Just-About-Right (JAR) Scales:

Design, Usage, Benefits and Risks

Lori Rothman
Merry Jo Parker
Editors

ASTM INTERNATIONAL
Standards Worldwide

Just-About-Right (JAR) Scales: Design, Usage, Benefits, and Risks

Lori Rothman and Merry Jo Parker

ASTM Stock Number: MNL63

ASTM International
100 Barr Harbor Drive
PO Box C700
West Conshohocken, PA 19428–2959

Printed in the U.S.A.

Library of Congress Cataloging-in-Publication Data

Rothman, Lori

Just-about-right (JAR) scales: design, usage, benefits, and risks / Lori Rothman and Merry Jo Parker [sponsored by Committee E18 on Sensory Evaluation].

p. cm. — (ASTM international manual series; MNL 63)

"ASTM stock number: MNL63."
Includes bibliographical references.
ISBN 978-0-8031-7010-0
1. Consumers—Research—Statistical methods—Handbooks, manuals, etc.
2. Consumer goods—Evaluation—Statistical methods—Handbooks, manuals, etc.
3. Quality of products—Handbooks, manuals, etc. I. Parker, Merry Jo, II. ASTM Committee E18 on Sensory Evaluation of Materials and Products. III. Title.
HF5415.3.R6786 2009
658.8'34015195—dc22 2008041327

Printed in Bridgeport, NJ
February 2009

Preface

When we volunteered for this assignment, we did not know that it would take 10+ years to complete. Or that this document would become Manual 63. Originally conceived as a Standard Guide, the completed document was more than 200 pages in length. Several comments were received during the subcommittee ballotting process that the document was too large to be a Standard Guide; thus the idea for a Manual was born. Our goal was to create THE definitive document on Just About Right scales in an easily understood, practical format and we think we have succeeded. Please tell us if you think otherwise.

Lori Rothman
Kraft Foods
801 Waukegan Road
Glenview, IL 60025

Merry Jo Parker
Food Perspectives
2880 Vicksburg Lane North
Plymouth, MN 55447

Dedication

This manual is dedicated to our hardworking colleagues at ASTM E18.04.26.

Acknowledgments

This manual represents the fruits of more than 10 years of labor by a wide range of individuals associated with ASTM International Committee E18. Through the years, E18 committee members helped 'grow' the manual, from setting the scope and creating the outline to writing the chapters themselves. Countless versions were edited by the E18.04.26 task group members; each and every editing session made the document more cohesive and compelling. We are particularly grateful to the case study authors, who worked with a common data set to produce the fantastic array of statistical options presented. Many of those authors ventured from their comfort zones, working with novel techniques to showcase them in a format easily read and understood. Task force members evaluated and edited these case studies, making sure that each technique was fairly presented with all its pros and cons. Janya Walsh provided the first draft of the Bibliography. A special 'thank you' goes to Bill Raynor of Kimberly Clark for reviewing all the statistical case studies. Tom Carr of Carr Consulting and Greg Stucky of Insights Now also deserve special recognition for their contributions. Richard Popper of P&K provided enormously helpful specific feedback on the entire document, an invaluable contribution. Scott Orthey of ASTM was our cheerleader extraordinaire, particularly as we became close to publishing. Thanks to the (changing) officers of E18 for their continual support and encouragement and to the publication staff of ASTM International, especially Kathy Dernoga and Monica Siperko. Finally, the editors thank each other for their continued dedication, dogged persistence and friendship. Neither of us could have done it alone.

Lori Rothman
Kraft Foods
801 Waukegan Road
Glenview, IL 60025

Merry Jo Parker
Food Perspectives
2880 Vicksburg Lane North
Plymouth, MN 55447

TABLE OF CONTENTS

Structure and Use of Just-About-Right Scales

Lori Rothman[1] and Merry Jo Parker[2]

Introduction

Just-about-right (JAR) scales are commonly used in consumer research to identify whether product attributes are perceived at levels that are too high, too low, or just about right for that product. While at first glance this type of scaling may appear simplistic, the process of developing and analyzing JAR scales is complex. In this manual, the history of JAR scale usage, the mechanics of JAR scale construction, inappropriate and appropriate uses of JAR scales, benefits and risks of use, and case studies that illustrate the many methods for data analysis of these scales are reviewed. Alternatives to JAR scales are also included. Finally, a summary and conclusions are presented.

This manual covers the application, construction, analysis, and interpretation of just-about-right (JAR) scales for use in testing with consumers. Defined as bipolar labeled attributes scales, JAR scales measure levels of a product's attribute relative to a respondent's theoretical ideal level. These scales have an anchored midpoint of "just about right" or "just right" and endpoints anchored to represent intensity levels of the attribute that are higher and lower than ideal. The ideal point model [1] will serve as the conceptual base for JAR scale interpretation.

This model is one of several available; however, it is not practical to cover all aspects of ideal point modeling nor to consider other conceptual frameworks, such as preference unfolding [2] in this manual.

This manual was developed by members of Task Group E18.04.26, which is part of the ASTM Committee E18 on Sensory Evaluation and is intended for sensory and market research professionals responsible for consumer testing and interpretation of consumer data. The scope of the task group was to develop a practical manual that covers the executional aspects of JAR scales (construction, placement, and analysis) and identifies and discusses issues of validity and interpretation.

This manual does not discuss in detail psychological or psychophysical processes related to scaling, basic practices in consumer testing, or general statistical principles.

Terminology

For definitions of terms relating to sensory analysis, see Terminology E253, and for terms relating to statistics, see Terminology E456.

Definitions of Terms Specific to this Manual:

- *bipolar scale*, *n*—a scale where the end anchors are semantic opposites, for example, "not nearly sweet enough" to "much too sweet," and there is an implied or anchored neutral mid-point.
- *just-about-right scale*, *n*—bipolar scale used to measure the level of an attribute relative to an assessor's ideal level, having a midpoint labeled "just about right" or "just right."
- *optimum*, *n*—amount or degree of something that is most favorable to some end.

Significance and Use

Quantitative data obtained from consumers are often used to aid product development. Types of quantitative data collected may include hedonic or acceptability measures, preference measures, attribute intensity or strength measures, and just-about-right (JAR) measures.

While hedonic and attribute intensity scales provide distinct classes of information, JAR scales are intended to combine intensity and acceptability to relate the perceived strength of specific attributes to the respondent's theoretical optimum.

Market researchers have used JAR scales routinely since the 1960s [3,4], although they were not always referred to as JAR scales. More recently [5], JAR scales have been used in conjunction with preference or acceptance measures as a diagnostic tool to understand the basis for these hedonic responses, with the aim of providing information concerning which product attributes to adjust and in which direction (increase or decrease). In this way, JAR scales provide guidance to the product developer.

Because JAR scales combine attribute strength and appropriateness in one scale, questionnaire length can be minimized. JAR scales should be easily interpretable by respondents when the attributes themselves are easy to understand (salty, dark, tight, and so forth).

JAR scales are not without controversy as to their usefulness and validity in market and product development guidance research. As with all other types of quantitative scales, JAR scales have issues and limitations regarding their usefulness and interpretability: the ability of respondents to understand the attribute endpoint and interval labels, number of scale points, physical representation of the scale on the questionnaire (horizontal or vertical orientation), and the psychological issues common to scaling tasks in general [6].

In addition, JAR scale data collection may be specifically hampered by the researcher's ability to construct bipolar scales for the attributes of interest and demand characteristics of the task (decoding and recording an attribute's strength and acceptability at the same time), which may prove difficult for some respondents.

Another important limitation concerns the validity of data obtained using JAR scales when it is suspected that the responses are a result of cognitive rather than sensory processing within the respondent. For example, a respondent may like a dessert's flavor profile "as is," but when asked, she may rate it as "too sweet" because she believes that sweet foods are "bad."

A number of statistical techniques are available for analyzing data from JAR scales; they range from simple to very complex. The statistical analysis of JAR scale data should be

[1] *Kraft Foods, 801 Waukegan Road, Glenview, IL 60025.*
[2] *Food Perspectives, 2880 Vicksburg Lane North, Plymouth, MN 55447.*

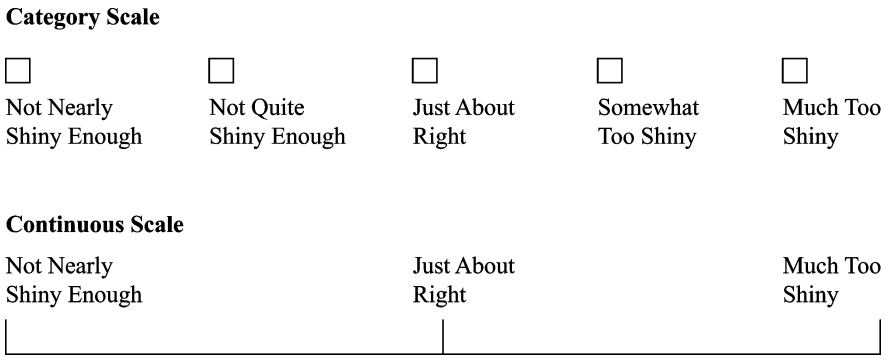

Fig. 1—Examples of Scale Types

chosen to match the objective of the portion of the research that concerns the JAR scales. Examples of objectives are listed in Table 1.

History of Jar Scale Usage

The genesis of JAR scaling is not well documented in published marketing or sensory research literature. A probable origin can be found in the attitude-measurement scaling methodologies developed by L. L. Thurstone and Rensis Likert in the 1920s and 1930s [7,8]. Thurstone contributed a differentiated attitude scale along a continuum, and Likert attached numbers to the levels of meaning to that scale. The Likert scale (five-point scale from "strongly agree" to "strongly disagree" with a "neutral" center point) is still widely used today.

It was not until the 1950s with the advent of Osgood's concept of "semantic space" and the development of "semantic differential scaling" that semantic opposites (bipolar scales such as "good-bad," "weak-strong," "soft-hard," or "rough-smooth") began to appear in measurement scales. These scales add three components of judgment: evaluative, potency, and activity that are used to measure attitudes and certain aspects of a product. In 1957, Osgood published his findings on the quantification of connotative semantic meaning in *The Measurement of Meaning* [9]. Semantic differential scales, however, are continuous and, unlike JAR scales, do not identify the "ideal point" for an individual.

It is unclear as to when and who began using the unique middle anchor of "just about right." Early references to discussions regarding the middle "ideal point" category date as far back as the 1950s [3,10–12]. In 1972, Moskowitz presented the idea of using JAR scales and self-designed ideals in the optimization of perceptual dimensions in food [13]. By the early 1980s, the use of scales with the center anchored with "just right" were reported by others [14–16].

From the 1980s to the present, sensory research on the use of JAR scaling continues. Of particular interest has been the relationship between JAR scales and hedonic and intensity scales, preferences, and consumption patterns. A number of authors have proposed the use of JAR scales as an alternative to hedonic scaling for determining the ideal level of an ingredient [16–20]. Others, however, either did not find agreement between JAR and hedonic scores [21] or found cause to question the order effect of one upon the other [22].

By 1999, JAR scales were reported to be widely used for product guidance.[3] However, their use and interpretation remains a controversial topic in sensory science [23], partly because of the type of judgments the respondents must make.

Mechanics of JAR Scale Construction

Questionnaire Development Using JAR Scales–When developing a questionnaire that will contain JAR scales, the researcher should consider what other scale types to include and their relative locations on the questionnaire. These will depend on the overall test objective and how the data will be used. A typical questionnaire for a consumer test may include hedonic as well as JAR responses. While it is possible to construct a questionnaire consisting solely of JAR scales, without a concomitant collection of hedonic information the researcher will be unable to relate JAR scale data to hedonic ratings. If the researcher wishes to understand respondent ratings of attribute strength, intensity scales may be included on the questionnaire along with JAR scales. Alternatively, descriptive data from a trained panel may be collected separately and examined with the JAR scale ratings.

Issues Related to Scaling

Scale Type–Sometimes known as "directionals," JAR scales are constructed as bipolar category or continuous line scales, with a midpoint labeled Just Right or Just About Right. In the case of a category scale, responses are limited to the number of points chosen for the scale; responses for continuous scales are theoretically infinite (see Fig. 1).

Number of Scale Points–If used as a category scale, the minimum number of scale points for a JAR is three. Because the center point of the scale is the respondent's optimum (ideal) point, the number of scale points is always odd. The scale is "balanced" in that there are an equal number of points on either side of the midpoint anchor. While there is no absolute maximum, in practice, the number of scale points is rarely greater than nine, unless a continuous line scale is used. There may be a law of diminishing returns associated with large numbers of scale points [24]. Although three is the minimum number of points, many researchers are uncomfortable with only three points because of "end

[3] Moskowitz H. R., "On the Analysis of Product Test Results: The Relation Among Liking, Sensory and Directional Attributes," unpublished, http://www.mji-designlab.com/articles/lang6htm

3 pt. JAR	5 pt. JAR	7 pt. JAR	9 pt. JAR
Too Weak	Much Too Weak	Much Too Weak	Extremely Too Weak
Just About Right	Somewhat Too Weak	Moderately Too Weak	Much Too Weak
Too Strong	Just About Right	Slightly Too Weak	Moderately Too Weak
	Somewhat Too Strong	Just About Right	Slightly Too Weak
	Much Too Strong	Slightly Too Strong	Just About Right
		Moderately Too Strong	Slightly Too Strong
		Much Too Strong	Moderately Too Strong
			Much Too Strong
			Extremely Too Strong

Fig. 2—Examples of Number of Scale Points

avoidance" [24], which can force respondents to the center of the scale. A 1999 ASTM International survey[4] of sensory professionals revealed that 52 % of those responding use a five-point JAR scale exclusively and 32 % use more than one type of JAR scale depending on the respondents and test objectives. Respondent qualifications may lead one toward a specific number of points, for example, when testing with young children, researchers may use three-point JAR scales to simplify the task [24] (see Fig. 2).

