Maximum versus Meaningful Discrimination in Scale Response: Implications for Validity of Measurement of Consumer Perceptions about Products

Madhubalan Viswanathan
Department of Business Administration, University of Illinois at Urbana, 61 Wohlers, MC 706, 1206 S. Sixth, Champaign, IL 61820, USA

Seymour Sudman
Department of Business Administration, University of Illinois at Urbana, 61 Wohlers, MC 706, 1206 S. Sixth, Champaign, IL 61820, USA

Michael Johnson
School of Business Administration, University of Michigan, 701 Tappan Street, Ann Arbor, MI, 48109-1234, USA

Abstract

This paper argues for the use of the number of response categories that are meaningful to respondents as a criterion in designing attribute rating scales in marketing in contrast to a focus in past research on using scales to maximize the discrimination elicited from respondents. Whereas scales eliciting a maximum level of discrimination may be more reliable than scales eliciting a meaningful level of discrimination, the latter are argued to be more valid in measuring sameness and difference between brands that are meaningful to respondents. Specifically, a distinction is drawn in this paper between the maximum number of categories that consumers can *discriminate* between and the number of categories that are *meaningful* to them. The meaningful number of categories refers to the number of categories that individuals typically use in thinking about an attribute in such situations as making a choice or judgment. Thus, the unique perspective of consumer behavior with its central focus on phenomena such as product judgment and choice is incorporated into the measurement of consumers' perceptions about attributes. Several studies were conducted to test hypotheses generated on the basis of the notion of the meaningful number of categories. The first study used an open-ended method (i.e., a sorting task) to measure the number of categories that are meaningful to consumers for specific attributes. Using the results of the first study as a basis, two studies demonstrated the effect of the meaningful number of categories for an attribute on scale response such that fewer scale points were used to rate products on attributes with fewer meaningful numbers of categories. Another study showed that a scale with the meaningful number of categories might be more accurate than other scales in predicting sameness and difference between brands that are meaningful to consumers. The significance of using scales with a meaningful number of categories is in validly measuring differences between products that are meaningful to consumers.

1. Introduction

A consumer completing a product survey with 7-point attribute rating scales rates a brand as a 6 out of 7 on a key attribute and another brand as a 7 out of 7. An issue of importance for the measurement of consumer perceptions is whether a difference between a 6 and a 7 on a rating scale really matters to this consumer for purposes of evaluating these brands. In practice, 7-point scales are usually employed without determining how customers typically think about specific attributes when judging products. The consumer may have discriminated between the two brands merely because of the need to respond using a 7-point scale. Perhaps, the consumer thought about the attribute in question using only three levels, i.e., high, medium, and low, and a 3-point scale should have been used. Then, both brands may have been rated a 3 out of 3 along the attribute in question. In this situation, a scale with an appropriate number of response categories may have captured sameness and differences between brands along attributes that were meaningful to consumers. As a result, valid inferences could have been drawn about the decision-making process of consumer for purposes of understanding and predicting product choice or judgment.

This example illustrates the importance of understanding how consumers think about product attributes, specifically in terms of the number of levels or magnitudes used by consumers to categorize an attribute continuum. Research outside of consumer behavior in the area of scaling has examined the issue of the optimal number of response categories to use in scales and focused on eliciting maximum levels of discriminations from respondents in keeping with the goals of enhancing reliability from a measurement perspective. This research stream, having developed in the general area of scaling rather than in the specific area of consumer behavior, has not examined how consumers actually categorize a product attribute with a view to using such insight to tailor scales for specific product attributes.

This paper argues for the use of the number of response categories that are meaningful to respondents as a criterion in designing attribute rating scales in marketing in contrast to a focus in past research on using scales to maximize the discrimination elicited from respondents. Whereas scales eliciting a maximum level of discrimination may be more reliable than scales eliciting a meaningful level of discrimination, the latter are argued to be more valid in measuring sameness and difference between brands that are meaningful to respondents. In contrast to a focus in past research in scaling on maximizing the discrimination elicited from consumers, a distinction is drawn in this paper between the maximum number of categories that consumers can *discriminate* between and the number of categories that are *meaningful* to them. The meaningful number of categories refers to the number of categories that individuals typically use in thinking about an attribute in such situations as making a choice or judgment. This distinction is import- ant because meaningful discriminations made by consumers form the basis for phenomena of interest such as choice and purchase behavior whereas maximum discriminations may merely be an artifact of the measurement scale.

From a measure development perspective, researchers have distinguished between subject- and stimulus-centered scales (Cox, 1980). Different types of validity can be evaluated for subject-centered scales employed to capture differences among individuals, as well as for stimulus-centered scales such as attribute ratings employed to capture differences between products. However, stimulus-centered scales such as attribute ratings, are often used to gauge sameness and difference between products whereas subject-centered scales are typically employed to place individuals along some continuum. A neglected issue in validity arises in such situations and is the focus of this paper, i.e., the validity of inferences of sameness and difference drawn from stimulus- centered scales with a certain number of response categories. This type of validity is distinct from validity established through correlational means to show covariation between scale ratings and other measures as evidence of predictive, nomological, convergent, or discriminant validity.

The approach taken in this paper was to (1) develop a method of determining the number of categories that are meaningful to consumers on specific product attributes (i.e., a sorting task), (2) test whether the number of meaningful categories for an attribute impacts responses to scales used to rate brands on that attribute, and (3) test whether a scale with a meaningful number of response categories is more accurate than other scales in predicting sameness and difference between brands that is meaningful to consumers. This research has important implications for research in marketing. Consumer perceptions based on rating scales are often used as diagnostic information in marketing research to make inferences about the decision-making process of consumers. From a methodological perspective, the valid measurement of consumers' perceptions of products along attributes is crucial for conducting research in marketing. For practitioners, as illustrated in the example, marketers need to know whether observed differences in ratings between their brand and competing brands or new formulations of brands would lead consumers to differentiate between brands or can safely be ignored. The paper is organized as follows. Following a discussion of relevant literature and development of hypotheses, four studies conducted to test the hypotheses are described. Finally, the implications of this research are discussed.

2. Literature review and conceptual framework

2.1 Review of past research on the optimal number of response categories

The literature on the optimal number of response categories to use in a scale has been characterized by several approaches including the information theoretic approach, metric approaches, and the absolute judgment paradigm (Cox, 1980). The information theoretic approach (cf. Garner and Hake, 1951) has related the number of response categories of a scale to the amount of information transmit- ted by it, with greater utilization of various response categories translating into higher information transmission by a scale. Information transmitted by a scale has been found to increase with an increase in

the number of response categories with some studies reporting that the amount of information transmitted levels off beyond a certain number of response categories in a scale (cf. Cox, 1980). Metric approaches to the optimal number of response categories have assessed the reliability of scales with different numbers of response categories. Churchill and Peter (1984) con- ducted a meta-analysis and found that the reliability of a scale increased with an increase in the number of response categories. However, some researchers have not found such a relationship (cf. Bendig, 1954). Research on absolute judgments has focused on perceptual judgments along dimensions such as loudness (Miller, 1956). This research has been cited as providing support for the use of approximately seven response categories in a scale due to a limit to human ability to discriminate beyond seven levels.

A common emphasis in past approaches has been on eliciting the *maximum level of reliable discrimination* from consumers. The implicit or explicit trade-off in each of these approaches has been in terms of the information generated by a scale for purposes of the researcher and limits to human ability to discriminate beyond a certain number of response categories. Researchers have pointed out that a scale with too few categories does not allow sufficient discrimination by consumers whereas a scale with too many categories may be beyond the consumers' ability to discriminate (cf. Komorita and Graham, 1965). This constraint is reflected by limits to increases in reliability with increases in the number of response categories in a scale in metric approaches, and by limits to the information generated by a scale in information theoretic approaches. Whereas a general rule about the number of response categories to use in a scale has been suggested (i.e., 7 ± 2 categories), researchers have pointed out that the optimal number of response categories could vary anywhere from 2 or 3 (Jacoby and Matell, 1971) to 25 (Champney and Marshall, 1939). In reviewing this area of research, Cox (1980) pointed out an immediate need to develop methods at the pretesting stage to evaluate the nature of information being collected using scales with different numbers of response categories.

The emphasis in past approaches on generating general rules for the optimal number of response categories to be used is consistent with the context of such research, which was in the general area of scaling. However, past research in scaling does not take into account the unique perspective of consumer behavior with its central focus on phenomena such as product judgment and choice. Consumers' perceptions along product attributes form the basis for higher level product decisions. Therefore, in research in marketing, attribute ratings are often used to understand and to predict product choice and judgment, an issue that forms the basis for the distinction between maximum and meaningful dis- crimination described next.

2.2 Conceptual framework

2.2.1 The distinction between maximum and meaningful discrimination

In this section, a distinction is drawn between the *maximum level to which respondents can discriminate* along some attribute continuum as emphasized in past research, and the *meaningful level of discrimination* along that continuum, followed by a discussion of the underlying assumptions of this distinction. The meaningful level of discrimination is used here to refer to the number of categories that individuals typically use in thinking about an attribute in situations involving the use of information on that product attribute in making a choice or judgment. This distinction is important in that meaningful discriminations made by respondents may underlie phenomena such as choice and purchase behavior whereas maximum discriminations may merely be an artifact of the measurement scale. More discriminating scales (i.e., scales with more response categories) may elicit finer discriminations across brands than would typically be employed by respondents in decision-making tasks involving attribute information. In this regard, researchers have pointed out the informative function served by response scales (cf. Sudman and Brad-burn, 1982; Matell and Jacoby, 1972). Considering the example at the beginning of the paper, if consumers con- sider three levels of an attribute to be meaningful to them (such as high, medium, and low), sameness and differences between brands on the 7-point scale may not provide *a valid basis for making inferences about the decision-making process of the respondent*. It should be noted that a scale that elicits the maximum level of discrimination might be more reliable than a scale that has fewer response categories and, hence, preferred for correlational analyses such as regressions. However, the latter may be more valid in measuring differences in perceptions that are meaningful to respondents. The measurement of meaningful discrimination is important for scales used to draw inferences that are not based on strength of correlations but on sameness or differences between ratings of stimuli (such as stimulus centered scales that measure differences between stimuli).

