six stimuli differences, reflecting the degree to which the set of stimuli coordinates matches across the objectively based and subjectively based maps. A stimuli order that leads to a lower configuration difference is preferred to an order that leads to a higher configuration difference.

Distances. Distance correlations were then evaluated, which are distances between stimuli produced by a map based on subject judgments correlated to distances between stimuli produced by a map based on objective stimuli differences. A stimulus order that generates data with a higher correlation is preferred to a stimulus order that generates data with a lower correlation.

Time. Accuracy may be related to time: a confusing stimuli order could require more time and lead to less accurate results. The reverse could also apply: an overly simplified stimuli order could engender less time and less accuracy. However, if accuracy is even across stimuli orders (as ascertained by the first three criteria), then the stimuli order that requires less time from subjects is generally considered preferable.[25]

Revisions. After subjects provide all judgments, do they feel that their own judgments were accurate? Allowing subjects to revise initial judgments is one way to discern whether they

Table 1: Attribute levels for the six dictionary descriptions compared

Dictionary	*Number of entries*	*Number of sheet music pages*	*Condition (has defects, yes/no)*	*Year*
1	10 000	150	No	2003
2	20 000	150	No	2003
3	15 000	50	No	2003
4	15 000	200	Yes	2003
5	20 000	10	Yes	2003
6	10 000	10	Yes	2003

felt that they were accurate. A stimuli order that leads to fewer and smaller revisions suggests a better ability to provide initial judgments, relative to a stimuli order that leads to a greater number of revisions and revisions that are larger in magnitude.

Qualitative feedback. Finally, another way to evaluate whether subjects feel that their judgments are accurate is by asking them, and thereby collecting qualitative data about how they judged stimuli, how their judgment process evolved, and why revisions were needed. Qualitative statements may also help to understand whether/why subjects felt that their initial judgments were correct.

EMPIRICAL TEST

A variation of stimuli tested by Hsee was used, to create stimuli fairly similar to that used in that prior study. Hsee had subjects compare music dictionary descriptions that differed on two attributes: condition of dictionary and number of entries in the dictionary. In this study, a third attribute was added: the number of sheet music pages supplied with the dictionary. Adding this attribute allowed the presentation of brand-pairs in a manner that delays variation on one attribute (to test the fourth stimulus condition). Table 1 lists the combinations of attributes used for each of the dictionary descriptions.

Stimuli orders were balanced across subjects in all conditions. This balancing was easy to achieve for all groups except the group that also involved a balance order within subjects. For this group, Ross's prescribed stimulus order was updated to achieve balance both within and across subjects. Power tests determined the need for a minimum sample size of 28 subjects per data-ordering method to make the results statistically reliable, and a target of 30 subjects per condition was set. With this target, a design was developed that achieves complete balance across and within 15 subjects (see Table 2). As can be seen in this table, 15 pairs appear, and each number (1–15) appears exactly once in each column and once in each row. Pairs of dictionary descriptions were presented on cards, and each pair described one dictionary on the top of the card and one dictionary on the bottom of the card. To balance across 30 subjects, the top and bottom descriptions were reversed for subjects 16–30.

Data were collected from 136 undergraduate student subjects, with 34 subjects in each of the four conditions. This slightly exceeded the targeted 30-subject sample size desired for statistical reasons, and in the balanced group subjects 31–34 received stimuli in the same order as received by subjects 1–4, and thus the balance cycle started over for these subjects.

Cards with pairs of descriptions were arranged according to the four stimulus order conditions. For the balanced, stationary and evaluability groups, the subjects were told to progress through the cards in the order in which they were received. The flexible focus group was the simultaneous condition, and the subjects in this group reviewed all pairs before making any judgments, and then determined for themselves which pairs to judge first. Following recommendations by Borg and Groenen (p. 93),[25] a practice set of ratings was used to familiarize subjects with the task. A 20-point scale was used for ratings. Each subject recorded the time they started and finished their initial ratings. After the subjects finished providing initial similarity judgments, qualitative data were collected. Specifically, subjects were asked to

> Notice the music dictionaries were all published in year 2003. But the dictionaries differ in three