Anchors–The midpoint of a JAR scale is reserved for the rating Just About Right or Just Right. Some researchers feel that the midpoint should be labeled Just Right, which implies that any deviation from "ideal" should be captured by a point outside of the center; other researchers feel that Just Right entails too strong a commitment on the part of the respondent [24], so the center choice is rephrased Just About Right. The scale end points are anchored in either direction from the scale midpoint with additional labeling of points as desired by the researcher. One side of the scale is the "less than just right" side, while the other is the "greater than just right" side. Some researchers have demonstrated greater scale reliability as a function of additional anchoring [25], while others suggest that word anchors may not represent equal spacing between points and avoid anchoring all but the end and center points [23]. When presented as a category scale, the scale points are presented as being equidistant from each other, although this may not be true psychologically. For example, the psychological distance between "somewhat too salty" and "much too salty" may be perceived as greater than the distance between "just right" and "slightly too salty." Scales that incorporate true equal interval spacing can be developed using Thurstonian methods [26]. When selecting the appropriate scale anchors, several approaches may be used. One approach is to use terms with clearly defined and understood semantic opposites that are consistent with the makeup of the product and general consumer recognition. When dealing with the fit of a garment, for example, "too tight" and "too loose" are generally accepted as semantic opposites, as are "too thin" and "too thick" when dealing with food texture. "Too sweet" and "too sour," on the other hand, would not be considered semantic opposites. The lack of opposites occurs for the chemical senses of olfaction, taste, and trigeminal, where zero or no intensity is a common occurrence for some products. Where no semantic opposite exists, one approach is to use the same attribute term on both sides of "just about right" ("not sweet enough," "too sweet," see Fig. 3). Another approach is to use generic terms as scale anchors, such as "too weak" and "too

Sweetness

Not Sweet Enough	Just Right	Too Sweet
OR		
Too Weak	Just Right	Too Strong

Sourness

Not Sour Enough	Just Right	Too Sour
OR		
Too Weak	Just Right	Too Strong

Fig. 3—Attribute Examples

[4] This survey was conducted in 1999 by the E18.04.26 Just About Right task group Vice Chair Merry Jo Parker. The survey was sent to all E18 members and was posted on Sensory.org. There were 77 responses to this survey.

strong," and "not enough" and "too much/many." The specific attribute would be positioned above the scale, as illustrated in Fig. 3.

Degree-of-Change Scale–The degree-of-change scale uses alternate instructions to the respondent and different anchors from the typical JAR scale to change the scale from being evaluative to action oriented. Instead of being instructed to provide an opinion as to the degree of saltiness ("not nearly salty enough" to "much too salty"), the respondent is asked how he would change the saltiness of the product. An example of a nine-point degree-of-change scale would include the question:

How would you change the saltiness of this product? The responses are:

- Decrease it extremely
- Decrease it very much
- Decrease it moderately
- Decrease it slightly
- Leave it the same
- Increase it slightly
- Increase it moderately
- Increase it very much
- Increase it extremely

Pokorny and Davidek [27] present a similar scale for attribute optimization. It has been suggested that this scale may be easier for the respondent to understand than a typical JAR scale as it is action oriented.[5] While the modes of analysis for these scales do not differ from those for JAR scales, the psychological processes involved with use of such scales are not known and have not been thoroughly researched. Therefore, researchers are encouraged to proceed with caution.

Issues Related to Attributes

Attribute Selection–Attributes for questionnaire inclusion should be easily understood by respondents (common language); technical or industrial jargon should be avoided ("too rheopectic"). Terms should be as specific as possible to avoid confusion; for example, the attribute "amount of chocolate" in chocolate ice cream that contains chocolate chips may confuse the respondent as to whether the researcher is asking about the strength of chocolate flavor in the base ice cream, the number of chocolate chips, or some combination of the two. Other terms that connote multiple meanings or sensory modalities such as "hot" or "spicy" should be avoided or explained further on the questionnaire ("heat/burn felt in the mouth," "strength of spice flavor").

Source of Attributes–The proper selection of attributes is central to obtaining usable results. Users of JAR scales develop a list of product attributes that are of interest to the researcher. The source of these attributes may include those that have been shown to be important based on prior product testing (including information obtained from descriptive panels), perceived characteristics of key ingredients, attributes that are suspected of interacting with other ingredients, or attributes taken from product advertising or claims. Additionally, qualitative research or other techniques (focus groups [28] repertory grid [29], or free choice profiling [30] may be used to elicit attributes from respondents before testing.

Single Attributes–Single attributes refers to scales with the same adjective on both end points ("too sweet," "not sweet enough"). The research aim is to select attributes whose intensity increases on one continuum, avoiding complex terms that relate to multiple sensory properties of the product or that may have more than one meaning to respondents. Such terms include "creamy," "rich," and "chewy." "Creamy" can relate to product appearance (color, opacity), flavor (cream, dairy), feel in mouth (smooth, thick), or some combination. One exception would be when the research goal is to understand consumer use of a complex term compared to terms used by a trained panel for descriptive analysis, or if prior studies have confirmed consumer use of the term, in which case use of the complex term is acceptable. Another exception may be to use the term in more than one location on the questionnaire with a specific modifier, such as "creamy appearance" or "creaminess in the mouth." If complex attribute scales are included, additional simpler terms may be included on the questionnaire to understand, during analysis, whether and how the simpler terms relate to the more complex terms. An example would be to include "cloudiness," "dairy flavor," "thickness," and "smoothness" along with "creaminess" to understand how the former terms contribute to the latter. Keep in mind that such research necessitates an increase in questionnaire length. There are other ways to understand complex attribute terms, including combining results of descriptive analysis with the data from the complex term JAR scale.

Combined Attribute Scales–Combined attribute scales denote scales whose endpoint anchors differ. These are problematic in JAR scale construction and should be avoided, except in cases in which there is a clearly defined semantic opposite ("too loose"/"too tight"). Combined attribute scales require assumptions about the relationship between those attributes that may or may not be true and eliminate the possibility that both qualities vary independently from "just about right." Consider the following combined attribute scales:

- Too sour-JAR-Too sweet
- Too soft-JAR-Too rough
- Too dry-JAR-Too greasy

A literal reading of these attribute scales reveals the following assumptions:

- A product that is "not sour enough" is "too sweet,"
- A product that is "not soft enough" is "too rough," and
- A product that is "too dry" cannot also be "too greasy."

Although these relationships may be true for a given product, they are not recommended for use in JAR scales except in situations in which the assumption is true. An example of such a scale would be "too loose" to "too tight," where, in fact, something that is not "too loose" or "just right" is "too tight." An alternative to combined JAR scales would be the use of two scales; in the first example, one scale ranging from "too sweet" to "not sweet enough" and a second scale ranging from "too sour" to "not sour enough." While this procedure will increase the length of the questionnaire, it is far better to have additional questions than to collect data that are not interpretable.

[5] Private communication.

Attributes with Negative or Positive Connotations–Attributes with inherent negative connotations should be used with caution. Respondents may find it difficult to rate coffee or chocolate as "not bitter enough" or salad dressing as "not sour enough" because of a perception that "less is better," and the JAR scale may lose its usefulness. Alternative procedures exist for handling inherently negative attributes. These include respondents rating attribute intensity levels with subsequent analysis to determine impact on overall liking, relating descriptive data to acceptability, or direct hedonic comparison of products with differences in the attribute of interest. Similarly, attributes with positive connotations (natural, blended), particularly those with no clear or direct association with specific sensory attributes of the product, should not be used with JAR scales.

Number of Attributes–In most cases, several attributes are studied in a single test for a variety of reasons: to provide direction for multi-attribute products, to identify attributes that have an impact on acceptability if their intensities are not "just right," to avoid using combined attribute scales, and to ensure all of the key attributes are included to avoid misattribution of perceptions (see p. 7, *Misattribution of Perceptions*). The overall length of the questionnaire and the number of attributes to be included should be considered to avoid issues of respondent psychological and sensory fatigue. Redundant attributes should not be included (such as scales "not thin enough" to "too thin" and "not thick enough" to "too thick," for example, where "thick" and "thin" could be considered as semantic opposites) unless one purpose of the research is to identify redundant attributes.

Location of JAR Scales on the Questionnaire

JAR scales are typically located on the questionnaire after hedonic ratings such as overall liking or liking of specific product attributes. Some researchers, however, place the overall liking subsequent to the liking ratings of specific product attributes; in these cases, the JAR attributes would precede the overall liking question. Gacula et al. [31] found JAR data to be statistically uncorrelated with liking data, suggesting that JAR scales can be placed prior to overall liking on the consumer ballot. It is common practice to have the order of the JAR scales roughly coincide with respondents' sensory experience of the product: for example, in the case of food items, appearance attributes, then aroma attributes, then flavor attributes, then oral texture attributes. In the case of nonfood (for example, facial tissue), the order might be appearance, feel by hand, aroma, and feel in use, followed by specific usage attributes. If intensity and JAR scales are included on the same questionnaire, it is common practice that respondents answer all questions relating to the same attribute, whether intensity or JAR, before moving to the next attribute, to ensure respondent focus and avoid confusion or fatigue. For example, it would be better to have the attributes ordered such as sweet flavor strength, sweet flavor JAR, salty flavor strength, and salty flavor JAR as opposed to sweet flavor strength, salty flavor strength, sweet flavor JAR, and salty flavor JAR. Some researchers additionally include specific attribute liking questions. In this case, a commonly used question order would be attribute liking; attribute strength; attribute JAR, for example sweetness liking; sweetness strength; and sweetness JAR. There are, however, some researchers who group all attribute liking questions together before any intensities or JAR scales.

Appropriate/Inappropriate Use of JAR Scales

JAR scales are appropriately used when the objective is to understand the respondents' reactions to the intensity or strength of specific attributes, compared to the level that they believe would be "ideal." These scales are particularly useful when product component levels have not been varied systematically, when testing competitive products or for single prototype evaluation. In these cases, respondents cannot respond directly to changing component levels via their hedonic responses. When samples are varied in an experimental design [43], JAR scales may not be necessary. In these systematic studies, the researcher is often interested in direct intensity or strength measurements by respondents to demonstrate the impact of the variable ranges on the product. The "optimal" product will be inferred from modeling overall liking (see p. 8, *Designed Experiments*), and JAR scales may provide no additional benefit. It has recently been demonstrated, however, that JAR data may be used to predict optimal ingredient levels within the sensory space of a designed experiment and thus may be a useful tool in this regard [32].

JAR scales are not intended for use with trained panels because the trained panel is used as an unbiased instrument for scaling attribute intensities, not for judging whether the intensities are too high or too low.

Benefits of Use

JAR Scales Provide Formulation Guidance

JAR scales are particularly useful in situations in which product attributes cannot be varied systematically. JAR scales will provide guidance as to which product attributes are at an "ideal" level and which are not. When an attribute is not JAR, the JAR scale can provide information as to the direction of possible change.

Through the use of JAR scales, one response can represent the combination of attribute strength and hedonic judgment that can be combined to provide directional information for product formulation or optimization [24]. If, for example, a tested product received a low hedonic score and was rated "too salty" by 50 % of respondents (and it had been inferred or determined that salt level impacted the hedonic score), the researcher would most likely lower the salt level prior to subsequent testing. In this way, information from JAR scales provides actionable guidance for product development.

JAR Scales Are Easily Understood

Another benefit is that respondents easily understand the scale itself, assuming that the JAR scale has been designed correctly and the attributes are properly chosen. The scale is also generally understood by other stakeholders in or clients of the organization conducting the research, making results easy to communicate.

Additional Benefits of Using JAR Scales

JAR scales may also assist in defining or identifying consumer segments in which segmentation is based on a product's "ideal" sensory profile. If JAR scale data are bimodal (a

product is rated both "too cheesy" and "not cheesy enough"), one hypothesis is that respondents differ in their ideal attribute levels indicating the possibility of sensory-based consumer segments. Bimodal responses could also point to an issue with the specific type or character of the attribute (Swiss versus cheddar cheese, for example) and the need for subsequent reformulation. Issues associated with bimodal responses may be more easily identified with JAR scales than with other scale types, such as intensity scales.

JARS Aid in Understanding Which Attributes Impact Liking

When used in conjunction with liking scales, JAR scales have the potential to aid in understanding which attributes have the greatest impact on liking; additional information on which end of the JAR scale to avoid (too strong or too weak) can also be determined. Researchers can use this information to prioritize which attributes to adjust during reformulation, and how to adjust them. Additionally, attributes that are not JAR but which have no impact on product acceptance can be ignored.

JAR Scale Data Analyses Range from Simple to Complex

With a variety of methods available to analyze JAR scales, ranging from simple graphical representation to complex statistical methods, researchers can customize the analysis depending upon project objectives, level of expertise, the amount of statistical resources available, and objectives of the project.

Risks of Use

Risks Common to All Scales

Attribute scaling involves the assignment of numbers to quantify sensory experience [24]. The following serves as a brief review and reminder that many of the risks associated with JAR scale usage are common to all scaling techniques. Potential remedies are also mentioned for each risk. See Lawless and Heymann [24] for a more thorough discussion of scaling risks.

Halo/Horns–With this risk, the respondent assumes that an overall positive or negative impression extends to the rating of subsequent attributes, whereby high product acceptability may lead to false ratings of "just about right." In this case, the respondent may wish to avoid the cognitive dissonance that may result from seemingly disparate ratings. Similarly, if the respondent does not like the product overall, assigning a positive rating on any attribute may be difficult. While we assume that respondents separate liking of a product from attribute evaluations, the inherent nature of JAR scales implies a degree of affective response that may make halo/horns a bigger risk for JAR scales than for other types of scales. Analyzing JAR scales in conjunction with liking data can aid in differentiating the halo/horns effect from attribute imbalances that affect liking.