2.2.2 Assumptions of the distinction between meaningful and maximum discrimination

A key premise made in drawing the distinction between meaningful and maximum levels of discrimination is that respondents naturally use a stable number of categories in thinking about a specific attribute during decision-making (i.e., the meaningful number of categories), hence the need to measure it as an input in scale development. A related premise made here is that it is possible to *reliably* discriminate to a greater degree than is meaningful to respondents (i.e., meaningful discrimination is not the same as maximum discrimination), thereby making it possible for scales to elicit finer discriminations than may be meaningful to respondents. In comparison, scale responses are assumed to be based on the relatively spontaneous formation of a rating in response to an item and its response format (i.e., the scale). Consequently, responses to attribute ratings scales may be more or less discriminating depending on the number of response categories in the response scale. Each of these premises is examined later.

From a cognitive perspective, the maximum level to which consumers can discriminate may be a function of their knowledge of variations of products along attributes in the marketing environment. For example, consumers may be aware that two brands of breakfast cereals have calorie contents of 120 and 125 calories per serving, discriminating information that is often available to them. However, all variations that consumers are aware of among products may not necessarily be meaningful to them in decision-making, and consequently, may not translate into differences in the utility that consumers attach to those variations during decision-making. Rather, the level of discrimination that is meaningful to consumers may be related to the *usefulness* of variations in products for purposes of making a choice or judgment. Therefore, some consumers may consider both 120 and 125 calories to be "high" in calorie content for purposes of decision-making although they are aware of the difference in calorie content.

Research in consumer behavior and psychology provides support for these premises. Specifically, past research suggests that consumers may classify information along a dimension by deriving cognitive categories from more discriminating information, such that these cognitive categories are relatively stable. Park (1978) argued that consumers deal with complex information along a dimension by recoding it into chunks or categories, thereby conserving their capacity to process information (e.g., "unacceptable" if gas mileage < 15, "acceptable" if gas mileage is between 15 and 25, and "excellent" if gas mileage > 25). In research on product evaluations, Chattopadhyay and Alba (1988) argued that consumers might derive single-fact interpretations (such as "rapid acceleration") from factual details (such as "goes from 0 to 60 in 7.5 seconds"). Single-fact interpretations are argued to be less specific than factual details but capture its meaning. Other research in marketing and psychology on the processing of numerical information (which is usually precise and discriminating) also supports the notion that consumers may derive cognitive categories from more discriminating information (cf. Hinrichs and Novick, 1982; Viswanathan and Childers, 1992, 1996). Further, Park (1978) as well as other researchers (cf. Park and Lessig, 1981; Johnson and Fornell, 1987; Viswanathan and Childers, 1992) have presented a theoretical rationale for such processing in terms conserving processing capacity. For example, Park and Lessig (1981) argue that the use of a greater number of categories to classify information along a dimension would require greater effort in attaching utility to a larger number of categories.

Because information along dimensions may be classified by deriving categories from more discriminating information, consumers may retain such information and, therefore, be able to discriminate to a greater degree than is meaningful to them. Research in marketing (Viswanathan and Childers, 1992) and psychology (Holyoak and Mah, 1982) provides support for this line of reasoning. In cognitive psychology, researchers who have studied comparative judgments (i.e., tasks that require subjects to compare stimuli along some dimension and choose the stimulus with the higher or lower magnitude) suggest that categorical information may be used to make comparisons between stimuli, but more precise

information may be available in memory, if the need for a finer discrimination arises (cf. Holyoak and Mah, 1982). At a broader level, such a line of reasoning is also consistent with several sources of past research in consumer information processing that information used in decision-making may be different from information available in memory. Whereas memory for product variations (and the ability to discriminate) may be influenced by factors that are prominent during learning of information, whether such discriminating information is used in decision-making may be a function of other factors. For example, Feldman and Lynch (1988) argue that the likelihood that information in memory will be used in decision-making is a function of the ease of recalling it from memory as well as its diagnosticity. Adapted to the present context, availability of maximally discriminating information in memory will not necessarily lead to usage of such information in decision-making.

In support of the assumption that scale responses are relatively spontaneous, past research suggests that small differences in response scales can elicit differing levels of discrimination from respondents. As reviewed earlier, past research (Churchill and Peter, 1984) suggests that finer reliable discriminations can be elicited from respondents with an increase in the number of response categories of a scale. The earlier review of different approaches in terms of increases in reliability and information transmission with an increase in the number of response categories illustrates the effects of response format on scale responses. In the broader domain of scale responses outside marketing, a large body of evidence has been documented regarding differences in responses obtained as a function of differences in response scales (cf. Poulton, 1989) that points to the influence of aspects of the response scale on scale responses. In a similar vein, Parducci (1965) presented the frequency– range model that suggests that responses involve a trade-off between the frequency principle (wherein respondents use each category for a fixed proportion of judgments) and the range principle (wherein respondents use response categories to divide the range, which is the difference between the extreme values). In fact, a body of research provides evidence that responses may change with changes in the response scale, such as the frequency range of response category labels (Schwarz and Bienas, 1990; Schwarz *et al*., 1985). In summary, past research in marketing and psychology provides support for the assumptions underlying the distinction between meaningful and maximum discrimination.

2.3 Hypotheses

2.3.1 Overview

A logical sequence of objectives with implications for the measurement of consumer perceptions were assessed through tests of several hypotheses. The first objective was to use a method to determine the number of categories that are meaningful to consumers on specific product attributes in Study 1. The aim was to operationalize the notion of meaningful discrimination along attributes. Using the first study

as a basis, the second objective was to test whether the number of meaningful categories for an attribute impacts responses to scales used to rate brands on that attribute such that fewer response categories are used for ratings along attributes with fewer meaningful numbers of categories. This objective was developed as Hypothesis 1 and tested in Studies 2 and 3. The next step in the research was to examine whether meaningful differences between products could best be measured using a scale with the number of categories that are meaningful to consumers. This objective was developed as Hypothesis 2 and tested in Study 4.

2.3.2 Relationship between meaningful discrimination and scale response

In terms of the relationship between meaningful discrimination and scale response, if there is no discernible effect of the number of response categories in a scale, this effect being overwhelmed by meaningful discrimination, scales with a range of different response categories could be used to measure meaningful differences among products. If there is no discernible effect of the meaningful number of categories on scale response, this factor being overwhelmed by aspects of the response format such as the number of response categories in the scale, a scale with the number of categories that are meaningful to consumers could be used to measure meaningful differences. If both the response format and the meaningful number of categories impact scale response, a scale with the number of categories that are meaningful to consumers could be used to measure meaningful differences. Alternately, the relationship between scale response and meaningful discrimination could be examined to determine meaningful differences from scale responses. This issue is developed as Hypothesis 1.

Consider a scenario where the meaningful number of categories for attributes is less than the number of response categories in scales used to measure brand ratings along attributes. Past research would suggest that a larger number of response categories in a scale would elicit finer discriminations (cf. Parducci, 1965; Churchill and Peter, 1984). However, if scale responses are influenced by the meaningful number of categories on an attribute continuum, and if this meaningful number is less than the number of response categories in a scale, then it would exert a downward influence on the number of categories used by consumers in scale response. Further, if different attributes have different meaningful number of categories (say, 3 vs. 4), a greater downward influence would be exerted on the number of response categories used in scale response for attributes with fewer meaningful numbers of categories. Consequently, fewer response categories in a scale would be used to rate a set of products on attributes with fewer meaningful numbers of categories when compared to attributes with higher meaningful numbers of categories. This argument stated as Hypothesis 1 was tested in Studies 2 and 3.

Hypothesis 1: Ratings of products along attributes with fewer meaningful numbers of categories will involve the use of fewer response categories in a rating scale when compared to ratings of products along attributes with higher meaningful numbers of categories.

2.3.3 Accuracy of scales in measuring meaningful discrimination

The next hypothesis examined whether meaningful differences between products could best be measured using a scale with the number of categories that are meaningful to consumers. As suggested by the example at the beginning of the paper, scales that match the meaningful number of categories in terms of the number of response categories may be more accurate than other scales in measuring sameness and difference between brands that are meaningful to consumers. On the other hand, even when the number of response categories match the meaningful number of categories, scale responses may be influenced by factors other than meaningful differences. For example, Parducci (1965) presented the frequency principle (wherein consumers use each category for a fixed proportion of judgments) and the range principle (wherein consumers use response categories to divide the range). These factors may influence scale response even for scales with a meaningful number of categories such that the responses do not reflect meaningful differences. In such a situation, an approach may be to use some method such as a sorting task rather than attribute rating scales to assess meaningful differences. This issue is developed as Hypothesis 2.