Table 2: Pattern of stimuli presentation for the balance group

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
1	8	12	10	11	3	4	14	9	6	13	5	7	2	15
2	15	1	8	12	10	11	3	4	14	9	6	13	5	7
3	4	14	9	6	13	5	7	2	15	1	8	12	10	11
4	14	9	6	13	5	7	2	15	1	8	12	10	11	3
5	7	2	15	1	8	12	10	11	3	4	14	9	6	13
6	13	5	7	2	15	1	8	12	10	11	3	4	14	9
7	2	15	1	8	12	10	11	3	4	14	9	6	13	5
8	12	10	11	3	4	14	9	6	13	5	7	2	15	1
9	6	13	5	7	2	15	1	8	12	10	11	3	4	14
10	11	3	4	14	9	6	13	5	7	2	15	1	8	12
11	3	4	14	9	6	13	5	7	2	15	1	8	12	10
12	10	11	3	4	14	9	6	13	5	7	2	15	1	8
13	5	7	2	15	1	8	12	10	11	3	4	14	9	6
14	9	6	13	5	7	2	15	1	8	12	10	11	3	4
15	1	8	12	10	11	3	4	14	9	6	13	5	7	2

Notes: This pattern is based on Ross's prescription (1934). Row 1 lists 15 subjects. Subject 1 sees stimuli in the order presented by Ross. Brand-pairs are then further balanced across subjects. Thus, each brand-pair occurs exactly once in each row. Each brand-pair is described on a card, with one brand description at the top of the card and one at the bottom. For subjects 16–30, the same stimuli order is used, but with brand descriptions switched, so that the top-of-card descriptions become bottom-of-card descriptions.

> aspects: (1) number of entries, (2) number of sheet songs, and (3) condition of dictionary. In the space below, briefly describe how you rated the dictionaries. Were all three aspects equally important? Or was one aspect most critical, where similarity would be high or low mainly because of similarity on this aspect? Or was any aspect not used, so similarity judgments were not based on this aspect?

A space for qualitative comments was provided. After the subjects provided answers, they then arranged their brand-pair ratings so that they had 15 cards, arranged from pairs judged as being most similar to pairs judged as being least similar. The subjects were then asked to review their ratings and re-rate any similarity scores they felt could be improved. A different-colored pen was used, and thus later each subject submitted their 15 cards and each card had an initial rating and some cards had updated ratings from this second effort. This re-scoring corresponds to the fifth metric used to evaluate the stimulus orders: the number and severity of revisions. Fewer and less severe revisions reflect subjects who felt that their initial ratings were relatively accurate.

Finally, after the subjects re-rated pairs, they were asked, 'If you made any changes, please briefly indicate why you made changes … in other words, what did you miss the first time around that you are now fixing?' This final qualitative question was followed by a space for providing an answer. Therefore, in total the procedure was designed to gather (1) initial similarity ratings; (2) time used to make judgments; (3) updated ratings; and (4) qualitative comments about how similarity judgments were made and updated.

RESULTS

Before comparing stimuli orders, each map produced from subjective judgments was evaluated, to make sure that a three-dimensional solution fit the map better than a two-dimensional solution, which could occur if subjects ignored an attribute. Using the Bayesian Information Criterion (BIC) statistic,[32] for each map a three-dimensional solution indeed fit better than a two-dimensional solution. Progression to main tests occurred next, as discussed below.

Scale. The stimulus order that performed the best was the third group tested: the flexible focus group. This group yielded the lowest standard deviation, 0.167, indicating that the overall size of the mapped solution created from using the flexible focus group judgments most closely matched the size of the mapped solution created from using objective stimuli differences. The next-best performing groups were the stationary group and the balanced group, with standard deviations of 0.221 and 0.226, respectively (consider as tied). The worst performing group was the evaluability group, which produced a standard deviation of 0.257, which is almost 54 per cent worse than the performance produced by the flexible focus group (0.257/0.167–1).

Configuration. The stimulus order that performed the best using the configuration was again the third group tested: the flexible focus group. The overall configuration difference between stimuli locations on the map produced from this group's judgments versus the map produced from objective attribute differences was 2.98. The next-best performing groups were the stationary group and the balanced group, with differences of 3.48 and 4.24, respectively. The worst performing group was the evaluability group, which produced a difference of 4.36.