Relevance–An attribute that has low relevance in the respondent's evaluation criteria set may receive a rating of "just about right" simply because it is unimportant, a way out of needing to make a determination. Prior knowledge or research can help to minimize the inclusion of irrelevant attributes.

Expectation Errors–A respondent may react to an attribute based on expectations, as opposed to perception. An example would be a rating of "not salty enough" to a hot dog that contained less than average salt. The level of salt may actually be "just right" for the respondent, but she may rate it as "not salty enough" because her expectation is for a saltier hot dog. Depending upon the test objective, the addition of a concept or positioning statement to the test may reduce this effect ("Now hot dogs with less salt!").

Contrast Effects–A contrast effect occurs when a respondent rates an attribute based on prior products evaluated, thereby exaggerating differences between products. For example, a respondent may receive an extremely large product, followed by a moderately large product. The first product might be rated "too large," with the second product being rated as "too small," because of the contrast between the first and second products received. Served monadically, the second product might be rated "just about right." This could lead to an inappropriate "fix" of the product by the developer. Combating contrast effects typically involves rotating the order of product evaluations. However, this presupposes that the contrast effects "balance out" over evaluation order, which may or may not be true. A total avoidance of contrast effect would involve pure monadic product evaluation, which is generally not a cost-effective option. However, examination of data by position served could aid in interpretation. Additionally, having the samples spaced equally apart in terms of expected attribute strength is desirable. For example, in a set of low-sodium samples, including one sample at a more typical sodium level will tend to lessen the rated saltiness differences among the lower sodium samples and thus should be avoided.

Context Effects–Similar to contrast effects, context effects refer to the influence of the specific set of samples included within the evaluation framework, even when balancing the sample serving orders. This effect typically occurs after evaluating several samples that are of similar attribute strength; while the respondent may not experience sensory specific fatigue, the attribute level in question may seem more "normal" or "average" than it would if experienced in a different set of samples. Consider a set of reduced salt products; samples rated earlier in the evaluation may be rated as "not salty enough," while samples served later in the order may be rated more "just about right," as the respondents frame of reference shifts more towards less salty products as being "typical" within the set. This would occur regardless of the evaluation order, as all of the samples would have reduced salt. Similarly, if a set of samples all with very high sugar content were evaluated, after a number of sample evaluations, the sweetness might not "stand out" as much as it had in the prior samples. Potential remedies for this risk include limiting the number of product evaluations in one session or including a wide range of attribute levels so that the context for the attribute encompasses "all" reasonable levels. This last remedy, however, may not be practical for most research studies.

Range/Frequency–The relative spacing and frequency of stimulus/attribute levels within a product test may result in context effects that produce biases in category scales [33]. Models for these biases have been described [34]. Centering bias may occur when products with different intensities of

an attribute are evaluated in one test [14]. "Just about right" scales are not exempt from centering and frequency biases. When this occurs, the respondent may rate the intermediate strength product as "just about right," misrepresenting the true optimum strength. Alternately, within a given stimulus range, bias in determining the "just right" stimulus or attribute level may occur when more samples are presented on one side of the stimulus range [35]. In that case, the bias would have respondents rate as "just right" samples at the more heavily represented end of the range. Parducci and Perrett [33] also discuss the tendency among respondents to use different parts of the scale with equal frequency. Johnson and Vickers [36] have confirmed methods suggested by Poulton [34] and McBride [14] to avoid the centering bias. When possible, common sense would dictate that attribute intensities be equally spaced, in an attempt to counteract the frequency bias. While "just about right" scales are not exempt from biases common to all attribute scales, careful attention to the range and frequency of attribute intensities in the prospective products, coupled with a prior awareness of the possibility of such biases, will assist in robust data collection and evaluation.

Attribute Interpretation–It is assumed that the respondent interprets the attribute as the researcher intends. Misunderstanding of the attribute by the respondent may lead to false conclusions concerning the acceptability of attribute intensities or a bimodal distribution of responses. For example, a product may be rated as "too bitter" because respondents may confuse bitterness with sourness or astringency. This could lead to unnecessary reformulation. Or respondents may not like a particular flavor *character*, even though the flavor *level* may be appropriate. Consider a product with a bimodal distribution of "not fruity enough" and "too fruity" because the fruit note, while moderate in strength, is "unpleasant." In this case, some respondents would rate the fruity note as "too strong" because of its unpleasantness, while others would rate it as "too weak" because there is not enough "pleasant fruity taste." In fact, a bimodal response distribution in JAR scales may be an indication of misinterpretation of an attribute. Care should be used in reaching this conclusion, however, because a bimodal response distribution may also indicate the presence of consumer segments (see p. 8, *Consumer Segments and JAR Ratings*). A bimodal response distribution may also indicate differential sensitivity in attribute detection among respondents.

Misattribution of Perceptions–The lack of opportunity to rate an attribute that is important to the respondent can lead to misattribution, whereby perceptions not explicitly measured result in perceptions being expressed in one or more unintended attributes. An example of this may occur if a respondent perceives a product as being too sweet and too sour in which the only relevant attribute on the ballot is sweetness JAR. In this case, the respondent may be confused; if she rates the product "too sweet" she might be afraid that the product will be inferred to be "not sour enough." To combat this, she rates the product as being "too salty" to capture the sourness strength, making this rating subconsciously or perhaps second guessing her perception. Although this may be an extreme example, it illustrates the necessity of thoroughly understanding the perceptible attributes of the product under study and deciding which attributes need to be included on the questionnaire.

Risks Unique to JAR Scales

Because JAR scales are intended to *combine* intensity and acceptability to relate the perceived strengths of specific attributes to the respondents' theoretical optimum, there are additional risks unique to JAR scales that researchers should consider. Potential remedies for each risk are included.

Cognition Versus Perception–Respondent confusion of perception with cognition can happen when an attribute carries negative connotations. Examples may include attributes such as "salty" or "sweet," which may have negative health connotations. A respondent may rate a product as "too salty" or "too sweet" because the respondent believes that ingredients that cause sweet and salty tastes are "unhealthy;" however, this product may in actuality be preferred to a less sweet or salty product by the same respondent. In this case, the JAR scale would not have provided valid direction for product development. Understanding the relationship between attribute skews and overall liking would be helpful in teasing out this misperception, as there may be no impact on overall liking associated with the attribute skew. Another example occurs when nonsensory judgments are considered along with the attribute in question. For example, a respondent may rate the tongue burn of a mouthwash as "too strong;" however, that level of burn may be associated with efficacy. In this case, a reduction in burn may result in a reduction in perceived efficacy. The inclusion of efficacy ratings would be appropriate in helping to understand this trend.

Never Enough "Just About Right" Attributes–There are some attributes that respondents typically rate as "not enough" on JAR scales, such as inclusions like nuts in chocolate bars or ice cream, toppings like pepperoni on a pizza, or characterizing flavors such as "cheesy." This may lead to an effect, similar to that discussed previously, in which JAR ratings are given based on cognition versus perception. Increasing the component in question may, in fact, decrease product acceptability, as the increased component may lead to flavor or texture imbalance. Linking the JAR ratings with hedonic measures may help identify attributes in which this effect exists.

Respondent Biases–Respondents may assume that at least one product under evaluation is "just about right" or that it is unlikely that all products are "too strong" or "too weak"' on some attribute, when in reality, that may be the case. They may assume that a product that is well liked is "just about right" on all attributes. Or, a respondent may assume that if she rates a product as "just about right" on all attributes, then he/she shall like it more than a product that she rated as "too high" or "too low" on one or more attributes. Including a sufficient number of respondents in the research will lessen the impact of this effect.

Attribute Interdependence–For JAR data to be actionable, the researcher should understand how formulation elements interact to influence attribute ratings. For example, a researcher may have several alternative solutions to the problem of a "too sweet" product, including adjusting the sweetness, sourness, saltiness, texture, or flavor system as a whole. Not understanding these interactions may limit the

developer's ability to respond appropriately to data obtained from JAR scales. Additionally, making changes to one attribute to move it closer to "just about right" may move another attribute away from optimum. Moskowitz [37] created models for product optimization using liking scores generated from an array of products that constituted an experimental design. The predicted optimal product did not generate estimated mid-point JAR scale values ("just about right") for many product attributes. Similarly, when the data were reverse engineered to set JAR scales to scale midpoints (optimal attribute levels), the estimated product profile was not that of the optimal product. Apart from the fact that these results were derived from mathematical formulas (although the original data were obtained via experimentation), it is possible that the disparities in reaching optimal overall acceptability versus optimal individual attribute levels is due to what Moskowitz et al. [23] referred to as "Contradictions and Trade-offs." When respondents rate a product on a number of JAR scale attributes, each attribute is considered separately. However, when altering the formulas in response to respondent input, there are often trade-offs between respondents' desires and formulation constraints. The result is that some attributes may not be delivered at the target level. A thorough understanding of the interactions between the product components and anticipated changes during reformulation will aid in lessening the number of off-target attributes.

Respondents May Not Know Their Ideal Attribute Level–Some researchers question if the respondent really knows his or her ideal level of an attribute. It is possible that a respondent may think that he would prefer a darker product, but if such a product were delivered, it might be "too dark" even if the other product attributes remain unchanged. One remedy for this risk is to understand, through prior consumer testing history, which attributes are prone to these effects.

Relating Responses to Level of Change Desired–A developer may incorrectly assume that a large number of "too much" responses suggests a larger decrease in an ingredient than a smaller number of "too much" responses, or that a larger number of "too much" responses suggests a larger impact on overall liking or preference than a smaller number of "too much" responses. It is difficult, if not impossible, to relate the level of adjustment to the distance of the attribute from "just right" [38].

Temporal Aspects of Product Testing–The amount of product evaluated and the time period over which a product is evaluated may influence the JAR scale ratings. An attribute may be "just right" when consuming a small amount of product, but may prove to be "too strong" or "too weak" when consuming a full serving or with continued exposure. Or, a product may have a strong initial impact, but be more balanced over a typical consumption pattern. These aspects of product testing should be considered when examining the data from JAR scales. Additionally, products that are not well blended (such as fat-free or reduced fat products) may have flavor "spikes" causing responses on both sides of "just right." These aspects of product testing should be considered when examining data from JAR scales.

Effect of Product Reformulation on Subsequent Respondent Ratings–Based on JAR scale data, products may be reformulated to satisfy respondents who found the product not "just about right." These reformulations may not take into account the effect of the reformulations on those respondents who originally rated the attribute as "just about right" or who liked the attribute strength in earlier evaluations. There are statistical methods of accounting for this potential issue (see Appendix P and Appendix Q case studies).

Consumer Segments and JAR Ratings–Lawless and Heymann [24] suggest that JAR scale ratings may mislead product developers to conclude that a homogeneous population exists where one does not. Consider the case in which two groups of respondents rate a product as "just about right;" one of the groups thinks the product is strong and the other group thinks the product is mild, although both believe the level is just about right. The solution for this risk would be to include intensity or strength evaluation along with the JAR scale ratings. However, knowing that there are consumer segments that vary in perception as opposed to acceptability would likely not affect a decision concerning the product. It is true, however, that without the inclusion of intensity or strength ratings, respondents' perceptions of attribute levels are not known.

Effect of JAR Scale Inclusion on Hedonic Ratings–There is some evidence [22,39] that the inclusion of JAR scales on the same questionnaire with hedonic scales may alter overall liking ratings compared to those ratings generated in the absence of JAR scales. This effect did not appear when intensity scales were included (as opposed to JAR scales), using the same attributes. Subsequent researchers [40,41] did not find this same effect consistently. The effect of JAR scale inclusion on hedonic scaling needs further study.

Remedies for the Risks–While the prior sections on risks may make a researcher cautious in using JAR scales, these risks can be minimized through judicious ballot construction, data analysis and interpretation, and taking the steps recommended in each risk section.

Alternatives to JAR Scales

Even with cautious questionnaire construction, careful data analysis, and interpretation, not all researchers will feel comfortable using JAR scales for product testing; alternatives exist that obviate the need for JAR scales. Examples of analysis of data sets that use alternatives to JAR scales are included in Appendixes X, Y, and Z.

Collecting Intensity and Hedonic Information as Separate Scales–The information obtained from JAR scales can be approximated by collecting data from a series of products in which attribute strength levels and attribute liking are collected for each product. The attribute intensities can be regressed on the attribute (or overall) liking scores. In this way, the attribute intensity level that is associated with the highest attribute (or overall) liking can be determined [42]

Designed Experiments–Designed experiments may be used to optimize a single product attribute or to optimize simultaneously multiple product attributes that may interact with one another. For a single attribute, a series of products is prepared that systematically varies a formulation or processing variable that would be expected to vary the perceived intensity of the attribute. At a minimum, overall liking is collected, although many researchers also collect attribute strength and attribute liking ratings. Data are analyzed to

identify optimal attribute strength and formulation variable. There are a number of experimental designs that are used from a simple design with four samples to very complex designs with a large number of samples [43].

Ideal Scaling–In place of attempting to gauge attribute intensity and acceptability in one scale, ideal point modeling involves separating out the hedonic component of the response from the intensity evaluation [41]. In essence, the respondent is asked how strong the attribute is on an intensity scale, followed by how strong they would like that attribute to be. The flow of questions would therefore be:

- How sweet is this product? (intensity scale of choice)
- How sweet is your ideal product? (intensity scale of choice)

Mean product attribute ratings are then compared to the mean "ideal" product attribute ratings, which serve as product benchmarks and provide product developers with direction. A comparison of the responses to these two questions gives an indication of the direction in which the attribute should be moved and, to some extent, the magnitude of the desired shift. It is postulated that the greater the distance between the perceived and ideal intensities, the greater the change that must be made to adjust the attribute.