Comparing two scales with three and seven response categories, respectively, from a purely probabilistic stand- point, the 3-point scale would be expected to overpredict sameness whereas the 7-point scale would overpredict difference between brands. An independent criterion such as meaningful discrimination offers a means of comparing these two scales to determine their accuracies in predicting same- ness and difference at a level of discrimination that is meaningful to consumers. Consider a scenario where seven categories are meaningful to consumers for a particular attribute (i.e., say, seven categories are used in attribute sorting). Using the sorting task as a measure of the meaningful level of discrimination, and collecting both sortings on attributes and ratings of brands on attributes on the 7-point scale versus, say, a 3-point scale, the accuracy of these two scales in predicting sameness versus difference between brands at a level that is meaningful to consumers could be assessed. Although there are trade-offs between sameness and difference accuracies for 3- versus 7-point scales from a probabilistic standpoint, a 7-point scale will lead to higher total accuracy (i.e., the average of sameness and difference accuracies) because it matches the meaningful level of discrimination, the criterion used to measure sameness and difference accuracy. In other words, a scale with a meaningful number of categories has an advantage in total accuracy because an advantage in one form of accuracy compensates for a smaller disadvantage in the other form of accuracy. Considering the reverse scenario where the

meaningful number of categories is three, a 7-point scale would outperform a 3-point scale on difference accuracy, but a 3-point scale would outperform a 7-point scale in sameness accuracy and in total accuracy. This argument is stated as Hypothesis 2 and was tested in Study 4.

> *Hypothesis 2:* A scale with a meaningful number of categories will be more accurate than other scales in predicting sameness and difference between brands along attributes at a level that is meaningful to consumers.

3 Study 1

The aim of this study was to operationalize meaningful discrimination along attributes.

3.1 Method

3.1.1 Overview

In this study, we used a sorting task because of several characteristics that make it a very suitable alternative for examining how consumers think about product attributes (Viswanathan *et al*., 1999). First, it is open-ended and, unlike category scaling, does not impose a certain number of categories on the respondents. Second, this method directly assesses the number of magnitudes used by consumers to think about product attributes. Furthermore, the sorting task has been used in research on breadth of categorization in psychology to examine somewhat similar issues as described later. Breadth of category has been defined as "the range of stimuli that are placed in the same class or category and share a common label" (Bruner and Tajfel, 1961, p. 231). A task used in breadth of categorization requires subjects to sort objects into categories or groups on a specified dimension (Block *et al*., 1981). The number of categories or groups employed in sorting has been used as a measure of conceptual differentiation (Gardner and Schoen, 1962). The categorization of objects on a specified dimension provides a means of understanding the number of groups used to think about a continuum. Using a similar approach for a product attribute continuum, the sorting of products on specified attributes was used here.

3.1.2 Stimulus materials

The product categories (four in all: candy bars, snack foods, soft drinks, and beverages), the specific products (12 in each category), as well as the attributes used for each product category (three attributes for each product category for a total of 12 product– attribute combinations: crunchiness, chocolaty flavor, and caramel flavor for candy bars; sweetness, saltiness, and "how good a snack" for snack foods; sweetness, caffeine content, and "how refreshing" for soft drinks; and sweetness, coldness, and "how refreshing" for beverages) were chosen from past research (Johnson *et al*., 1990). Johnson *et al*.

(1990) elicited lists of attributes from respondents for each product in each product category. The top three attributes for each category on the basis of frequency of elicitation were selected for this study. The four categories consisted of two sets of subordinate– superordinate pairs, candy bars– snack foods, and soft drinks– beverages. In a pilot test where subjects were asked to sort brands of candy bars and then describe how they performed the sorting task, written descriptions and responses to scales completed by subjects after each sorting task suggested that they were adhering to instructions and performing the task with relative ease.

3.1.3 Procedures

An overview of the procedures is presented in Fig. 1. Subjects sorted sets of 12 products in to groups along specified attributes. In past research using the sorting task, subjects have been instructed to sort objects into groups that "go together" based on certain dimensions (cf. Block *et al*., 1981). A similar approach was employed here. A particular attribute was specified and subjects were required to sort 12 products/brands into groups that go together. Subjects were instructed that they could use any number of groups that seemed appropriate. Instructions stressed that subjects should perform the sorting only along the specified attribute. Subjects were presented with a list of the 12 products and asked to write down the names of products and draw circles around them to indicate groups. After each sorting task, subjects wrote descriptions of how they performed the sorting task and completed scales relating to the task One hundred twenty students from a midwestern university participated in the study. Each subject sorted each of the four different product stimulus sets (i.e., candy bars, snack foods, soft drinks, and beverages) on the basis of a specified attribute for each category (such as sweetness of soft drinks, caffeine content of beverages, etc.), with the attributes used for each product category being different across the three groups of subjects. One attribute for each category mentioned previously was selected to form a set of four attributes and three such sets were formed. Three groups of 40 subjects each completed a questionnaire corresponding to each set. The four categories consisted of two sets of subordinate– superordinate pairs. A constraint in choosing attributes to comprise a set was that similar attributes were not included in the same set (e.g., sweetness of candy bars and sweetness of snack foods were not placed in the same set). The use of similar attributes was avoided to minimize lack of independence across sortings. The sequence of categories for sorting were counterbalanced within each set with the constraint that no two categories from a subordinate– superordinate pair were presented in a consecutive sequence. On this basis, eight versions of the questionnaire were prepared for each set of 40 subjects with five subjects assigned to each version.

3.2 Results

3.2.1. Assessment of the sorting task

Responses to scales employed to assess the sorting task suggested that subjects were adhering to instructions in performing the task by concentrating on the specified attribute, a central requirement in order to measure the meaningful number of categories for a specific product attribute (see Viswanathan *et al*., 1999).

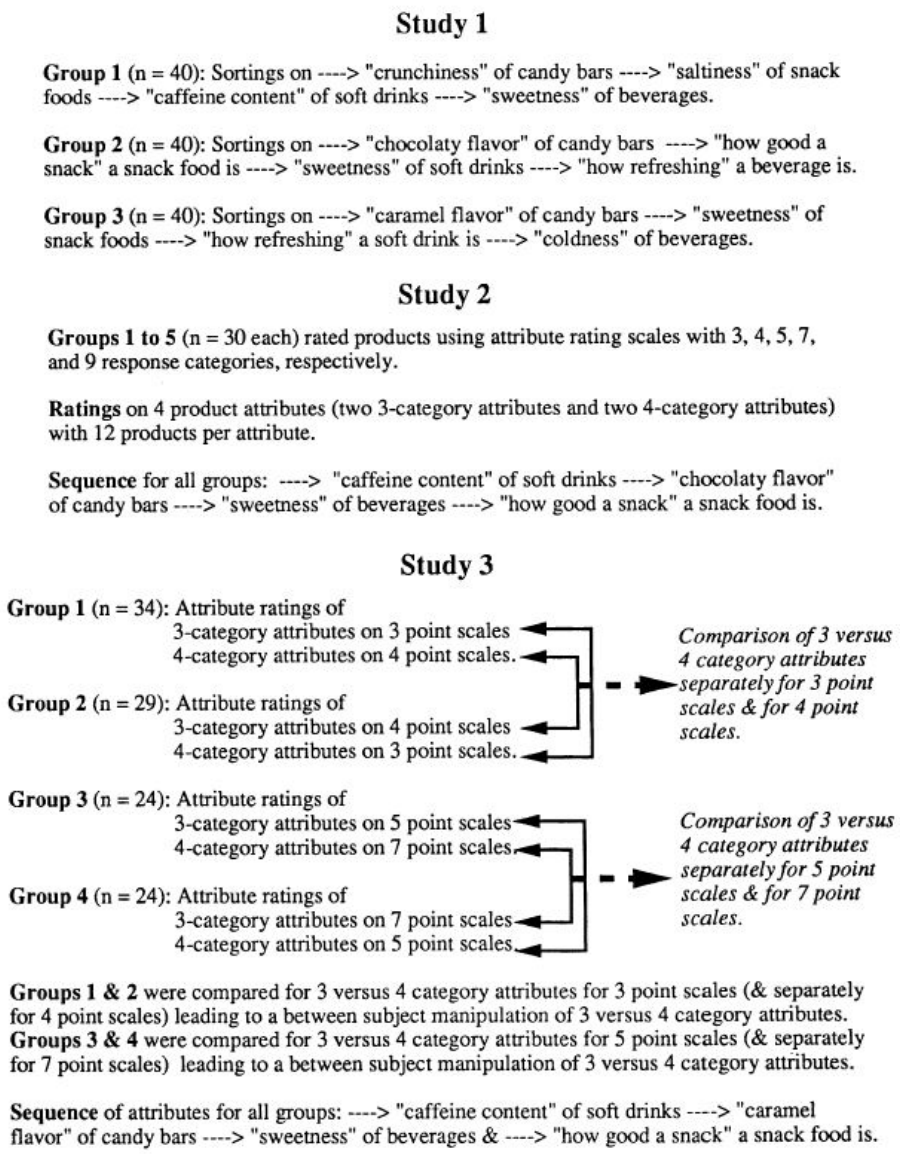


Fig. 1. Overview of Studies 1–3.