Distances. The stimulus order that performed the best (highest correlation: $r=0.9148$) was achieved in the flexible focus group, in which subjects saw all stimuli before rating any. The stimulus order based on the balanced design achieved the next highest correlation: $r=0.8902$. The stimulus order based on the stationary property achieved the third highest correlation: $r=0.8849$. The lowest correlation, $r=0.8697$, was achieved with the last stimulus order, based on the evaluability property.

Time. Time variations across groups were not statistically significant (all P-values>0.05). Thus, sample times do not clearly support one stimulus order over another. However, the group requiring the least amount of time was the group that saw all brands before making any judgments, again finding best performance from the flexible focus group. The mean time used across subjects was 4.76 min to complete the 15 judgments. The next highest mean time was 4.94 min, used by subjects in the stationary group. The mean time was slightly higher in the balanced condition, at 5.11 min. The mean time in the final condition, 5.32, was highest, again finding the worst performance from the stimulus order in which variation on an attribute was delayed in being presented.

Table 3: Revisions across groups

Group (stimulus order)	*Percent of judgments revised*	*Mean absolute revision*
Balance	57.84	2.818
Stationary	59.61	2.820
Flexible focus	39.22	2.606
Evaluability	53.53	2.608

Revisions. After the subjects judged similarities, they were asked to put the 15 judged pairs in order, from most similar to least similar. The subjects were then asked to revise the judgments as needed. This test criterion is based on the notion that fewer and less severe revisions reflect a better stimulus order, because initial judgments were viewed as correct. Table 3 shows the results. Again, the stimulus order that achieves the best results is an order in which all brands are viewed first, based on the flexible focus property. This group had the fewest number of revisions: 39.22 per cent of all judgments were revised. This group had significantly fewer revisions than other groups ($P<0.05$). However, significant differences did not exist across the other groups, with the percent of judgments revised in other groups ranging from 53.53 per cent to 59.61 per cent. Many revisions were small. On the 20-point scale, quite a few revisions were simply one or two points. The rightmost column of Table 3 shows the mean absolute revision, only among judgments that were revised. Judgments that were not revised were omitted (with all judgments included, again the flexible focus group had the lowest mean revision because many judgments were not revised in that group). Mean absolute revisions are not significantly different across groups, ranging from a low of 2.606 in the flexible focus group to a high of 2.82 in the stationary group.

Qualitative comments. The subjects provided qualitative comments twice: once after finishing initial judgments and once after making revisions. Some comments were interesting in terms of data order, and reflect the ability (or inability) to evaluate dimensions accurately during the rating process. These qualitative comments were evaluated by two researchers, one of whom had not worked on other aspects of the project.

Table 4: Aggregate results

Group	*Scale objective/ subjective*	*Configuration objective/ subjective*	*Distances objective/ subjective*	*Time*	*Percent of judgments revised*	*Qualitative comments*
Balance	2.5	3	2	3	3	2.5
Stationary	2.5	2	3	2	4	2.5
Flexible focus	1	1	1	1	1	1
Evaluability	4	4	4	4	2	4

Notes: Four data-ordering methods were compared across six evaluative criteria. Four groups are listed in the leftmost column. In each subsequent column, they are ranked in terms of their relative performance. For example, for the scale test, the flexible focus group performed best (1 denotes first place), and the evaluability group performed the worst (4 denotes last place). In two instances performance was tied.

The conclusions were the same across the evaluators. Multiple subjects in the balance, stationary and evaluability groups commented on feeling that they should judge more 'harshly' as they progressed through the ratings. This did not occur in the flexible focus group, in which subjects saw all dictionary descriptions before the rating task. Four subjects in the evaluability group commented on the importance of dictionary condition changing during the initial rating process, but none of the subjects in the other groups commented on the importance of condition changing. This reflects the delay in seeing variation in the dictionary condition attribute that was unique to subjects in the evaluability group. In terms of the comments the subjects made after revising ratings, the comments from subjects in the flexible focus group reflect the low number of changes in that group. Compared to other groups, more comments were found in the flexible focus group about making only minimal changes. As might be expected, across all groups some subjects expressed a tendency to be more methodical (comments varied, but subjects used words and phrases such as methodical, being more consistent, and being more uniform in their judgments).