Asking liking of the strength of the attribute may precede these questions. In this case, the flow of questions would be:

- How much do you like or dislike the "attribute" of this product?
- How "attribute" is this product?
- How "attribute" should this product be?

Response to the prior liking question, if asked, may suggest the significance of the discrepancy between the perceived and ideal intensities with respect to product acceptance and may provide product developers with a broader picture of product performance. When using this technique, it is possible that respondents may rate the ideal level of certain attributes as unrealistically high or low, which may lead to erroneous reformulation.

Data Analysis Techniques

There are a number of methods available to the researcher for analysis of JAR scale data and several for analyzing data with alternatives to JAR scales. As with any scaling technique, proper analysis is critical for drawing appropriate conclusions from the data. While analysis of JAR scale data ranges from simple to very complex, the interpretation of such data should be considered carefully.

Data Analysis of JAR Scales–When choosing a method for JAR scale analysis, it is important to consider what specific question(s) the researcher wants to answer. As a first step, the researcher should consider whether the question(s) to be answered involves only the JAR data or whether the question is based on relationships between JAR and other data, such as product acceptability. Table 1 lists commonly used methods for JAR scale data analysis in two columns depending upon whether the question and subsequent analysis method involves only JAR data or the relationship between JAR and other data. The specific question that each method of analysis intends to answer is also included. Each method for analysis is presented in the same format: an introduction and objective section, which gives background for the analysis as well as what the method purports to determine; a requirements section, which specifies the type of data needed; a "how to" section, which describes the mechanics of the analysis; a results and conclusions section, which discusses the results of the analysis when applied to the case study data; a pros and cons section, which underscores benefits and caveats associated with the analysis; and finally, a recommendation section, which discusses when the method should be used (if at all).

Description of the Dataset Parameters

These data are from a five-sample sequential monadic test with $N = 119$. The data from three samples were used for the analyses. Sample 170 is bimodally distributed for overall liking, Sample 896 is normally distributed, and Sample 914 is not fully bimodal but is also not normal.

The samples are variations of the same type of product.

Attribute descriptions were changed to generalize the product, but the liking and JAR data are related to the same attributes. Figure 4 outlines the dataset attributes and scale anchors.

Methods of Analysis Involving Only the JAR Scales

Graphical Methods–These methods involve visual examination only:

- Graphical data display (see Appendix A).
- Graphical scaling (see Appendix B).

Nongraphical Methods–These methods require computation:

- Percent difference from norm and percent difference from just right (see Appendix C).
- The mean (see Appendix D).
- Mean directional and mean absolute deviation (see Appendix E).
- Mean versus scale midpoint (see Appendix F).
- Cochram-Mantel-Haenszel (CMH), Stuart Maxwell, McNemar, and Chi-square (see Appendix G).
- Proportional odds/hazards model (see Appendix H).
- *t*-tests (see Appendix I).
- Analysis of variance (ANOVA) (see Appendix J).
- Thurstonian ideal point modeling (see Appendix K).

Methods of Analysis Involving Data Relationships

The following methods relate JAR to data obtained from other scale types, most commonly liking measures.

- Penalty or mean drop analysis (see Appendix L).
- Adding significance measures to penalty analysis (see Appendixesdevand O).

The case studies, Appendix N and Appendix O, provide several methods for testing the significance of an observed mean drop (in other words, they determine whether the difference in mean scores among those rating the product "just about right"" and those rating the product "too weak" (for example) is statistically significant). Neither case study requires that the response data be normally distributed because they directly approximate the variability of the mean drops; however, interval level data is required.

The first case study, Appendix N, transforms the JAR variable to dummy variables, and then creates a regression model for each variable. The regression coefficients from these models are taken to be unweighted penalties. The case

TABLE 1—Methods for analysis of JAR Scale Data.

Questions Involving Only JAR Data	Questions Involving Data Relationships
What does the JAR Distribution Look like? • Graphical Data Display (Appendix A)[a] • Graphical Scaling (Appendix B)[a]	How did liking change among those not finding the product JAR? • Penalty or Mean Drop Analysis (Appendix L)[b] • Significance of Penalties (Appendix N)[b] • Bootstrapping Penalty Analysis (Appendix O)[b]
Are there enough JAR Responses compared to what is expected? • % Difference From Norm (Appendix C)[a] • % Difference From Just Right (Appendix C)[a]	What is the potential effect of attribute adjustment on attribute likers? • Opportunity Analysis (Appendix P)[b]
Are there an equal number of responses on either side of just right? • Graphical Scaling (Appendix B)[a] • Mean (Appendix D)[a] • Mean Directional (Appendix E)[a] • Mean Absolute Deviation (Appendix E)[a] • Mean vs. Scale Midpoint (Appendix F)[a]	What is the predicted effect on the entire distribution of JAR scores if the non-JAR attribute skews are reduced? • Product Improvement Analysis (PRIMO) (Appendix Q)[c]
Are the distributions of JAR scores similar between products? • Cochran-Mantel-Haenszel (CMH) (Appendix G)[b] • Stuart Maxwell (Appendix G)[b] • McNemar (Appendix G)[b] • Chi-square (Appendix G)[a] • Proportional Odds/Hazards Model (POM/PHM) (Appendix H)[c]	Did the people who don't find the product JAR rate it lower? • Chi Square (Appendix R)[a]
Do the JAR score means differ between products? • T-Tests (Appendix I)[a] • Analysis of Variance (ANOVA) (Appendix J)[a]	What are the spatial relationships between JAR and other attributes? • Biplots (Appendix S)[b] • Correspondence Analysis (Appendix S)[b] • Principle Components Analysis (Appendix S)[b]
How do the JAR scale distributions compare to those of the theoretical ideal distribution? • Thurstonian Ideal Point Modeling (Appendix K)[c]	What is the linear relationship between the JAR rating and liking? • Correlation (Appendix T)[a]
	Can the relationship between JAR and other data be modeled? • Linear Regression (Appendix U)[b] • Multivariate Adaptive Regression Splines (MARS) (Appendix V)[c] • Partial Least Squares Dummy (PLS) (Appendix W)[b]

[a]Basic statistical skills needed, most statistics packages include
[b]Advanced statistical skills or special software needed
[c]Advanced statistical skills and specialized software needed

study then provides four methods of significance testing of these coefficients. Three of the four methods use a *t*-test on the coefficient, calculating the coefficient standard error either directly from the model or using a jackknife or bootstrap procedure. The method based directly on the model is termed "parametric." The other two methods are termed "semi-parametric" and derive from leave-one-out cross-validation (jackknife) or bootstrap resampling of the data. The fourth method presented is the percentile bootstrap, which is nonparametric and produces a confidence interval that is used to determine significance.

The second case study, Appendix O, provides a single method of significance testing. This method is similar to the semi-parametric bootstrapping method in the first case study; however, it uses bootstrap resampling directly on the mean drops and not on the regression coefficients. Both the jackknife and bootstrap require some programming skill to implement, but the results are straightforward to use. All the presented methods of estimating the variance are well grounded in modern statistical theory and should be of interest to researchers interested in adding significance tests to their penalty analyses.

- Opportunity analysis (see Appendix P).
- Product improvement analysis (PRIMO) (see Appendix O).
- Chi square (see Appendix R).
- Factor analysis and biplots (see Appendix S).

Methods of Analysis Involving Correlation and Regression

The following case studies use correlation or regression analysis to relate the consumers' JAR ratings to overall liking. Although all of the case studies (except the case study that only covers correlation) share regression analysis as their data analysis technique, the approaches are very different from each other. They differ in the assumptions they make concerning the statistical properties of the JAR data and, more basically, in the questions they answer concerning the relationship of the JAR ratings to overall liking. The methods cannot be used interchangeably. The objective of each analysis shall be considered to select an approach that meets the needs of the researcher.

The first case study uses correlation analysis to relate the JAR data to overall liking. Correlation is a widely used

Sample ID n=119	JUST ABOUT RIGHT (JAR) ATTRIBUTES	JAR ANCHORS (5 point scale)	LIKING ATTRIBUTES	LIKING ANCHORS (9 point hedonic scale)
170	Size	Much too Small Slightly Too Small Just About Right Slightly Too Large Much too Large	Liking of Size	Dislike Extremely
896	Color	Much too Light Slightly Too Light Just About Right Slightly Too Dark Much too Dark	Liking of Color	Dislike Very Much
914	Flavor	Much too Weak Slightly Too Weak Just About Right Slightly Too Strong Much too Strong	Liking of Flavor	Dislike Moderately
	Thickness	Much too Thin Slightly Too Thin Just About Right Slightly Too Thick Much too Thick	Liking of Texture	Dislike Slightly
	Stickiness	Not Nearly Sticky Enough Slightly Not Sticky Enough Just About Right Slightly Too Sticky Much too Sticky	Overall Liking	Neither Like nor Dislike
				Like Slightly
				Like Moderately
				Like Very Much
				Like Extremely

Fig. 4—Date Set Attributes and Scale Anchors

and familiar statistical technique. It is, however, limited in the depth of information it can provide concerning the relationship of JAR data to overall liking, and it is among the most restrictive in terms of the assumptions it makes about the nature of JAR scale data. Correlation analysis assumes that JAR ratings are interval scale data arising from, at a minimum, a unimodal, symmetric probability distribution. Further, correlation assumes that the relationship between the JAR ratings and overall liking can be adequately summarized using a straight line. All of these assumptions are suspect. However, the wide availability of software that can perform correlation analysis and its ease of use and interpretation make it a seemingly desirable technique. When using correlation analysis, each JAR scale is analyzed separately. Correlation analysis does not reveal how important it is to be JAR. However, correlation analysis reveals if it is worse to be above JAR than it is to be below JAR or vice versa. If the correlation is positive then it is worse to be "Not Enough". If the correlation is negative it is worse to be "Too Much". If that is the only goal of the analysis, correlation analysis may be a solution. However, there exist other widely available, easy-to-use techniques that can reveal more about the nature of the JAR/liking relationship.

The second case study uses standard regression analysis to relate the JAR data to overall liking. Like correlation analysis, regression analysis is a familiar and widely available technique. Regression makes the same assumptions concerning the nature of JAR data as correlation. However, regression has several advantages over correlation. It provides predicted liking ratings based on the JAR scale ratings and all of the JAR scales can be analyzed simultaneously. More importantly, the simple linear regression model can be extended to fit curvilinear relationships between the JAR ratings and overall liking. These curvilinear relationships come closer to the expectation that overall liking should be higher at the middle of a JAR scale and they are capable of revealing if it is better to be on one side of JAR than the other.

Other, more sophisticated, regression techniques avoid the assumptions that JAR scales produce data that are interval scale, unimodal, and symmetric. For example, the third

case study uses multivariate adaptive regression splines (MARS) analysis to relate JAR ratings to overall liking. MARS selects the JAR variables that are significantly related to overall liking and determines the cost associated with being above or below the JAR level for specific attributes. Unlike penalty analysis that uses a collapsed three-point JAR scale, MARS uses the information from all of the JAR scale categories. Beyond determining if it is better to be on one side of JAR than the other, MARS estimates how much overall liking is decreased by not being JAR. One drawback to MARS is that the software required to perform the analysis is not widely available.

Another limitation of MARS (and all of the other techniques discussed thus far) is that they assume the JAR ratings are independent of each other. This is seldom the case. In almost all product categories, many sensory attributes rise and fall either together or opposite each other and, therefore, are intercorrelated. The fourth case study overcomes this limitation. Partial least squares (PLS) regression with dummy variables possesses all of the advantages of MARS but does not require that the JAR attributes be independent of each other. Although the output of a PLS regression can be difficult to interpret, the approach provides as much, if not more, information about the relationship between the JAR ratings and overall liking while making the fewest assumptions about the nature of the JAR data.

A previously mentioned regression case study presents two related techniques: proportional odds model (POM) and proportional hazards model (PHM). Although both are a type of regression analysis, they deliver results more similar to ANOVA than standard regression. Rather than focusing on predicting overall liking, POM and PHM provide overall tests for differences among the test samples and pair-wise comparisons to determine which samples are significantly different from each other. Both approaches take into account the ordinal nature of JAR scale data, but neither gives any special treatment to the middle "JAR" category on the scale. If statistical comparisons of the test products are the primary objective of the analysis, POM and PHM could be considered. While these methods are regression based, their objective of differentiating products' JAR distributions places them in Table 1 under "Questions Involving Only JAR Data," specifically under the question, "Are the distributions of JAR scores similar between products?"

- Correlation (see Appendix T)
- Regression (see Appendix U)
- MARS (see Appendix V)
- Partial least squares dummy (PLS) (see Appendix W).

Data Analysis for Methods Alternative to JAR Scales

- Collecting intensity and hedonic information separately (see Appendix X).
- Designed experiment (See Appendix Y).
- Ideal point scaling (see Appendix Z).

Summary and Conclusions

This manual has provided an in-depth look at JAR scales, including their application, construction, and analysis. After an introductory section, a brief history of the origin and evolution of JAR scale usage was presented. This was followed by a practical discussion covering JAR scale construction including number of scale points, identification and placement of scale anchors, attribute selection, and location on the ballot. A section on appropriate and inappropriate uses of JAR scales followed. An extensive review of benefits and risks was then presented, including risks that are common to all scales as well as those risks that are unique to JAR scale usage. Alternatives to JAR scales were included for the researcher that chooses other means of obtaining product diagnostic information. Finally, over 25 methods for analysis of JAR scale data were presented in the form of case studies using raw data from a common dataset. The case studies, all in similar format, described the objective, requirements, computations, output, and interpretation of each method, followed by a section on pros and cons with a final recommendation on usage. These case study analyses ranged from simple graphical representations to complex computations that required advanced statistical knowledge or specialized software or both. Case studies for alternatives to JAR scales were presented as well, although these are based on unique datasets.