3.2.2 Assessment of meaningful number of categories

The mean number of categories employed for each product–attribute combination was computed across subjects and ranged from 2.75 to 4.18 for the 12 product–across subjects and ranged from 2.75 to 4.18 for the 12 product–attribute combinations (see Table 1). The frequency distribution of responses is also presented to illustrate the heterogeneity in responses, an issue that is discussed in the concluding discussion. Based on these results, four product– attribute combinations were chosen for subsequent studies such that two would represent three-category attributes and two would represent four-category attributes. The product – attribute combinations chosen with attributes shown in quotations were sweetness of beverages, and how good a snack a snack food is as four-category attributes, with the mean number of categories employed of 4.13 and 4.18, respectively, and caffeine content of soft drinks, and chocolaty flavor of candy bars as three-category attributes, with the mean number of categories employed of 2.95 and 3.50, respectively. Because several product attributes appeared to be appropriate for use as three- or four-category attributes, the choice of these four specific product attributes was somewhat arbitrary. One attribute was chosen from each of the four product categories. Comparisons of the number of categories employed for four- versus three- category attributes using *t* tests suggested significant differences. The number of categories for how good a snack a snack food is was significantly greater than the number of categories for chocolaty flavor of candy bars [$t(39) = 3.08$, $P < .01$] and for caffeine content of soft drinks [$t(77) = 5.00$, $P < .001$]. The number of categories for sweetness of beverages was significantly greater than the number of categories for caffeine content of soft drinks [$t(38) = 4.74$, $P < .001$] and for chocolaty flavor of candy bars [$t(77) = 2.05$, $P < .05$]. Therefore, Study 1 provided a set of product– attribute combinations with different meaningful numbers of categories that were used in subsequent studies. A noteworthy outcome of this study was that a larger number of meaningful categories were used for attributes of superordinate when compared to subordinate categories.

4 Studies 2 and 3

Study 1 illustrated a method for operationalizing meaningful discrimination with the results suggesting the very likely conclusion that the meaningful number of categories differs for different attributes. In light of these differences, the next step in the research was to assess whether meaningful discrimination influences scale response, i.e., test Hypothesis 1. As mentioned earlier, several possibilities exist in this regard with implications for the measurement of consumer perceptions.

4.1 Study 2

4.1.1 Method

In Study 2, the product– attribute combinations chosen from Study 1 were used to collect ratings of products along attributes using scales with different numbers of response categories. The procedure for Study 2 is summarized in Fig. 1. One hundred fifty students at a midwestern university participated in Study 2 with an approximately equal number of subjects being assigned to each of five groups, which differed in the number of response categories used for the rating scales. Three, four, five, seven, and nine response categories, respectively, were used for the five groups to test Hypothesis 1 across a range of commonly used scales with different numbers of response categories. Subjects rated products on the set of four product– attribute combinations chosen from Study 1 (i.e., 12 products per attribute for a total of 48 ratings). Scales were appropriately end-anchored in terms of the attribute that was being measured.

4.1.2 Data analyses

An important issue in data analyses was to develop an appropriate indicator of the frequency of *usage* of response categories in order to test Hypothesis 1. Hypothesis 1 states that fewer response categories will be used to rate three- category attributes when compared to four-category attributes, hence the need to measure usage of response categories. The frequency of usage of response categories has been aggregated across subjects in past research in order to investigate the degree of utilization of specific scale positions (cf. Wyatt and Meyers, 1987). However, the goal here was to assess the frequency of usage of response categories *without regard to specific scale positions*. As stated previously, Hypothesis 1 relates to the number of response categories used without regard to specific scale positions.

Table 1
Number of categories used in sorting

Product attribute	Number of categories employed						Mean	S.D.
	1	2	3	4	5	>5		
	% of Respondents							
Study 1								
"Caramel flavor" of candy bars	0.0	45.0	37.5	15.0	2.5	0.0	2.75	0.81
"Caffeine content" of soft drinks	2.6	35.9	33.3	20.5	7.7	0.0	2.95	1.00
"How refreshing" a soft drink is	2.5	17.5	47.5	22.5	5.0	5.0	3.25	1.06
"Coldness" of beverages	0.0	12.5	55.0	12.5	17.5	2.5	3.45	1.09
"Chocolaty flavor" of Candy bars	2.5	15.0	35.0	32.5	10.0	5.0	3.50	1.18
"Saltiness" of snack foods	0.0	20.5	33.3	25.6	12.8	7.7	3.54	1.19
"How refreshing" a beverage is	2.5	20.0	27.5	22.5	25.0	2.5	3.55	1.22
"Crunchiness" of candy bars	0.0	5.1	56.4	17.9	15.4	5.1	3.59	0.99
"Sweetness" of soft drinks	0.0	15.0	37.5	30.0	10.0	7.5	3.60	1.17
"Sweetness" of snack foods	0.0	0.0	47.5	32.5	15.0	5.0	3.83	1.04
"Sweetness" of beverages	2.6	7.7	25.6	30.8	17.9	15.4	4.13	1.53
"How good a snack" a snack food is	0.0	0.0	32.5	35.0	22.5	10.0	4.18	1.17
Study 4								
"Caramel flavor" of candy bars								
Group 1[a]	2.0	31.4	54.9	11.8	0.0	0.0	2.77	0.68
Group 2	0.0	38.0	50.0	10.0	2.0	0.0	2.76	0.72
Group 3	0.0	32.6	47.8	17.4	2.2	0.0	2.89	0.77
Total	0.7	34.0	51.0	12.9	1.4	0.0	2.80	0.72
"Crunchiness" of candy bars								
Group 1	2.0	15.7	56.9	21.6	3.9	0.0	3.10	0.7
Group 2	0.0	18.0	62.0	14.0	6.0	0.0	3.08	0.75
Group 3	0.0	6.4	57.4	25.5	10.6	0.0	3.40	0.77
Total	0.7	13.5	58.8	20.3	6.8	0.0	3.19	0.78
"Chocolaty flavor" of candy bars								
Group 1	5.9	31.4	51.0	9.8	2.0	0.0	2.71	0.81
Group 2	0.0	42.0	42.0	14.0	2.0	0.0	2.76	0.77
Group 3	0.0	29.8	40.4	25.5	4.3	0.0	3.04	0.86
Total	2.0	34.5	44.6	16.2	2.7	0.0	2.83	0.82
"Caffeine content" of soft drinks								
Group 1	2.0	30.0	52.0	16.0	0.0	0.0	2.77	0.82
Group 2	4.2	27.1	47.9	18.8	2.1	0.0	2.88	0.84
Group 3	0.0	34.0	44.7	17.0	4.3	0.0	2.92	0.83
Total	2.1	30.3	48.3	17.2	2.1	0.0	2.85	0.83
"Sweetness" of soft drinks								
Group 1	0.0	24.0	54.0	18.0	4.0	0.0	2.96	0.87
Group 2	0.0	26.5	46.9	20.4	4.1	2.0	3.08	0.91
Group 3	0.0	17.0	55.3	14.9	10.6	2.1	3.28	1.02
Total	0.0	22.6	52.1	17.8	6.2	1.4	3.10	0.93
"How refreshing" a soft drink is								
Group 1	3.9	25.5	41.2	25.5	2.0	2.0	3.04	1.04
Group 2	0.0	30.0	50.0	16.0	4.0	0.0	2.94	0.79
Group 3	4.2	22.9	52.1	14.6	4.2	2.1	2.98	0.96
Total	2.7	26.2	47.7	18.8	3.4	1.4	2.99	0.93

[a] Groups 1–3 represent data from the three groups of 50 subjects each in Study 4 as shown in Fig. 3.

An aggregation of frequency of usage of specific scale positions across subjects was inappropriate for the purpose at hand as illustrated by the following example. Consider an extreme scenario where all consumers used a single response category to rate a range of stimuli on a 7-point scale but were equally likely to use any one of the seven response categories. An aggregation across consumers would result in a uniform distribution across the seven response categories suggesting the use of seven categories.

However, the appropriate indicator of the number of response categories utilized by each consumer is one. By extension, similar problems arise with the use of two or more response categories by consumers. Therefore, a different type of analysis was performed.

Numerical examples are shown in Fig. 2 for a single consumer and used to demonstrate the rationale and procedures for data analyses. The frequency of usage of each response category was computed for each consumer and ordered from highest to lowest. For example, on the 7-point scale in Fig. 2, ratings of 12 brands on a three-category attribute involve the use of three response categories, scale point 1 a total of five times, scale point 4 a total of four times, and scale point 7 a total of three times. Therefore, the ordered frequencies for the most, second most, and third most frequently used response categories are Categories 5, 4, and 3, respectively. Each set of ordered frequencies for a consumer reflected the extent to which each response category was used from the most to least frequently used.

These frequencies can be meaningfully aggregated across subjects. This is because the aggregation across subjects is on the basis of a ranking on frequency of usage such that the most frequently used response category is aggregated across subjects followed by the two most frequently used response categories and so on. For example, in Fig. 2, cumulative frequencies for the three-category attribute of the most frequently used response category, the two most frequently used category, and the three most frequently used category are 5, 9 (i.e., 5 + 4), and 12 (5 + 4 + 3), respectively.

The number of response categories required to capture certain percentages of ratings has been used in past research (Ramsay, 1973). Adapting such an indicator to the purpose at hand, the dependent variable computed using these data was the proportion of the total of 12 products captured by a certain number of most frequently used response categories. Hypothesis 1 states that fewer response categories will be used to rate three-category attributes when compared to four-category attributes. Restating Hypothesis 1 in terms of the dependent variable described previously, the proportion of products captured by, say, the most frequently used response category is hypothesized to be lower for four-category attributes when compared to three-category attributes. This is because a greater number of response categories is hypothesized to be used for four- category attributes. Consequently, a lower proportion of products would be captured by a certain number of response categories for the four-category attributes when compared to the three-category attributes. For example, in Fig. 2, the proportion of products captured by the most frequently used category is .42 versus .33 for three- versus four-category attributes.