DISCUSSION

In practice, market researchers gathering consumer feedback may present brands to consumers in different orders. A common current practice is to strive for balance across the order of brands presented to consumers, evenly rotating the sequence presented to subjects. The aim of this study is to show that, when balance is maintained across subjects, balance applied within may produce worse results than results achieved by using other stimulus orders. Ultimately, marketers can improve the accuracy of their research results and corresponding managerial decisions by using non-balanced stimuli orders when engaging studies where each subject evaluates multiple brands.

Table 4 shows results across the four groups tested, and the six evaluation criteria. The overall 'best' stimulus order was achieved with the group based on the flexible focus property, in which subjects saw all brands before making any judgments. This group performed best on all six criteria. Further, it seems clear that the 'worst' stimulus order was achieved with group 4, based on extending Hsee's evaluability property, indicating that the bias found by Hsee is extendable to other applications. The groups based on the balance property and the stationary property performed about evenly.

A main result of this study is the calling into question of the need for balance within subjects, provided balance across subjects is maintained. The results indicate that advantages may be achieved by showing subjects all stimuli first, and subsequently asking for ratings. This approach allows subjects to focus attention in a flexible manner, so that judgments can be made in an order that seems 'correct' to each subject. The findings also show that valid results are achieved and require less time among subjects when stimuli are ordered based on the stationary property. Thus, in time-consuming contexts such as evaluating very novel or complex brands, if revealing all brands initially is impractical, this result indicates that marketers should use the non-balanced order prescribed by Kulpe. Another conclusion this study makes involves the importance of allowing consumers to understand new features early during an evaluation process. Although marketers certainly wish to avoid confusing consumers, the poor performance of subjects in the evaluability group suggests that market researchers should reveal new product attributes as quickly as possible to subjects, to increase the validity of consumer evaluations.

Regarding the limitations of this study, the six evaluative criteria produced non-uniform results. This means that the best stimuli order might be context-dependent, which was not addressed. Further, Table 4 presents overall results, but some results are equivocal. For example, the time metric did not yield statistically significant differences. Therefore, conclusions are made by gauging all results collectively, but some researchers might prefer to focus on one evaluative criterion more intensively.

Other limitations involve the context of the empirical work. Analysis was restricted to similarity judgments and MDS analysis. Similarity judgments allow us to use desirable testing criteria. Nevertheless, conclusions are most clearly warranted for applications that involve similarity judgments, which are very common in certain types of market research, such as positioning studies and new product studies.

Several extensions of this study are worth pursuing. First, emotional involvement and salience were unexplored in this study, but can bring about dramatic effects on consumer evaluations. For example, if consumer evaluations develop through an affective response, then it may be important to identify a presentation order of brands that produces affective responses similar to what consumers experience while shopping. Expanding the notion of optimal orders to contexts involving emotion and high product salience could lead to providing more complete prescriptions to research practitioners. A second extension could be applied to survey research methods. Respondents could be prompted to review all questions first, before answering any. Exploration in a survey context may lead to improved survey practices. A third extension is to more clearly understand contexts in which the evaluability property exists in practice and may lead to invalid data and wrong conclusions. A survey of research practitioners may be helpful, to at least gauge their understanding of this potential problem, and to determine whether practices exist that should be corrected. The fourth and fifth extensions of this study would be to test additional stimuli orders beyond the four tested here, and to improve the criteria used to compare stimuli orders. It is possible that some other stimuli orders exist, or that combinations of instructions could be found that further improve the accuracy of judgments.

REFERENCES

1 Gladwell, M. (2005) *Blink*. New York: Little, Brown and Company.

2 Mitchell, M. and Jolley, J. (2007) *Research Design Explained*, 6th edn. Belmont, CA: Wadsworth.

3 Krishna, A. (2006) Interaction of senses: The effect of vision versus touch on the elongation bias. *Journal of Consumer Research* 32(4): 557–566.

4 Thomas, M. and Morwitz, V. (2005) Penny wise and pound foolish: The left-digit effect in price cognition. *Journal of Consumer Research* 32(1): 54–64.

5 Greenleaf, E. (2004) Reserves, regret, and rejoicing in open English auctions. *Journal of Consumer Research* 31(2): 264–273.