The case studies demonstrated that, with proper analysis and interpretation, JAR scales provide actionable guidance for product development.

References

[1] Ennis, D. M., *Analytic Approaches to Accounting for Individual Ideal Points*, IFPress, Vol. 8 No. 2, 2005, pp. 2–3.

[2] Moskowitz, H. R., Munoz, M. S., and Gacula, M. C., *Viewpoints and Controversies in Sensory Science and Consumer Product Testing*, Food & Nutrition Press, Inc., Trumbull, CT, 2003, pp. 416–430.

[3] Coombs, C. H., *A Theory of Data*, John Wiley & Sons, Inc., New York, 1964.

[4] Kruskal, J. B., "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis," *Psychometrika*, Vol. 29, 1964, pp. 1–27.

[5] Schutz, H. G., "Consumer DataSense and Nonsense," *Food Quality Preference*, Vol. 10, 1999, pp. 245–251.

[6] Riskey, D. R., "Use and Abuses of Category Scales in Sensory Measurement," *J. Sens. Stud.*, Vol. 1, 1986, pp. 217–236.

[7] Likert, R., "A Technique for the Measurement of Attitudes," *Arch. Psychol.*, 1932, p. 140.

[8] Thurstone, L. L., "Attitudes Can Be Measured," *Am. J. Sociol.*, Vol. 33, No. 4, 1928, pp. 529–552.

[9] Osgood, C. E., Suci, G. J., and Tannenbaum, P. H., *The Measurement of Meaning*, University of Illinois Press, Urbana, IL, 1957, pp. 155–159.

[10] Coombs, C. H., "Psychological Scaling Without a Unit of Measurement," *Psychol. Rev.*, Vol. 57, 1950, pp. 145–158.

[11] Jones, L. V., Peryam, D. R., and Thurstone, L. L., "Development of a Scale for Measuring Soldiers' Food Preferences," *Food Res.*, Vol. 20, 1955, pp. 512–520.

[12] Gridgeman, N. T., "A Comparison of Some Taste-Test Methods," *J. Food Sci.*, Vol. 16, 1961, pp. 171–177.

[13] Moskowitz, H. R., "Subjective Ideals and Sensory Optimization in Evaluating Perceptual Dimensions in Food," *J. Appl. Psychol.*, Vol. 56, 1972, p. 60.

[14] McBride, R. L., "Range Bias in Sensory Evaluation," *J. Food Technol.*, Vol. 17, 1982, pp. 405–410.

[15] Frijters, J. E. R., and Rasmussen-Conrad, E. L., *J. Gen. Psy-*

chol., Vol. 107, 1982, pp. 233–247.

[16] Booth, D. A., Thompson, A., and Shahedian, B., "A Robust, Brief Measure of an Individual's Most Preferred Level of Salt in an Ordinary Foodstuff," *Appetite: J. Intake Res.*, Vol. 4, 1983, pp. 301–312.

[17] McBride, R. L., and Booth, D. A., "Using Classical Psychophysics to Determine Ideal Flavor Intensity," *J. Food Technol.*, Vol. 21, 1986, pp. 775–780.

[18] Connor, M. T., and Booth, D. A., "Preferred Sweetness of a Lime Drink and Preference for Sweet Over Non-Sweet Foods," *Appetite*, Vol. 10, 1988, pp. 25–35.

[19] Sheperd, R., Farleigh, C. A., Land, D. G., and Franklin, J. G., "Validity of Relative-to-Ideal Rating Procedure Compared with Hedonic Rating," in *Progress in Flavor Research, 4th Weurman Flavor Research Symposium*, Elsevier Science Publishers, Amsterdam, 1984, pp. 103–110.

[20] Vickers, Z., Holton, E., and Wang, J., "Effect of Ideal-Relative Sweetness on Yogurt Consumption," *Food Quality Preference*, Vol. 12, No. 8, 2001, pp. 521–526.

[21] Epler, S., Chambers, E., IV, and Kemp, K., "Just About Right Scales are Not Useful for Measuring Sweetness in Lemonade," *J. Sens. Stud.*, Vol. 13, 1998, pp. 191–198.

[22] Earthy, P. J., MacFie, J. H., and Duncan, H., "Effect of Question Order on Sensory Perception and Preference in Central Location Trials," *J. Sens. Stud.*, Vol. 12, 1997, pp. 215–237.

[23] Moskowitz, H. R., Munoz, M. S., and Gacula, M. C., *Viewpoints and Controversies in Sensory Science and Consumer Product Testing*, Food & Nutrition Press, Inc., Trumbull, CT, 2003.

[24] Lawless, H. T., and Heyman, H., *Sensory Evaluation of Food—Principles and Practices*, Chapman and Hall, New York, 1998.

[25] Bendig, A. W., and Hughes, J. B., II, "Effect of Amount of Verbal Anchoring and Number of Rating-Scale Categories Upon Transmitted Information," *J. Exp. Psychol.*, Vol. 40, No. 2, 1953, pp. 87–90.

[26] Kim, W., Ennis, D., and O'Mahony, M., "A New Approach to Category Scales of Intensity II: Use of d′ Values," *J. Sens. Stud.*, Vol. 13, 1998, pp. 251–267.

[27] Pokorny, J., and Davidek, J., "Application of Hedonic Sensory Profiles for the Characterization of Food Quality," *Die Nahrung*, Vol. 8, 1986, pp. 757–763.

[28] Stewart, D. W., Shamdasani, P. N., and Rook, D. W., *Focus Groups: Theory and Practice*, 2nd ed., Thousand Oaks California: Sage Publications, 2007.

[29] Baxter, I. A., and Jack-Schroder, M. J. A., "The Use of Repertory Grid Method to Elicit Perceptual Data from Primary School Children," *Food Quality Preference*, Vol. 9, 1999, pp. 73–80.

[30] Tang, C., and Heymann, H., "Multidimensional Sorting, Similarity Scaling and Free-Choice Profiling of Grape Jellies," *J. Sens. Stud.*, Vol. 17, No. 6, 2002, pp. 493–509.

[31] Gacula, M. C., Jr., Mohan, P., Fuller, J., Pollack, L., and Moskowitz, H.R., "Questionnaire practice: What happens when the JAR scale is placed between two "overall" acceptance scales?"*J. Sens. Stud.*, Vol. 23, 2008, pp. 136–147.

[32] Lovely, C., and Meullenet, J. F., "A Comparison of Statistical Approaches for the Optimization of Strawberry Yogurt Formulation," *7th Pangborn Sensory Science Symposium*, 2007.

[33] Parducci, A., and Perrett, L. F., "Category Rating Scales: Effects of Relative Spacing and Frequency of Stimulus Values," *J. Exp. Psychol. Monogr.*, Vol. 89, 1971, pp. 427–452.

[34] Poulton, E. C., "Models for Biases in Judging Sensory Magnitude," *Psychol. Bull.*, Vol. 86, 1979, pp. 777–803.

[35] Riskey, D. R., Parducci, A., and Beauchamp, G. K., "Effects of Context in Judgments of Sweetness and Pleasantness," *Percept. Psychophys.*, Vol. 26, 1979, pp. 171–176.

[36] Johnson, J., and Vickers, Z., "Avoiding the Centering Bias or Range Effect when Determining an Optimum Level of Sweetness in Lemonade," *J. Sens. Stud.*, Vol. 2, 1987, pp. 283–292.

[37] Moskowitz, H. R., *Food Concepts and Products: Just In Time Development*, Food & Nutrition Press, Inc., Trumbull, CT, 1994.

[38] Moskowitz, H. R., "Just About Right (JAR) Directionality and the Wandering Sensory Unit in Data Analysis Workshop: Getting the Most Out of Just-About-Right Data," *Food Quality Preference*, Vol. 15, 2004, pp. 891–899.

[39] Popper, R., Rosenstock, W., Schraidt, M., and Kroll, B. J., "The Effect of Attribute Questions on Overall Liking Ratings," *Food Quality Preference*, Vol. 15, 2004, pp. 853–858.

[40] Popper, R., Schraidt, M., and Kroll, B. J., "When Do Attribute Ratings Affect Overall Liking Ratings," presented at the *6th Pangborn Sensory Sciences Symposium*, Harrogate International Center, York, 7–11 Aug. 2005.

[41] van Trip, H., Punter, P., Mickartz, F., and Kruithof, L., "The Quest for the Ideal Product," *J. Food Quality Preference*, 2007.

[42] Moskowitz, H. R., "Learning from the Competition Through Category Appraisal: One Practitioner's Keys to Faster and More Efficient Product Development," *Food Service Technol.*, Vol. 1, No. 2, 2001, pp. 103–118.

[43] Gacula, M. C., Jr., Singh, J., Bi, J., and Altan, S., *Statistical Methods in Food and Consumer Research*, 2nd edition, Academic, San Diego, CA, 2009.

Appendix A: Graphical Data Display

Colleen Conley[1]

Introduction and Objectives

The graphical data display method provides a visual comparison of JAR data across products and attributes, or both. The objective of this method is to provide a method to visually assess the distribution of JAR data. This is a descriptive/summarization method, not an inferential one.

Requirements

This method requires the frequency distributions for each attribute for each product, and an application with bar graph capabilities.

"How to"

Summarize the distribution of the data in a frequency table for each product by attribute combination. Graph these summaries using either bar charts or cumulative bar charts of the frequencies. Both can be grouped either by scale category or by product as shown in the two examples below. The bar charts can also be displayed in a product X attribute grid.

Case Study Data Examples

The examples below use the flavor attribute for products 170, 896, and 914. Figures 1(a) and 1(b) display the frequency distribution grouped by scale category and by product, respectively. Figure 2 displays a cumulative bar chart.

Results and Conclusions

Example 1(a) Graph of Frequency Distributions (grouped by JAR scale category) comparing response patterns for three products.

From the graphics above 82 % (84/102) of the assessors scored flavor of Product 896 "About Right," whereas only 62 % (63/102) scored flavor of Product "About Right" and only 44 % (45/102) scored Product 914 "About Right."

Alternatively, bar graphs can be used as illustrated below, showing actual number of responses given in each category for all three products on the same graphic. It is easy to see in the graphic below that product 896 was considered "About Right" in flavor by the most assessors ($N = 84$), while product 914 was thought to have "Not Enough" flavor by 48 assessors and "About Right" by 45 assessors.

Pros and Cons

This method provides a succinct visual summary, and is intended for use as a summary method comparing multiple products/attributes. It is not a formal technique for testing hypotheses.

Recommendation

This method is recommended for all JAR scale data as an initial evaluation tool.

[1] *Director, Sensory Science, Flavor Technology R&D, Solae Company, LLC, P.O. Box 88940, St. Louis, MO 63188.*

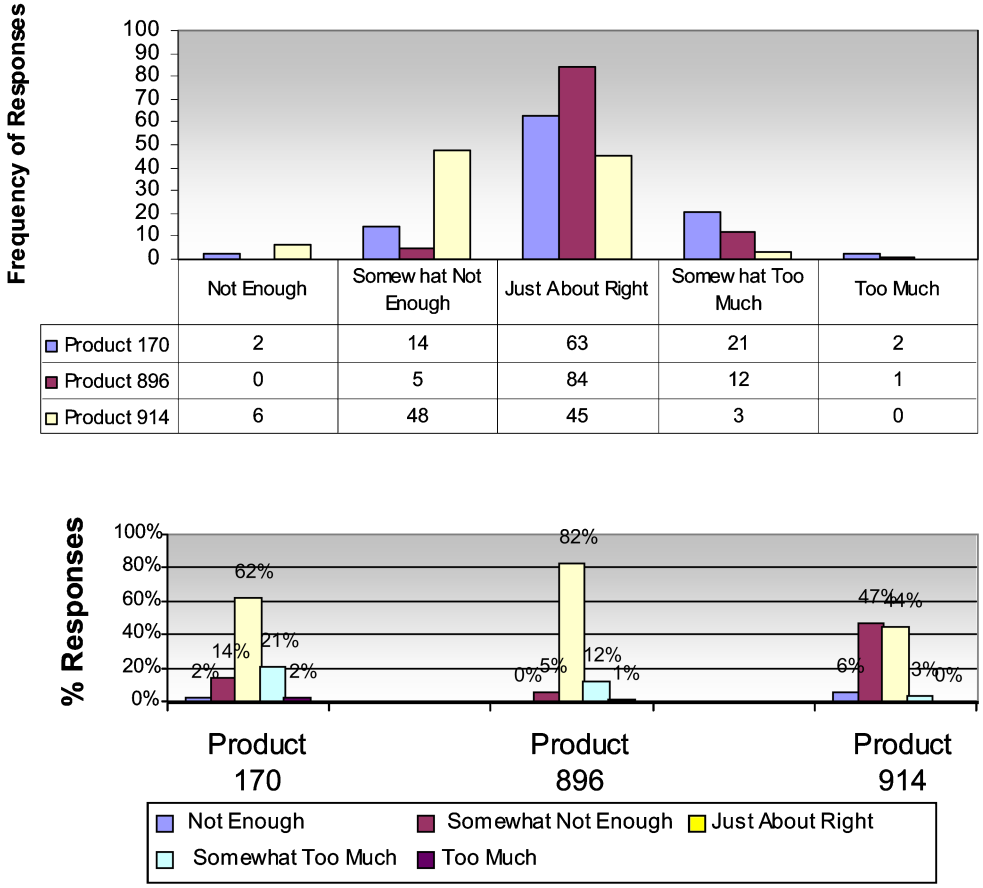

Fig. 1—(a) Flavor JAR frequencies. (b) Flavor JAR percent (%) responses.