4.1.3 Results and discussion

A 5 x 2 x 2 factorial ANOVA was run on the data from Study 2 based on the proportion of products captured by the most frequently used response category. The factors were the number of response categories in a scale (between subjects: three, four, five, seven, and nine), three- versus four-

category attributes (within subjects), and replication of attributes (within subjects). A significant main effect was found for the meaningful number of categories [$F(1,143) = 42.18$, $P < .001$], with higher proportions of products being captured by a certain number of response categories for three-category attributes in line with Hypothesis 1 (.45 versus .39). Contrasts comparing differences between three- and four-category attributes for each scale with a certain number of response categories were examined to test Hypothesis 1. As predicted, significantly higher proportions of products were captured by a certain number of response categories for three-category attributes when compared to four-category attributes for the 4-, 5-, 7-, and 9-point scales (see Table 2). A nonsignificant effect was obtained for the 3- point scale perhaps because of the availability of only three response categories to rate products along both three- and four-category attributes, i.e., because the number of response categories were the same as the meaningful number of categories for the three-category attributes.

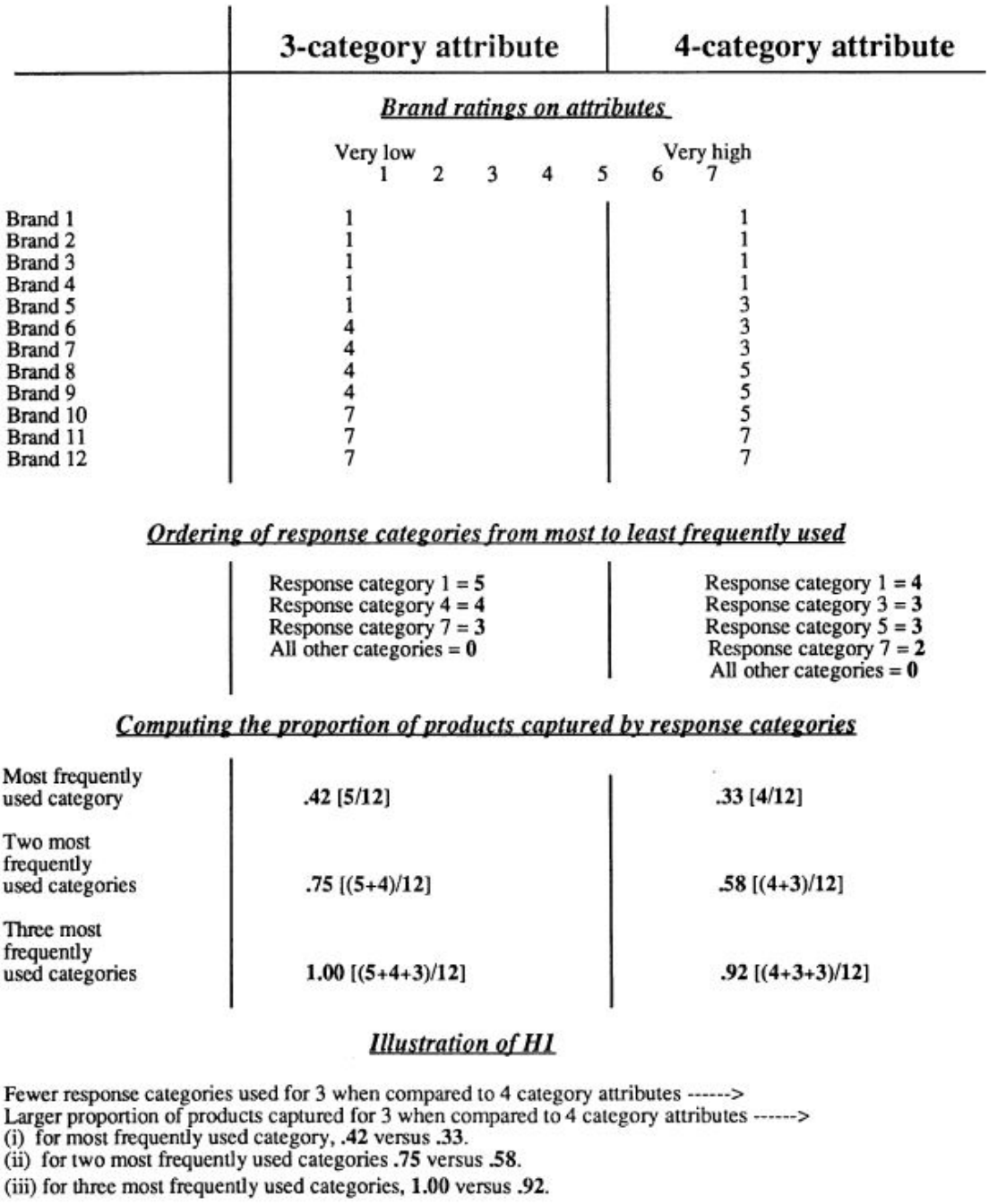


Fig. 2. Illustration of computation of frequency of usage—Studies 2 and 3.

A similar ANOVA was run on the proportion of products captured by the two most frequently used response categories. A significant main effect was found for the meaningful number of categories [$F(1,143) = 80.85$, $P < .001$], with higher proportions of products being captured by a certain number of response categories for three-category attributes in line with Hypothesis 1 (.74 versus .66). As predicted, significantly higher proportions of products were captured by a certain number of response categories for

three- category attributes when compared to four-category attributes for the 4-, 5-, 7-, and 9-point scales (see Table 2). A similar ANOVA was run on the proportion of products captured by the three most frequently used response categories. A significant main effect was found for the meaningful number of categories [$F(1,143) = 96.48$, $P < .001$], with higher proportions of products being captured by a certain number of response categories for three-category attributes in line with Hypothesis 1 (.90 versus .83). As predicted, significantly higher proportions of products were captured by a certain number of response categories for three-category attributes when compared to four-category attributes for the 4-, 5-, 7-, and 9-point scales (see Table 2). Mean proportions are also presented in Table 2 for 5-, 7-, and 9-point scales in terms of the four most frequently used categories, etc. The results provide strong support for Hypothesis 1. Nonsignificant effects were obtained in some instances when nearly all the products were captured by a certain number of response categories for both three- and four-category attributes.

4.2 Study 3

Study 3 was conducted to test Hypothesis 1 and attempt to replicate the results of Study 2 using a different design that is described later.

Table 2
Proportions of products as captured by response categories

Number of response categories and type of attribute	Number of most frequently used categories							
	1	2	3	4	5	6	7	8
Study 2								
3-Point scale								
Three-category attributes	.52	.86						
Four-category of attributes	.49	.82						
4-Point scale								
Three-category attributes	.48*	.79*	.96*					
Four-category attributes	.43	.72	.89					
5-Point scale								
Three-category attributes	.47*	.76*	.93*	.99*				
Four-category attributes	.39	.66	.84	.95				
7-Point scale								
Three-category attributes	.40*	.66*	.82*	.94*	.99	1.00		
Four-category attributes	.34	.58	.75	.87	.96	.99		
9-Point scale								
Three-category attributes	.38*	.64*	.80*	.91*	.96*	.98	.99	1.00
Four-category attributes	.32	.53	.68	.81	.89	.96	1.00	1.00
Study 3								
3-Point scale								
Three-category attributes	.57*	.87*						
Four-category attributes	.49	.82						
4-Point scale								
Three-category attributes	.55*	.82*	.94					
Four-category attributes	.45	.72	.90					
5-Point scale								
Three-category attributes	.52*	.79*	.93*	.99				
Four-category attributes	.39	.67	.85	.95				
7-Point scale								
Three-category attributes	.50*	.73*	.87*	.95*	.98	.99		
Four-category attributes	.38	.62	.79	.90	.96	1.00		

* Significantly higher proportion for the three-category attributes when compared to the four-category attributes at .05, .01, or .001 levels.

4.2.1 . Method

In Study 2, the same group of subjects rated products on three- versus four-category attributes for a scale with a certain number of response categories leading to a test of Hypothesis 1 within groups of subjects. In order to provide a stronger test of Hypothesis 1 in Study 3, a between-subject manipulation of three- versus four-category attributes was used. As a result, for a scale with a certain number of response categories, independent ratings of products for three- versus four-category attributes were collected from *different* groups of subjects. A different design was used to achieve an efficient between-subject manipulation of three- versus four-category attributes and is presented in Fig. 1. One group rated products on the three-category attributes using a 3 (5)-point scale and the four-category attributes using a 4 (7)-point scale while another group rated products on the three-category attributes using a 4 (7)-point scale and the four-category attributes using a 3 (5)-point scale. Comparisons between the three- versus four-category attributes were performed separately for the three-point scale (or the 4-point scale) by combining portions of data from each group of subjects consisting of ratings on the three-point scale (or the 4-point scale). Consequently, the design described previously led to a between-subject manipulation of three- versus four-category attributes for a scale with a certain number of response categories. One change in the set of product attributes used in Study 2 was the replacement of chocolaty flavor of candy bars in the three-category condition (mean number of sorted categories of 3.50 in Study 1), with caramel flavor of candy bars, which had a mean number of sorted categories that was closer to 3 and farther from 4 (i.e., 2.75 in Study 1).

4.2.2 Results

The 2 (three- versus four-category attributes; between subjects) x 2 (replication of attribute; within subjects) ANOVAs were run separately on the data for the 3-point scale for the most frequently used category. Similar ANOVAs were run for the two most frequently used categories, etc. Similar ANOVAs were also run separately for the 4-, 5-, and 7-point scales. Significant main effects were found for three- versus four-category attributes, as predicted (see Table 2). The findings were similar to those of Study 2 and provide strong support for Hypothesis 1 (Table 2).