6 Haugtvedt, C., Herr, P. and Kardes, F. (eds.) (2008) *Handbook of Consumer Psychology*. Boca Raton, FL: CRC Press.

7 Kumar, V. and Gaeth, G. (1991) Attribute order and product familiarity effects in decision tasks using conjoint analysis. *International Journal of Research in Marketing* 8(2): 113–124.

8 Kahneman, D. and Tversky, A. (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2): 263–291.

9 Scarpi, D. (2004) Effects of presentation order on product evaluation: An empirical analysis. *International Review of Retail, Distribution and Consumer Research* 14(3): 309–319.

10 Brunel, F. and Nelson, M. (2003) Message order effects and gender differences in advertising persuasion. *Journal of Advertising Research* 43(3): 330–341.

11 Woltman Elpers, J., Mukherjee, A. and Hoyer, W.D. (2004) Humor in television advertising: A moment-to-moment analysis. *Journal of Consumer Research* 31(3): 592–598.

12 Kivetz, R. and Simonson, I. (2000) The effects of incomplete information on consumer choice. *Journal of Marketing Research* 37(4): 427–448.

13 DeMoranville, C.W. and Bienstock, C. (2003) Question order effects in measuring service quality. *International Journal of Research in Marketing* 20(3): 217–231.

14 Johar, G.V. and Pham, M.T. (1999) Relatedness, prominence, and constructive sponsor indentification. *Journal of Marketing Research* 36(3): 299–312.

15 Janiszewski, C. and Cunha Jr., M. (2004) Within counterbalancing, the influence of price discount framing on the evaluation of a product bundle. *Journal of Consumer Research* 30(4): 534–546.

16 Forlani, D., Mullins, J.W. and Walker Jr, O.C. (2002) New product decision making: How chance and size of loss influence what marketing managers see and do. *Psychology and Marketing* 19(11): 957–981.

17 Appleton-Knapp, S.L., Bjork, R.A. and Wickens, T.D. (2005) Examining the spacing effect in advertising: Encoding variability, retrieval processes, and their interaction. *Journal of Consumer Research* 32(2): 266–276.

18 Brooks, C.M., Kaufmann, P.J. and Lichtenstein, D.R. (2004) Travel configuration on consumer trip-chained store choice. *Journal of Consumer Research* 31(2): 241–248.

19 Skurnik, I., Yoon, C., Park, D.C. and Schwarz, N. (2005) Balance and delay practice trials, how warnings about false claims become recommendations. *Journal of Consumer Research* 31(4): 713–724.

20 Janiszewski, C., Silk, T. and Cooke, A. (2003) Different scales for different frames: The role of subjective scales and experience in explaining attribute-framing effects. *Journal of Consumer Research* 30(3): 311–325.

21 Poulton, E.C. (1973) Unwanted range effects from using within-subject experimental designs. *Psychological Bulletin* 80(2): 113–121.

22 Poulton, E.C. (1975) Range effects in experiments on people. *American Journal of Psychology* 88(1): 3–32.

23 Strack, F., Martin, L. and Schwarz, N. (1988) Priming and communication: The social determinants of information use in judgments of life-satisfaction. *European Journal of Social Psychology* 18(5): 429–442.

24 Schwarz, N., Strack, F. and Mai, H. (1991) Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly* 55(Spring): 3–23.

25 Borg, I. and Groenen, P. (1997) *Modern Multidimensional Scaling: Theory and Applications*. New York: Springer-Verlag.

26 Thurstone, L.L. (1927) A law of comparative judgment. *Psychological Review* 34(4): 273–286.

27 Coombs, C.H. (1964) *A Theory of Data*. Ann Arbor, MI: Mathesis Press.

28 Ross, R.T. (1934) Optimum orders for the presentation of pairs in the method of paired comparisons. *The Journal of Educational Psychology* 25(5): 375–382.

29 Pincus, H.S. and Schmelkin, L.P. (2003) Faculty perceptions of academic dishonesty: A multidimensional scaling analysis. *The Journal of Higher Education* 74(2): 196–209.

30 Goldstone, R.L. (1994) An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers* 26(4): 381–386.

31 Hsee, C.K. (1996) The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes* 67(3): 247–257.

32 Schwartz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics* 6(2): 461–464.