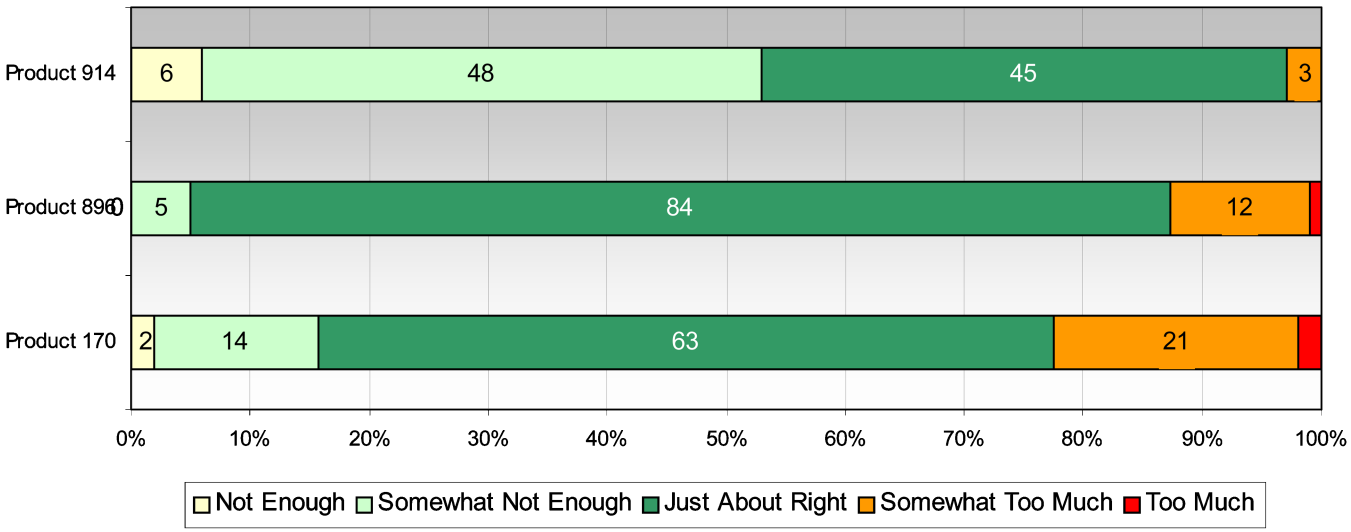

Fig. 2—Flavor JAR.

Appendix B: Graphical Scaling

Anne Goldman[1] and Jagoda Mazur[1]

Introduction and Objective

Graphical scaling is a technique for presenting the results for a JAR scale analysis that graphically illustrates any imbalance around the "Just Right" scale point.

Requirements

The percentage of respondents on both the "Not Enough" and "Too Much" sides of "Just Right" are summed up, for each product and attribute.

"How to"

Subtract the proportion of responses on the "Not Enough" side of the scale from the proportion on the "Too Much" side of the scale. The difference (the Net Effect) indicates the magnitude and direction of differences among test samples, and can be graphed for illustration of the product differences.

Case Study Data Example

This example uses the Size, Color, Flavor, Texture, and Stickiness ratings for products 170, 896, and 914. The data have been collapsed to a three-category scale, as illustrated in Table 1, and the differences plotted in the right hand column of Graph 1.1 This discussion assumes that the *a priori* criteria included a requirement for a "Just Right" of ⩾70 % and a Net Effect (Difference) ⩽20 %.

TABLE 1—Distribution of responses (% of respondents) for the three products.

	%	Sample 170	Sample 896	Sample 914
Size	"Too Small"	24	22	28
	"Just Right"	36	39	31
	"Too Large"	40	39	41
Color	"Not Enough"	11	1	21
	"Just Right"	83	96	79
	"Too Much"	6	3	0
Flavor	"Not Enough"	16	5	53
	"Just Right"	62	82	44
	"Too Much"	23	13	3
Texture	"Too Thin"	13	9	4
	"Just Right"	82	88	75
	"Too Thick"	5	3	21
Stickiness	"Not Enough"	4	1	11
	"Just Right"	79	85	80
	"Too Much"	17	14	9

Results and Conclusions

The ratings show that all three products scored below the 70 % "Just Right" criterion for size, and products 170 and 914 did not meet that criteria for flavor. The Net Effects plot clearly shows that sample 914 lacks flavor. Although Product 914 met the 70 % hurdle for color, the Net Effects suggest that there may be room for improvement by making the product a bit darker.

The actual ratings also reveal polarization of responses for Size for all three products and may suggest existence of more than one consumer segment in the sample. Due to polarization, the Net Effects do not provide clear direction for product changes for Size, as they are located below the 20 % Net Effect criterion in this example. Sample 170 scored slightly below the 70 % criterion for Flavor; however, the "Not Just Right" responses are similar for each end of the scale, again not giving obvious guidance for improvement.

Pros and Cons

The benefits of using this method include ease of calculation, visual presentation, and the simplicity of examination and interpretation of one summary statistic. In cases where the data on each side of "Just Right" are aggregated, the Net Effects may not be representative of the magnitude of differences among products, because information on the degree of "Not Enough" or "Too Much" are lost. Other disadvantages of this method are that if the JAR data are bimodally distributed owing to consumer segments (which may have different expectations concerning product intensities), the results may be misleading in that Net Effects may be artificially low. Another caveat revolves around the assumption that if a product is reformulated to address the skew on one side of "Just Right" that this would not cause a skew on the other side of "Just Right." In color for product 914, for example, if the color were reformulated to be darker, would some of the respondents currently rating the product "Just Right" now rate the color too dark?

Recommendation

The graphical scaling for JAR scale data is recommended as a presentation component of a JAR analysis, but is not recommended as a stand-alone method of analysis.

[1] *Applied Consumer and Clinical Evaluations, 2575 B. Dunwin Dr., Mississauga, ON L5L3N9.*

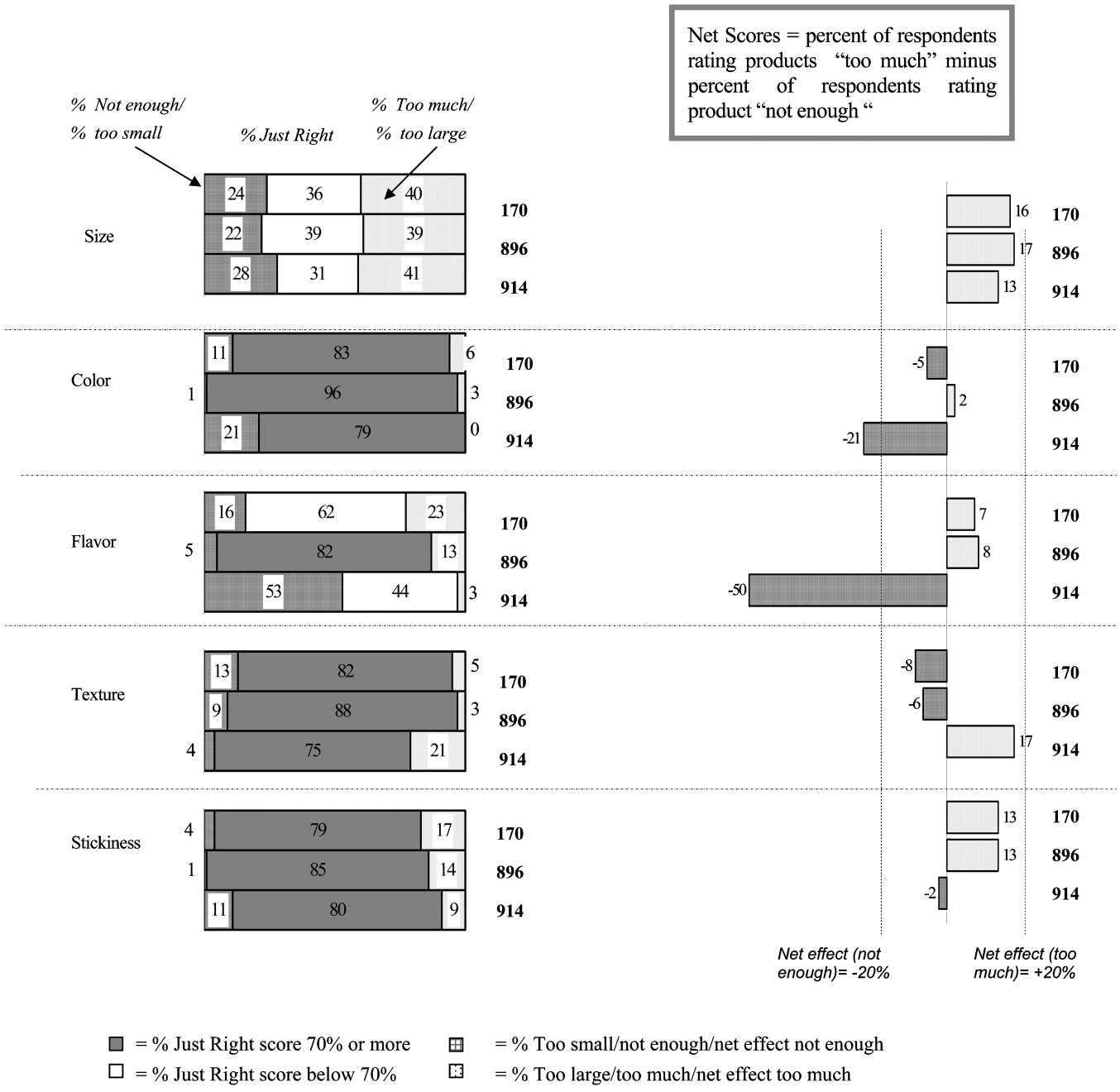

Fig. 1—Actual "Just Right" ratings and net "Just Right" scores.

Appendix C: Percent Difference from Norm and Percent Difference from Just Right

Gloria A. Gaskin[1] and Joni L. Keith[1]

Introduction and Objective

The Percent Difference from Norm analysis determines if the % JAR responses meet or exceed an established norm. If they do not, it also determines the direction of the imbalances of the JAR scale responses. The Percent Difference from "Just Right" analysis does not rely on an established norm, but compares the non-JAR responses to those in the "Just About Right" category. If a significant difference is found, the non-JAR responses are compared to each other. The analysis is conducted on one product-attribute combination at a time, but there is no limit as to how many products or attributes may be analyzed.

Requirement

These analyses require the base size (n) and marginal frequency distribution for each JAR scale attribute and product to be analyzed.

"How to"

There are two approaches to utilizing this technique depending upon whether a normative value for the desired "Percent Just Right" has been established.

Analysis A: Percent Difference from Norm used when a normative value for percent just right has been established

1. Determine the "norm" JAR% responses to be used in the analysis, e.g., 70 %;
2. Is the actual JAR% equal to or greater than the established norm?
3. a. If yes, no further analysis is required; if the JAR% is less than the established norm, continue with the analysis.
   b. Sum the number of responses at each end of the JAR scale (e.g., 1 + 2 and 4 + 5, when using a centered 5-point category scale) and perform a significance test to determine if the number of response at the two ends are significantly different using a binomial or chi-square test against an expected value of 50 %. The total sample size is the number of observations that are not "Just About Right." The confidence level of the test will commonly be a value such as 90 % or 95 %. The critical value can be determined using a binomial table, or a statistical function. Excel offers the CRITBINOM function.
   If a difference is found between the extremes on the "Just About Right" scale, the product should be changed towards the end which has the fewer responses.

Example from Case Study Data

The following tables (Tables 1–3) demonstrate the method for the Flavor, Color, and Size attributes from products 170, 896, and 914 from the case study data. In each case, the norm is assumed to be 70 % "Just About Right" or better. The columns in each table include

1. the product code,
2. the observed JAR value,
3. does the observed JAR meet the criterion,
4. the number of ratings below the "Just About Right" value,
5. the number of ratings above the "Just About Right" value,
6. the sum of columns 5 and 6,
7. the larger of columns 5 and 6,
8. the binomial critical value (the 95 % confidence level in this example),
9. the result of the comparison of columns 7 and 8,
10. the conclusion from that comparison.

Results and Conclusions from the Percent Difference from Norm analysis

Results of the case study using the Percent Difference from Norm analysis indicate that Product 914 does not have

TABLE 1—Attribute: Flavor, option A.

1	2	3	4	5	6	7	8	9	10
Product Code	JAR (n, %)	Is JAR >= 70 %?	1 and 2	4 and 5	Sum (or n)	Max[a]	0.05 Critical Value	Are 1 and 2 different than 4 and 5 at p=0.05?	Conclusions
170	63	No	16	23	39	23	27	No	
Percent	61.76		15.69	22.55					Product 914 does not have enough flavor; increase the flavor of Product 914
896	84	Yes	5	13	18	13	14	N/A, JAR >= 70 %	
Percent	82.35		4.90	12.75					
914	45	No	54	3	57	54	37	Yes	
Percent	44.12		52.94	2.94					

[a]Larger of 1+2 or 4+5

[1] *Bush Brothers and Company, 1016 E. Weisgarber Road, Knoxville, TN 37909.*

TABLE 2—Attribute: Flavor, option A.

1	2	3	4	5	6	7	8	9	10
Product Code	JAR (*n*, %)	Is JAR>= 70 %?	1 and 2	4 and 5	Sum (or *n*)	Max[a]	0.05 Critical Value	Are 1 and 2 different than 4 and 5 at $p=0.05$?	Conclusions
170	85	Yes	11	6	17	11	14	N/A, JAR>= 70 %	All products considered JAR; no action necessary
Percent	83.33		10.78	5.88					
896	98	Yes	1	3	4	3	N/A	N/A, JAR>= 70 %	
Percent	96.08		0.98	2.94					
914	81	Yes	21	0	21	21	16	N/A, JAR>= 70 %	
Percent	79.41		20.59	0.00					

[a]Larger of 1+2 or 4+5

enough flavor ($p=0.05$, $n=57$ [54 (not enough flavor) versus 3 (too much flavor)]. For the attribute Color, all of the products were considered just about right in color based on a JAR% of 70% or more. For the attribute Size, Product 170 was found to be much too large ($p=0.05$, $n=65$ [24("Not Large Enough") versus. 41 ("Much Too Large")]. Product 896 also was found to be much too large in size ($p=0.05$, $n=62$ [22("Not Large Enough") versus 40 ("Much Too Large style")].