4.3. Discussion of Studies 2 and 3

Two findings emerged from Study 2, which were replicated in Study 3. First, the results demonstrated that mean- *4.2.1. Method* In Study 2, the same group of subjects rated products on three- versus four-category attributes for a scale with a certain number of response categories leading to a test of

Hypothesis 1 within groups of subjects. In order to provide a stronger test of Hypothesis 1 in Study 3, a between-subject manipulation of three- versus four-category attributes was used. As a result, for a scale with a certain number of response categories, independent ratings of products for three- versus four-category attributes were collected from *different* groups of subjects. A different design was used to achieve an efficient between-subject manipulation of three- versus four-category attributes and is presented in Fig. 1. One group rated products on the three-category attributes using a 3 (5)-point scale and the four-category attributes using a 4 (7)-point scale while another group rated products on the three-category attributes using a 4 (7)-point scale and the four-category attributes using a 3 (5)-point scale. Comparisons between the three- versus four-category attributes were performed separately for the three-point scale (or the 4-point scale) by combining portions of data from each group of subjects consisting of ratings on the three-point scale (or the 4-point scale). Consequently, the design described previously led to a between-subject manipulation of three- versus four-category attributes for a scale with a certain number of response categories. One change in the set of product attributes used in Study 2 was the replacement of chocolaty flavor of candy bars in the three-category condition (mean number of sorted categories of 3.50 in Study 1), with caramel flavor of candy bars, which had a mean number of sorted categories that was closer to 3 and farther from 4 (i.e., 2.75 in Study 1).

4.2.2. Results

The 2 (three- versus four-category attributes; between subjects) x 2 (replication of attribute; within subjects) ANOVAs were run separately on the data for the 3-point scale for the most frequently used category. Similar ANOVAs were run for the two most frequently used categories, etc. Similar ANOVAs were also run separately for the 4-, 5-, and 7-point scales. Significant main effects were found for three- versus four-category attributes, as predicted (see Table 2). The findings were similar to those of Study 2 and provide strong support for Hypothesis 1 (Table 2).

4.3. Discussion of Studies 2 and 3

Two findings emerged from Study 2, which were replicated in Study 3. First, the results demonstrated that meaningfulness impacts scale response such that fewer response categories are used to rate products along attributes with fewer meaningful numbers of categories. The second key finding was that the higher the number of response categories in a scale, the greater the discrimination elicited, as suggested in past research. The mean proportion of products captured by the three most frequently used response categories in Study 2 for the 3-, 4-, 5-, 7-, and 9-point scales, were 1.00, .93, .89, .78, and .74, respectively, with all pairwise differences being significant at the .01 level. Similar results were obtained

for Study 3. Mean proportions for scales with different number of response categories in Table 2 demonstrate this pattern of findings.

These findings suggest that, though the number of response categories in a scale influence scale response by eliciting finer discriminations with increases in the number of response categories, the number of meaningful categories for an attribute also influences attribute ratings. Further-more, the results suggest the possibility of assessing meaningful discrimination from scale response. If, say, the meaningful number of categories for an attribute is three, then a certain proportion of products may be captured by three response categories for a 5-point scale. In other words, the empirical relationship between the number of meaningful categories and the proportions of products captured by a certain number of response categories for certain scales could be used to infer meaningful discrimination from scale response. Scale usage between attributes could also be compared to infer differences in meaningful number of categories for different attributes. Such differences could be used in interpreting whether differences in attribute ratings are meaningful.

5. Study 4

Studies 2 and 3 suggested that both the response format and the meaningful number of categories impact scale response. As suggested earlier, a scale with the number of categories that are meaningful to consumers can be employed in such a scenario to measure meaningful differences between products. The next step in the research was, therefore, to examine whether meaningful differences between products could best be measured using a scale with the number of categories that are meaningful to consumers, i.e., a test of Hypothesis 2.

5.1. Method

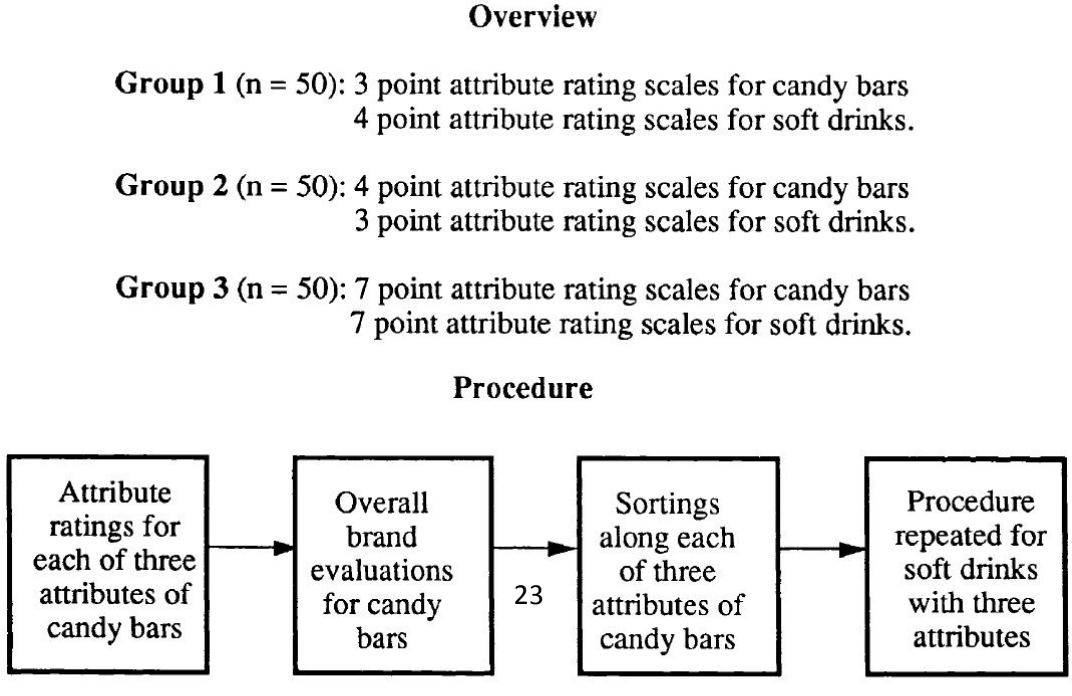
5.1.1. Overview

The method for Study 4 is summarized in Fig. 3 where subjects completed both sorting and rating tasks to assess Hypothesis 2. The aim here was to evaluate rating scales using sortings as the criteria for meaningful discrimination. Therefore, rating scales were compared on the degree to which they captured sameness and difference between brand pairs as measured by the sortings. In Study 4, two product categories were chosen from Studies 1 to 3, namely, candy bars and soft drinks. Three attributes each were chosen for candy bars (i.e., caramel flavor, chocolaty flavor, and crunchiness) and soft drinks (i.e., caffeine content, sweet- ness, and how refreshing a soft drink is) such that the mean number of categories for each attribute was approximately three based on the sorting task in Study 1 (i.e., ranging from 2.75 to 3.60; see Table 1). The number of response categories used for the attribute rating scales was manipulated between subjects to be 7 (for both soft drinks and candy bars), 3/4 (i.e., a 3-point scale for

soft drinks and a 4-point scale for candy bars), or 4/3 (i.e., a 4-point scale for soft drinks and a 3-point scale for candy bars) (Fig. 3). The 4- point scales were used in order to compare a scale with a meaningful number of categories (i.e., a 3-point scale) to a scale with an additional category.

5.1.2. Procedures

One hundred fifty undergraduate students at a midwestern university completed a questionnaire, with 50 students each being assigned to the 7-point scale condition and the two 3/4-point scale conditions, respectively (Fig. 3). Sub- jects first rated the 12 candy bars used in Studies 1 – 3 on each of three attributes: caramel flavor, crunchiness, and chocolaty flavor. The scales used to rate these attributes were appropriately anchored at the ends. After completing ratings on each attribute, subjects completed scales measuring comfort in using attribute rating scales, as well as satisfaction with scales and suitability of scales in capturing accurate responses. Next, subjects completed several scales relating to overall evaluations of brands of candy bars. Specifically, subjects filled out 7-point scales for each of the 12 brands based on brand liking (anchored dislike it very much to like it very much), brand rating (very bad to very good), and likelihood of purchase (not at all likely to very likely). Then they ranked the 12 brands on the basis of their liking and completed a brand choice task. These tasks served to separate the attribute rating tasks from the attribute sorting tasks that followed. They then completed a sorting task based on overall liking for candy bars where the instructions asked them to sort products ‘‘into groups that go together on the basis of how much you like them.’’ Next, in the attribute sorting tasks, subjects sorted the candy bars on each of three attributes; caramel flavor, crunchiness, and chocolaty flavor. After each sorting task, subjects completed scales relating to the task as in Study 1. Finally, subjects rated the importance of each of the three attributes in selecting a candy bar and also ranked them in the order of importance. The whole procedure was repeated for the product category, soft drinks.



Rather than ask subjects to sort the candy bars into groups that go together along specified attributes as in Study 1, subjects were asked to sort candy bars on a specific attribute as they would ‘‘if they were deciding how much they like each candy bar.’’ This change was incorporated into the procedure in order to measure the number of categories that was meaningful to consumers in the context of deciding how much they liked each brand. This modification was made because the concept of the meaningful number of categories was argued earlier to relate to how consumers think about a product attribute in the course of decision-making.