Analysis B: Percent Difference from "Just Right" used when a normative percent "Just Right" has not been established

This analysis determines if the non-JAR responses on the "Just About Right" scale (1, 2, 4, 5 on a centered 5-point category scale) are significantly different from the number of responses in the JAR category, for a given attribute and product,.

1 Determine n, the base size, which is the total of all responses.
2. Total the number of responses in the non-JAR categories of the "Just About Right" scale (1 + 2 + 4 + 5).

Compare the number of responses in the non-JAR categories to the number expected in a binomial distribution of the same total size having a binomial proportion of 0.5. Use a table of the binomial distribution to look up the critical value for the alpha = 0.05 level when n is equal to the base size calculated in step1, and p, the binomial proportion, is 0.5. The critical value may also be determined by using the following function in Excel: CRITBINOM(trials, probability_s, alpha) where trials is equal to n, as above, probability _s is 0.5 and alpha is 0.05.

If a difference is found between the total of the extreme categories and the number of JAR responses and if the number of JAR responses is less than or equal to the total of all other responses, proceed to Analysis A, Step 3b, to assess if the extreme categories are different from one another and if so, in which direction.

Example from Case Study Data

The following tables (Tables 4–6) demonstrate the method for the Flavor, Color, and Size attributes from products 170, 896, and 914 from the case study data. The columns in each table include

1. the product code,
2. the number of non-JAR ratings,
3. the observed JAR value,
4. the base size,
5. the binomial critical value (the 95% confidence level in this example),
6. the result of the comparison of columns 2 and 3,
7. the logical conclusion from that comparison,
8. a textual summary of the conclusion.

Results and Conclusions from Option B

Results of the case study using Analysis B, indicate that Product 914 does not have significantly more JAR responses than the combined data from the combined extreme categories ($p=0.05$, $n=102$ [57 ("Not Enough" flavor) versus 45 ("Just About Right" in fla-

TABLE 3—Attribute: Size, option A.

1	2	3	4	5	6	7	8	9	10
Product Code	JAR (*n*, %)	Is JAR>= 70 %?	1 and 2	4 and 5	Sum (or *n*)	Max[a]	0.05 Critical Value	Are 1 and 2 different than 4 and 5 at $p=0.05$?	Conclusions
170	37	No	24	41	65	41	41	Yes	Reduce the size of Products 170 and 914
Percent	36.27		23.53	40.20					
896	40	No	22	40	62	40	39	Yes	
Percent	39.22		21.57	39.22					
914	32	No	28	42	70	42	44	No	
Percent	31.37		27.45	41.18					

[a]Larger of 1+2 or 4+5

TABLE 4—Attribute: Flavor, option B.

1 Product Code	2 1,2,4,5	3 JAR	4 Sum (or *n*)	5 0.05 Critical Value	6 Are 1,2,4,5 different than JAR at $p=0.05$?	7 Is JAR≫than all others at $p=0.05$?	8 Conclusions
170	16+23 =39	63	102	62	Yes	Yes	
Percent	38.24	61.76				But not large enough. Go to Option A	Same conclusion as in Option A
896	5+13= 18	84	102	62	Yes	Yes	Same conclusion as in Option A; No adjustments in flavor necessary
Percent	17.65	82.35				No further analysis	
914	54+3= 57	45	102	62	No	No	Same conclusion as in Option A
Percent	55.88	44.12				Go to Option A	

vor)]. Consequently, further analysis was necessary. Proceed to Step 3b in Option A. Utilizing Option A, it is discovered that Product 914 does not have enough flavor ($p=0.05$, $n=57$ [54 ("Not Enough Flavor") versus 3 ("Too Much Flavor")].

For the attribute Color, all of the products were considered just about right in color based on their JAR % which in all cases was significantly larger than the total of the combined extreme categories ($p=0.05$). For the attribute Size, the JAR % was significantly different from the total of the combined extreme categories for all of the products ($p=0.05$). Additionally, the number of JAR responses for all products was *less than the total of all other responses*. This necessitates following with Option A. Hence, the results are consistent with using *Option A only*. Product 170 was found to be much too large ($p=0.05$, $n=65$ [24 ("Not Large Enough") versus 41 ("Much Too Large")]. Product 896 also was found to be much too large in Size ($p=0.05$, $n=62$ [22 ("Not Large Enough") versus 40 ("Much Too Large")].

TABLE 5—Attribute: Color, option B.

1 Product Code	2 1,2,4,5	3 JAR	4 Sum (or *n*)	5 0.05 Critical Value	6 Are 1,2,4,5 different than JAR at $p=0.05$?	7 Is JAR≫than all others at $p=0.05$?	8 Conclusions
170	11+6= 17	85	102	62	Yes	Yes	In all cases, same conclusions as in Option A.
Percent	16.67	83.33				No further analysis	
896	1+3+4	98	102	62	Yes	Yes	
Percent	3.92	96.08				No further analysis	
914	21+0= 21	81	102	62	Yes	Yes	
Percent	20.59	79.41				No further analysis	

TABLE 6—Attribute: Size, option B.

1 Product Code	2 1,2,4,5	3 JAR	4 Sum (or *n*)	5 0.05 Critical Value	6 Are 1,2,4,5 different than JAR at $p=0.05$?	7 Is JAR≫than all others at $p=0.05$?	8 Conclusions
170	21+41 =65	37	102	62	Yes	No	In all cases, same conclusions as in Option A.
Percent	63.73	36.27				Go to Option A	
896	22+40 =62	40	102	62	No	No	
Percent	60.78	39.22				Go to Option A	
914	28+42 =70	32	102	62	Yes	No	
Percent	68.63	31.37				Go to Option A	

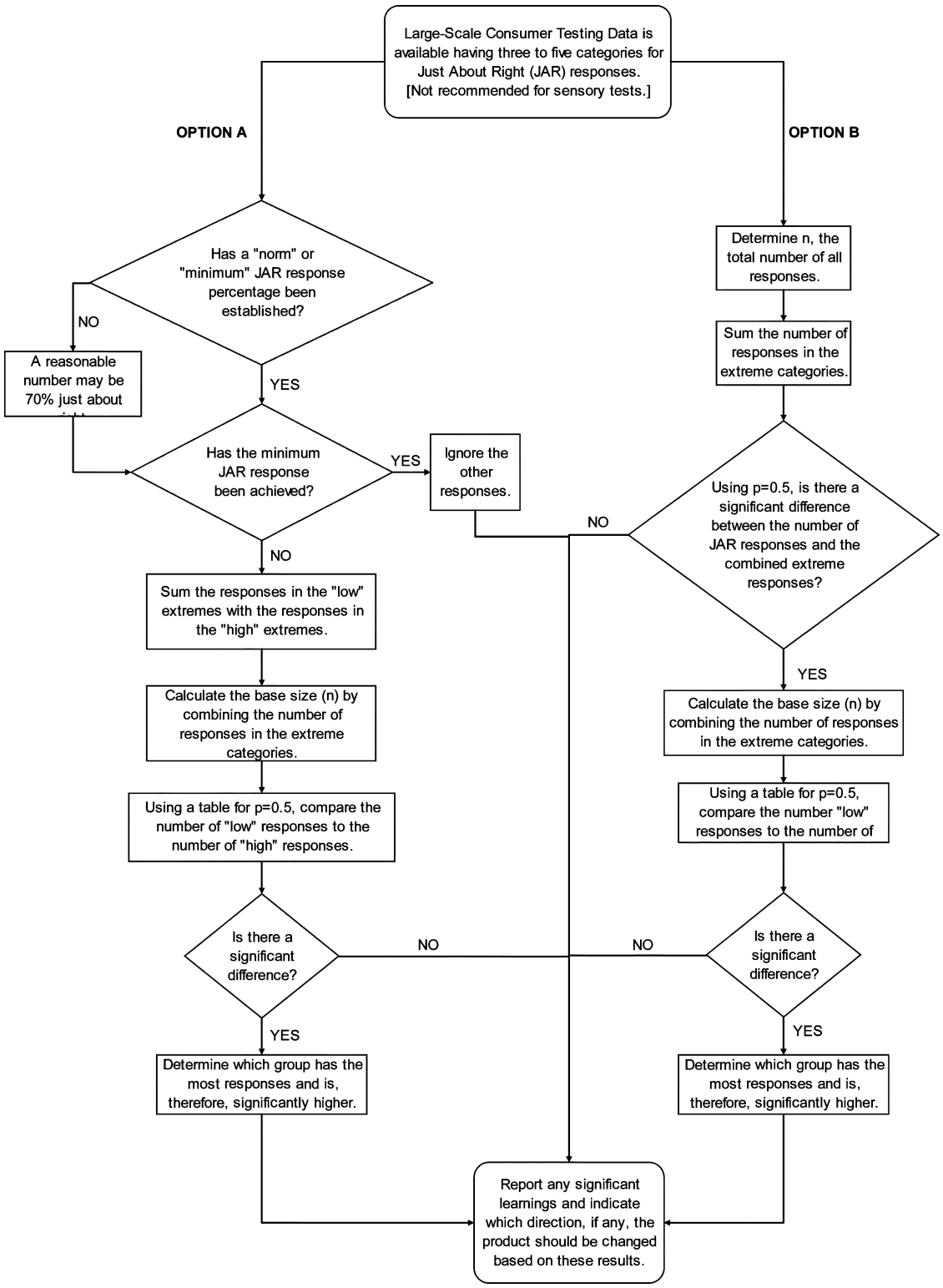

Fig. 1—Just about right scale: Percent difference from norm (flowchart for options A and B).

Pros and Cons

These analyses cover two situations. If a norm is established, Analysis A may be utilized. When a norm JAR % has not established, Analysis B may be used. When the responses are imbalanced, this analysis enables the researcher to determine the direction for optimization. (See Fig. 1.)

This technique does not, by itself, indicate how much of a physical change in an attribute is necessary in order to optimize the product. Neither does this analysis include the effect that other attribute, including Overall Liking, may have

on a particular attribute. These analyses will also suffer when the JAR values are multi-model in the target population.

Recommendation

These methods are recommended when the goal is to understand the JAR ratings for a particular product/attribute combination. When a norm has been established, it will determine if it meets that norm and how the product should be modified to meet that norm. When a norm has not been established, the second method can be used to evaluate the deviation from the "Just About Right" value and suggest how the product can be modified to improve the "Just About Right" score. Neither method is designed to compare products.

Appendix D: The Mean

Lori Rothman[1]

Introduction and Objectives

The objective of the mean analysis is to determine if, on average, the product scores "Just About Right," or whether there is a skew of "Too Much" or "Not Enough" of an attribute. This method of analysis uses only the JAR data. While calculation of the JAR means is completed individually for each product, these means can be compared (see Appendixes I and J). Determining whether the mean is statistically different from the scale midpoint is discussed in Appendix F.

Requirements

To conduct an analysis of the mean, the raw data scores for each respondent and product for the attributes of interest are needed.

"How to"

Each mean score is calculated by summing up the raw data values for each attribute/product combination, and then dividing the sum by number of responses for that attribute/product combination. The mean is then compared to the "Just About Right" value. For example on a symmetric 5-point scale, ranging from 1 to 5, the sample mean would be compared to the "Just About Right" value of 3.

Example From Case Study Data

Figure 1 displays the frequency distribution and mean of the JAR ratings for five attributes on product code 458.

Results and Conclusions

The JAR score means for sample 458 for the attributes Size, Color, Amount of Flavor, Thin/Thick, and Stickiness were calculated and are given below. Interpreting the sample mean for each attribute, a mean of 3.0 for Size implies that the average rating for Size is "Just About Right." A mean of 2.93 for Thin/Thick appears to be close enough to 3.0 to indicate that the thickness is "Just About Right." At a mean of 3.36, the Flavor appears to be "Too Strong," while a mean of 3.13 for Stickiness may indicate that the product is "Too Sticky." The mean of 2.87 for Color probably indicates that the product is "Too Light."

Pros and Cons

The benefits of using the scale means include ease of calculation and the simplicity of examination and interpretation of one summary statistic.

The benefits of this analysis are outweighed by the flawed conclusions that can result from this simplistic approach. Consider the scale mean of 3.0 for size, which implies that the size is "Just Right." Examination of the distribution of scores for this attribute indicates a large degree of bimodality, with 32 % of respondents rating the sample as "Too Large" and 31 % rating the sample as "Too Small." How can the sample be considered "Just Right" for size, when nearly two thirds of respondents rate it otherwise? It is this failure to account for the distribution of responses that make use of the mean unsuitable for JAR scale analysis.

Another limitation is the use of "eyeball" judgments about whether the mean is "close enough" to the scale midpoint when concluding that the attribute is "Just About Right." These judgments often neglect the variability and skewness about the mean. While the mean of 3.36 appears to be solidly in the "Too Strong" area, how confident is the researcher that the mean of 2.87 clearly indicates that the sample is "Too Light."