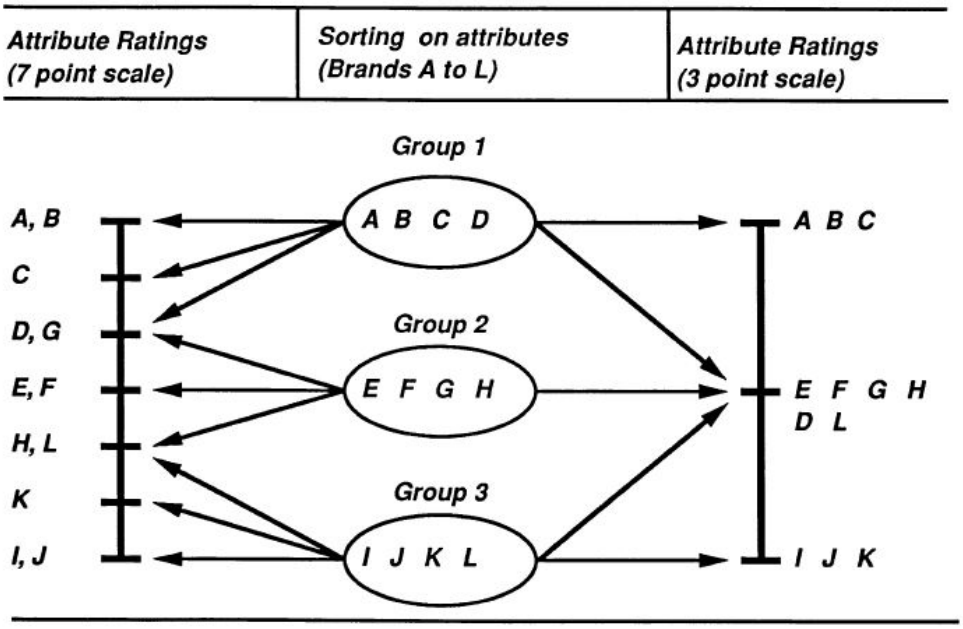
5.1.3. Preliminary analyses

Using small but significant differences in the number of meaningful categories between attributes on the basis of the sorting tasks used in Study 4, Hypothesis 1 was tested and supported in both studies. Preliminary analyses suggested that the mean number of categories was close to three for all six attributes in Study 4, for each group of 50 subjects as well as for the overall sample (bottom of Table 1). Differences in the mean number of categories were mostly nonsignificant across groups. Therefore, the 3-point scale used in the two 3/4 conditions for both soft drinks and candy bars had the meaningful number of categories for all attributes. Comparisons of means between 3-, 4-, and 7-point scales conditions suggested that the sorting method is resistant to the influence of the rating scales that preceded it in measuring the meaningful number of categories. Whereas some of the means in Study 4 were not significantly different from the means in Study 1, some of the means were significantly less than the means in Study 1. Perhaps, the use of a specific context in Study 4, i.e., a decisionmaking context, may have decreased the number of categories that consumers used in sorting. Means on scales measuring comfort, satisfaction, and suitability of attribute rating scales were directionally higher for 3-point scales when compared to 7-point scales.

As in Study 1, the sorting task was assessed by examining responses to scales. Mean responses for each subject for each of these scales across the six sorting tasks were computed and were very consistent with Study 1. As in Study 1, the sorting task was also assessed by treating the number of groups that subjects sorted products into for each of the six sortings as items in a multiple item scale. Consistent with Study 1, moderate intercorrelations between these items (average intercorrelation =.45) and a moderate reliability for the 6-item scale (average coefficient alpha= .83) point to the existence of individual differences in sorting.

5.2. Results

Hypothesis 2 states that a scale with a meaningful number of categories (i.e., a 3-point scale) will be more accurate than other scales in predicting sameness and difference between brands that is meaningful to consumers (i.e., on the basis of the sorting tasks). In order to compare the 3-, 4-, and 7-point scales in Study 4 to test Hypothesis 2, the degree of accuracy of these scales in predicting whether pairs of brands were perceived as being identical or different on specific attributes was examined using the attribute sortings as a criterion. Numerical examples are shown in Fig. 4 to illustrate the data analyses employed here. On a particular attribute, any brands could be sorted into the same category or into different categories. For example, in Fig. 4, brands A, B, C, and D were sorted into the same category. Therefore, brand pairs AB, AC, etc. were sorted into the same category. However, brand pairs AE, AF, etc. were sorted into different categories. Similarly, any brand pair could be rated as being the same or different on an attribute rating scale. For example, brand pair AB is rated as being the same on the 7-point scale whereas brand pair AC is rated as being different. All possible pairwise combinations of 12 brands in each category led to 66 possible brand pairs. On each attribute, all brand pairs out of the total of 66 that were sorted into the same category were identified (i.e., 18 in Fig. 4). Among these brand pairs, the proportion that were rated as being the same on the 3 (4 and 7)-point scale was computed as a measure of sameness accuracy of the scale (i.e., 12/18 for the 3-point scale in Fig. 4). This proportion is an appropriate measure of sameness accuracy because the criterion used to assess accuracy is meaningful discrimination as measured by the sorting task and represents the extent to which the rating scale is similar to the sorting task in measuring sameness. Similarly, on each attribute, all brand pairs out of the total of 66 that were sorted into different categories were identified (i.e., 48 in Fig. 3). Among these brand pairs, the proportion that were rated as being different on the 3 (4 and 7)-point scale was computed as a measure of difference accuracy of the scale (i.e., 39/48 for the 3-point scale in Fig. 3). This analysis was performed for each attribute in each product category. (If more than



Sameness accuracy

Number of brand pairs sorted into same category = 6 x 3 (AB, AC, etc.; EF, etc.) = **18**

Number of brand pairs among the 18 above that were rated as being the same:
7 point scale: 1(A,B)+ 1(E,F)+ 1(I,J) = **3**; **3 point scale**: 3(A,B,C)+ 6(E,F,G,H)+ 3(I,J,K) = **12**

Sameness accuracy, i.e., proportion of all brand pairs sorted into same category in sorting that were rated as being the same on a rating scale:
7 point scale: 3/18 = **0.17**; **3 point scale**: 12/18 = **0.67**.

Difference accuracy

Number of brand pairs sorted into different categories = 66 - 18 = **48**.

Number of brand pairs among the 48 above that were rated as being different :
7 point scale: 8 x 3 (A,B,C, with E thro' L) + 7 (D with E thro' L except G) + 3 x 4 (E,F,G, with I thro' L, etc.) + 3 (H with I,J,K) = **46**
3 point scale : 8 x 3 (A,B,C with E thro' L) + 3 (D with I,J ,K) + 4x3 (E,F,G,H with I,J,K) = **39**

Difference accuracy, i.e., proportion of all brand pairs sorted into different categories in sorting that were rated as being different on a rating scale:
7 point scale: 46/48= **0.96** ; **3 point scale**: 39/48 = **0.81**

Fig. 4. Illustration of data analyses for Study 4.

50% of the 66 possible comparisons were missing for a subject on an attribute, i.e., a majority of data along an attribute, data for that subject on that attribute was treated as missing, leading to the exclusion of approximately 1% of the data.)

In terms of overall means summed across the three attributes of each category and across both product categories, both the 4- and the 7-point scales were significantly higher than the 3-point scale in difference accuracy (Table 3). However, the 3-point scale was significantly higher than both the 4- and the 7-point scales in sameness accuracy and in total accuracy, providing support for Hypothesis 2. Each product category was also analyzed separately leading to a similar pattern of results (Table 3). For candy bars, the 4- and 7-point scales were significantly higher than the 3-point scale in difference accuracy. However, the 3-point scale was significantly higher than the 4- and 7-point scales in sameness accuracy, significantly higher than the 7-point scale for total accuracy, and directionally higher than the 4-point scale for total accuracy, providing support for Hypothesis 2. For soft drinks, the 7-point scale was significantly higher than the 3-point scale and the 4-point scale was directionally higher than the 3-point scale in difference accuracy. However, the 3-point scale was significantly higher than the 4- and 7-point scales in sameness accuracy, and in total accuracy, providing support for Hypothesis 2. This pattern of results in terms of total accuracy was found at a significant level for one out of three attributes of candy bars and two out of three attributes of soft drinks with directional support for the other attributes. Comparing 4- and 7-point scales in similar analyses, no significant difference in total accuracy was found in all analyses (Table 3). Trade-offs between sameness and difference accuracies were found at a significant level for candy bars and at a directional level for soft drinks. In overall analysis summed across product categories, a significant difference was found only for sameness accuracy. The results provide support for Hypothesis 2.

5.3. Discussion of results

The results of Study 4 provide support for the hypothesis that a scale with a meaningful number of categories will be more accurate than other scales. Other analyses suggested that differences on scales with a meaningful number of categories appeared to be more likely to be reflected in differences in overall evaluations than other scales.1 Additional analyses suggested that the use of a scale with a

meaningful number of categories might come without a sizable loss of performance in correlational approaches.

6. General discussion

This research addresses a neglected issue in validity, i.e., the validity of inferences of sameness and difference drawn from stimulus-centered scales with a certain number of response categories. In contrast to a focus in past research on maximizing the discrimination elicited from consumers, a distinction is drawn in this paper between the maximum number of categories that consumers can discriminate between and the number of categories that are meaningful to them. Thus, this research incorporates the unique perspective of consumer behavior with its central focus on phenomena such as product judgment and choice into the measurement of consumers' perceptions about attributes. The significance of using scales with a meaningful number of categories is in validly measuring differences between products that are meaningful to consumers. Such differences are more likely to be reflected in differences in consumer liking, purchase intent, and purchase behavior than differences obtained with some other scale that elicits greater discrimination. Therefore, valuable diagnostic information about consumer attitudes and behaviors can be gained.

Although this research was methodological in its orientation, both methodological and substantive research on magnitude information has implications for areas of marketing that relate to the usage of product information in judgments. Insight could be gained into the relationship between the number of meaningful categories and properties of attributes (for example, attributes of superordinate categories appear to have a larger number of meaningful categories than attributes of subordinate categories in Study 1) as well as characteristics of consumers (such as expertise). Research in areas such as attitude formation and memory-based judgments has not explicitly examined the nature of magnitudes characterizing attribute continua that underlie global attitudes and judgments. Attitude research has typically measured beliefs about product attributes using traditional attribute rating scales whereas the present research suggests that meaningful discrimination at the attribute level forms the basis for discrimination of products and the formation of global attitudes.

An advantage of using scales with meaningful number of categories may be that a difference of one unit is likely to be meaningful. Therefore, a norm for interpreting magnitudes of differences is available for such scales, albeit an approximate one based on an aggregation of data across consumers. The notion of meaningful discrimination could also be applied to more aggregate scales such as those measuring purchase intentions and overall judgments of products. The notion of meaningful discrimination could also be applied in determining how a scale should be labeled. Whereas researchers sometimes collapse underutilized end-anchors of scales, the approach suggested here would provide a

basis to determine how to label end-points and minimize their underutilization. Such labeling may also minimize other response errors such as overutilization of the middle point or of one end of a scale. Scale responses could also be analyzed using the scale utilization procedure employed here to identify whether too many or too few response categories may have been used and in interpreting mean ratings or in considering options such as collapsing response categories before computing means.

Because individual differences in the meaningful number of categories may exist between consumers, data on meaningful discrimination and scale utilization should be examined for such differences. If subgroups of customers have very different numbers of meaningful categories, smaller segments may need to be identified on the basis of the distinctly different decision-making processes used by different groups of consumers. In this regard, several individual difference variables may of relevance such as need for precision, a preference for engaging in fine-grained processing (Viswanathan, 1997). Expertise and product familiarity may also moderate the meaningful number of categories used by consumers.

An interesting pattern of results is that the number of meaningful categories was consistently lower than seven and even five for the product attributes used here. In particular, three appeared to be the meaningful number of categories for several attributes. Perhaps, consumers often view products as being ‘‘high,’’ ‘‘medium,’’ or ‘‘low’’ on attributes. This pattern argues for the use of shorter scales than the 5- or 7-point scales often used in marketing research. Such scales may be easier to administer particularly in telephone surveys, easier for consumers to complete, and also be easier to label than longer scales.

It should be noted that we chose fairly low involvement products and consequently the product attributes are likely to have been of low relevance or importance compared to attributes of some other products. Moreover, we used attributes such as caramel flavor of a qualitative/sensory nature about which precise information is not available in the marketing environment. Because of the lack of importance of these attributes and their qualitative nature, consumers may have been less likely to make finer discriminations than were meaningful to them. Our choice of stimuli may also have quite possibly led to some degree of error in the measure of meaningful discrimination. Both these outcomes are likely to have made the tests of hypotheses more stringent, yet support was found for the hypotheses. The results may be stronger for attributes of more important product categories such as gas mileage of automobiles or calorie content of foods where consumers may be more likely to make finer discriminations along attributes than are meaningful to them. Along these lines, the use of students’ samples, and paper-and-pencil procedures, suggest the need to conceptually replicate these findings using different samples and methods of administration.

This study represents a first attempt to measure the meaningful number of categories and additional research is needed in this area. Future research should examine different methods including the

sorting task that can be used to study how consumers think about magnitudes along product attributes and assess the degree of convergence. Some potential methods include magnitude estimation scaling (Viswanathan *et al*., 1995), similarity judgments (Johnson and Fornell, 1987), elicitation of verbal labels (see Zimmer, 1984), and category scaling (see Park and Lessig, 1981). Other lines of future research include a focus on the relationship between the number of meaningful categories and scale responses. Such analyses could lead to techniques to extract data from scale responses at a level of discrimination that is meaningful to consumers. Future research should also focus on the relationship between the number of meaningful categories and properties of attributes (such as concreteness–abstractness of attributes) as well as characteristics of consumers (such as expertise). Such research could provide guidelines for the meaningful number of categories for types of attributes and groups of consumers. In conclusion, this research demonstrates the importance of employing a meaningful number of response categories in scales in order validly measure differences between products that are meaningful to consumers.

Acknowledgements

Madhubalan Viswanathan and Michael Johnson are extremely grateful for the invaluable opportunity to collaborate with Seymour Sudman who passed away in May 2000. It was indeed a distinct honor and a privilege to be his coauthors.

References

Bendig AW. Reliability and the number of rating scale categories. J Appl Psychol 1954; 38(1):38–40.

Block J, Buss DM, Block JH, Gjerde PF. The cognitive style of breadth of categorization: longitudinal consistency of personality correlates. J Pers Soc Psychol 1981; 40(4):770–9.

Bruner JS, Tajfel H. Cognitive risk and environmental change. J Abnorm Soc Psychol 1961; 62(2):231
41.

Champney H, Marshall H. Rater's minimal discrimination as a criterion for determining the optimal refinement of a rating scale. J Appl Psychol 1939; 23:323–31.

Chattopadhyay A, Alba JW. The situational importance of recall and inference in consumer decision making. J Consum Res 1988; 15:1–12.

Churchill GA, Peter JP. Research design effects on the reliability of rating scales: a meta-analysis. J Mark Res 1984; 21:360–75.

Cox E. The optimal number of response alternatives in a scale: a review. J Mark Res 1980; 17:407–22.

Feldman JM, Lynch JG. Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. J Appl Psychol 1988; 73:421–35.

Gardner RW, Schoen RA. Differentiation and abstraction in concept formation. Psychol Monogr. 1962; 76 (41 Whole No. 560).

Garner WR, Hake HW. The amount of information in absolute judgments. Psychol Rev 1951; 58:446–59.

Hinrichs JV, Novick LR. Memory for numbers: nominal vs. magnitude information. Mem Cognit 1982; 10(5):479–86.

Holyoak KJ, Mah WA. Cognitive reference points in judgments of symbolic magnitude. Cognit Psychol 1982; 14:328–52.

Jacoby J, Matell MS. Three-point Likert scales are good enough. J Mark Res 1971; 8:495–500.

Johnson MD, Fornell C. The nature and methodological implications of the cognitive representation of products. J Consum Res 1987; 14:214–28.

Johnson MD, Lehmann DR, Horne DR. The effects of fatigue on judgments of interproduct similarity. Int J Res Mark 1990; 7:35–43.

Komorita SS, Graham WK. Number of scale points and the reliability of scale. Educ Psychol Meas 1965; 25(4):987–95.

Matell MS, Jacoby J. Is there an optimal number of alternatives for Likert scale items? J Appl Psychol 1972; 56:506–9.

Miller GA. The magical seven plus or minus two: some limits on our capacity for processing information. Psychol Rev 1956; 63:81–97.

Parducci A. Category judgment: a range frequency model. Psychol Rev 1965; 72(6):407–18.

Park CW. A conflict resolution choice model. J Consum Res 1978; 5:124–37.

Park CW, Lessig VP. Familiarity and its impact on consumer decision biases and heuristics. J Consum Res 1981; 8:223–30 (September). Poulton EC. Bias in quantifying judgments. Hillsdale; Erlbaum, 1989.

Ramsay JO. The effect of number of categories in rating scales on precision of estimation of scale values. Psychometrika 1973; 38(4):513–32.

Schwarz N, Bienias J. What mediates the impact of response alternatives on frequency reports of mundane behaviors. In: Appl Cognit Psychol 4(1). US: John Wiley & Sons Inc, Jan–Feb 1990. p. 61–72.

Schwarz N, Hippler H-J, Deutsch B, Strack F. Response scales: Effects of category range on reported behavior and comparative judgments. In: Pub Opin Quart 49(3). US: University of Chicago Press, Fall 1985. p. 388–95.

Sudman S, Bradburn NM. Asking questions San Francisco (CA): Jossey-Bass, 1982.
Viswanathan M. Individual Differences in Need for Precision. Pers Soc Psychol Bull 1997; 23(7):717–35.
Viswanathan M, Childers T. The encoding and utilization of magnitudes of product attributes: an investigation using numerical and verbal information. Faculty Working Paper No. 92-0105. Bureau of Economic and Business Research, University of Illinois, 1992.
Viswanathan M, Childers T. Processing of numerical and verbal product information. J Consum Psychol 1996; 5(4):359–85.
Viswanathan M, Childers T, Nagaraj S. Using magnitude estimation scaling in marketing research: an application to understand how consumers think about product attributes. In: Proceedings of the American Marketing Association Winter Educators Conference vol. 6. Chicago (IL): American Marketing Association, 1995. p. 483–9.
Viswanathan M, Johnson M, Sudman S. Understanding consumer usage of product magnitudes through sorting tasks. Psychol Mark 1999; 16(8): 643–57.
Wyatt RC, Meyers LS. Psychometric properties of four 5-point Likert-type response scales. Educ Psychol Meas 1987; 47:27–35.
Zimmer AC. A model for the interpretation of verbal predictions. Int J Man–Mach Stud 1984; 20:121–34.