A third caveat revolves around the finding of a skew in the data, for example, in the case of the Flavor of sample 458 having a mean of "Too Strong." If the researcher makes the Flavor less strong in response to this finding, what will happen to the respondents who rated the product initially as

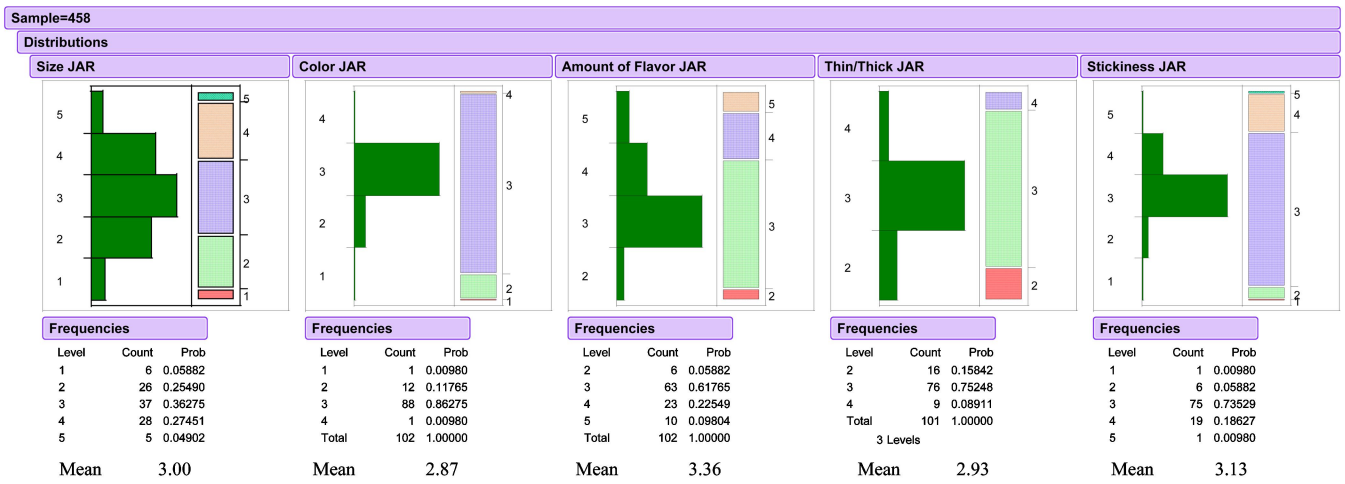

Fig. 1—Frequency distributions and JAR score means for sample 458.

[1] *Kraft Foods, 801 Waukegan Rd., Glenview IL 60025.*

"Just Right?" Will they then rate it "Too Weak" after the reformulation? Thus, examination of the mean alone does not consider what happens after the skew is adjusted.

Yet a fourth limitation is that the finding of a difference from the "Just About Right" value does not, by itself, indicate that the product was not well liked, nor does it indicate the effect that the difference has on Overall Liking. Finally, using only the mean implies an assumption of normality with a known variance, i.e., a bell shaped curve on the responses. If the data are bimodal, this assumption is violated, making this an inappropriate analysis. JAR data can be bimodal, in the presence of consumer segments, which may have different expectations concerning product attribute intensities.

Recommendation

The use of the mean rating for JAR scale data is recommended *only* when combined with additional information such as examination of the data distribution.

Appendix E: Mean Direction and Mean Absolute Deviation

Veronika Jones[1]

Introduction and Objectives

The mean directional and mean absolute deviations are statistics that can be used to summarize and check the balance of JAR data. The JAR data are analyzed separately for each product and each attribute.

The mean directional deviation is the average signed difference of the ratings from the "Just About Right" value. On a 5-point JAR scale, the mean directional ranges from –2 to +2 ("Just Right" =0). Scores that are closer to –2 indicate that respondents thought that attribute was "Too Low." Scores that are closer to +2 indicate that respondents thought that the attribute was "Too High." The mean direction deviation is a simple shift of the mean, covered in Appendix D.

The mean absolute deviation summarizes the spread of the ratings about the "Just About Right" value. On a 5-point scale, the mean absolute deviation ranges from 0, when all judgments are "Just About Right" to+2, when all judgments are at one or the other extreme end of the scale. Unlike percent "Just-Right" scores, absolute deviations can be calculated for each individual and the mean absolute deviation analyzed by any of the standard parametric statistical procedures.

The mean absolute deviation summarizes the average distance to the "Just About Right" value while the mean directional deviation summarizes the average direction the attribute is from the "Just About Right" value (i.e., "Too Low or Too High").

Requirements

To calculate the mean directional and absolute values you must have the distribution of the JAR scores for each product/attribute combination.

"How to"

Mean Directional Deviation

1) Compute the mean score.
2) Subtract the "Just About Right" value for the scale.

Mean Absolute Deviation

1) Subtract the "Just About Right" value from each score.
2) Compute the absolute value of each difference
3) Average the absolute values over assessors.

Example from Case Study Data

Table 1 provides a detailed example of the computations for Flavor ratings for Code 170 by subjects 49–52. Column 4 of that table displays the ratings, Column 5 subtracts 3, the "Just About Right" value for that scale, and Column 6 displays the absolute value of that difference. Table 2 summarizes the calculations for the Size, Color, Flavor, Thin/Thick, and Stickiness attributes for products 170, 896, and 914.

Results and Conclusions

From the mean scores shown in Table 2, we can draw the following conclusions.

- **Size:** There was not much difference between the samples and they were close to "Just Right," though they may all be slightly too large.
- **Color:** Sample 896 was closer to "Just Right" than the other samples. Sample 914 was too light.
- **Flavor:** Sample 896 was closer to just right than the other samples; sample 914 was the farthest from "Just Right." Sample 914 did not have enough flavor.
- **Thin/Thick:** Sample 896 was the closest to Just Right; sample 914 was the farthest from "Just Right." Sample 914 was too thick and the other two samples were slightly too thin.
- **Stickiness:** There was not much difference between the samples; sample 896 was slightly closer to just right than the other two samples. Sample 914 was slightly not sticky enough and samples 170 and 896 were slightly too sticky.
- **Product Comparisons**
  - Sample 896 was the closest to just right for all of the attributes compared to the other two samples.
  - Sample 914 was the farthest from just right for most attributes. It was too light, it did not have enough flavor, it was too thick and it was not sticky enough.
  - Sample 170 was nearly "Just Right" but it was not as satisfactory to respondents as sample was 896.

Pros and Cons

This analysis provides a measure of the spread around the "Just About Right" value and the direction of the average de-

TABLE 1—Example of directional deviation and absolute deviation scores for respondents 49-52 for Amount of Flavor JAR.

Resp. #	Serve	Code #	Amount of Flavor	Directional dev.	Absolute dev.
49	4	170	4	1	1
50	4	170	2	–1	1
51	4	170	3	0	0
52	4	170	5	2	2

[1] *Fonterra Reserch Center, Private Bag 11 029, Dairy Farm Road, Palmerston North, New Zealand.*

TABLE 2—Mean directional and mean absolute scores for example data set.

	Size		Color		Flavor		Thin/Thick		Stickiness	
Sample	Directional	Absolute	Directional	Absolute	Directional	Absolute	Directional	Absolute	Directional	Absolute
170	0.22	0.75	–0.05	0.17	0.07	0.42	–0.08	0.18	0.14	0.22
896	0.24	0.73	0.02	0.04	0.09	0.19	–0.07	0.13	0.14	0.16
914	0.10	0.82	–0.21	0.21	–0.56	0.62	0.17	0.25	–0.03	0.21

viation from that value. The latter can be used to suggest directions for improvements.

A potential limitation of the deviation statistic is the two-step computation required for the deviation statistic. While that may be trivial in packages such as SAS, SPSS, R, and JMP, it requires extra programming in systems such as Excel. This analysis could be improved by a graphical display.

Recommendation

This analysis is useful in summarizing the average shift and spread from the "Just About Right" value on a JAR scale., However, it does not provide any statistical means to judge the differences among products or the inherent variability in the measures.

Appendix F: Mean versus Scale Mid-Point

Sandra Pitts[1]

Introduction and Objective

This method describes a statistical procedure to compare a single mean to its JAR scale mid-point. It will determine if a product attribute is perceived to be significantly different from the ideal. When an attribute is significantly different from the mid-point, conclusions can also be made about the direction of the difference.

This analysis is designed to compare a single product/attribute combination to its center-point, and is not appropriate for the comparisons across products or attributes. Comparisons among multiple products in the same test are limited to statements as to how they each relate to the JAR scale mid-point or "Ideal" value. Note that the "Ideal" value refers to the JAR scale mid-point, and is not related to a specific gold standard product.

Requirements

The method requires the individual JAR ratings of each assessor. It is assumed that the scale values are coded as consecutive integers (e.g., 1 through 5 for the commonly used symmetrical 5-point JAR scale).

"How to"

First, examine the frequencies of distribution of the raw JAR data for each product. If the data appear approximately unimodal for each product, proceed with the analysis. Otherwise, when the data appear bimodal, consider an alternate analysis.

Subtract the scale mid-point value from the raw data (e.g., 3.0 on a 5-point 1–5 scale). Analyze these data using a one-sample, two-tailed *t*-test at the desired confidence level (e.g., 95 %). The analysis may be performed using Excel, SAS, MiniTab, SPSS, or another general statistical software package with *t*-test capabilities.

Case Study Data Examples

The tables below summarize the mean scores and the -*p* values from a one-sample *t*-test for all of the attributes and products in the Case Study. For each attribute-product combination, the individual ratings were compared to the JAR mid-point of 3.0. Recall that lack of significance does not prove equivalence, only that the data are not sufficient to reject the hypothesis of equivalence.

Results and Conclusions

See Tables 1 and 2. Product 170 could be considered "Just About Right" for Amount of Color, Amount of Flavor and Thickness, and was perceived to be "Too Large" in Size and "Too Sticky" as compared to the "Ideal."

Product 458 could be considered "Just About Right" for Size and Thickness, and was perceived to be "Too Low" in Color, "Too High" in Flavor Intensity, and "Too Sticky" as compared to the "Ideal."

TABLE 1—Mean Scores for five products for five JAR attributes. $N=102$. Mean scores in bold are significantly different from the mid-scale point of 3.0, at $\alpha=0.05$.

Product Code	Size	Color	Amount of Flavor	Thickness Thin/Thick	Stickiness
170	**3.2**	3.0	3.1	2.9	**3.1**
458	3.0	**2.9**	**3.4**	2.9	**3.1**
523	3.2	2.9	**3.2**	**2.9**	**3.1**
896	**3.2**	3.0	**3.1**	2.9	**3.1**
914	3.1	**2.8**	**2.4**	**3.2**	3.0

TABLE 2—*t*-test comparison of JAR mean scores versus scale mid-point for five products for five attributes (*p*-values by attribute).

Product Code	Size	Color	Amount of Flavor	Thickness Thin/Thick	Stickiness
170	0.026	0.227	0.329	0.059	0.004
458	1.000	0.001	0.000	0.163	0.023
523	0.106	0.134	0.001	0.004	0.001
896	0.015	0.320	0.049	0.070	0.001
914	0.347	0.000	0.000	0.001	0.551

Product 523 could be considered "Just About Right" for Size and Amount of Color, and was perceived to be "Too High" in Flavor Intensity, and "Too Thin" and "Too Sticky" as compared to the "Ideal."

Product 896 could be considered "Just About Right" for Amount of Color and Thickness, and was perceived as "Too Large," "Too High" in Flavor Intensity, and "Too Sticky" as compared to the "Ideal."

Product 914 could be considered "Just About Right" for Size and Stickiness, and was considered Too Low in Color, "Too Low" in Flavor Intensity, and "Too Thick" as compared to the "Ideal."

Pros and Cons

This is a simple analysis to perform, and can provide guidance to product development on how directional changes to a product attribute might increase its acceptability. This method is reasonable to use when there is no established norm for the expected percentage of "Just About Right" responses.

This method assumes a normal distribution of responses. Distribution of responses should be examined before performing this procedure (i.e., review the frequencies of distributions either as a table of numerical values or as a histogram); if there is a bimodal distribution, then a simple test of the mean is generally not appropriate (see Appendix D).

Comparison of the means to the scale mid-point pro-

[1] *Compusense Inc., 679 Southgate Drive, Guelph ON, Canada N1 G 4S2.*

vides directional guidance on product changes; additional data are required to draw conclusions about the absolute amount of change for the specified attribute to increase consumer acceptability.

This method does not allow direct comparisons among samples; additional analyses (see Appendix J) would be required in order to compare two or more products.

Recommendation

This method of analysis is recommended when there is only JAR data available from a unimodal population, to allow comparison of individual product attributes to the "Ideal."

Appendix G: Methods for Determining Whether JAR Distributions are Similar Among Products (Chi-Square, Cochran-Mantel-Haenszel (CMH), Stuart-Maxwell, McNemar)

Carl Fritz[1]

General Introduction and Objectives

The following methods can each be used to determine whether JAR score distributions are similar among a set of products:

- Chi-square method
- Cochran-Mantel-Haenszel (CMH) method [1], [2]
- Stuart-Maxwell method [3]
- McNemar method [4]

The chi-square method and CMH method are the most general of the four methods listed above. Both methods can be used for comparing JAR score distributions among any number of products for any number of JAR scale categories. The chi-square method differs from the other three methods with respect to the design of the consumer test for which it can be used. The use of the chi-square method requires independence among the assessors' responses. This limits the use of the chi-square method to situations where different groups of assessors evaluate each product.

The CMH, Stuart-Maxwell, and McNemar methods are all appropriate when the assessors evaluate all of the products. These methods provide an increased level of power over the chi-square method by taking advantage of the positive correlation that typically exists among an individual assessor's ratings on two or more products evaluated during the same consumer test.

The CMH method can be used to test the equality of JAR score distributions for multiple products using JAR scales with mulitple categories. The Stuart-Maxwell method is a special case of the CMH method and is used to compare JAR score distributions of two products for JAR scale with multiple categories. The McNemar method is a special case of the CMH and Stuart-Maxwell methods that is used for two products and exactly two response categories (e.g. "Just Right" and "Not Just Right" or "Too Thick" and "Not Too Thick"). The test statistics for the CMH, Stuart-Maxwell, and McNemar tests are identical for the situation where there are two products and two scale categories.

These methods are appropriate for either complete block designs, where each assessor evaluates a sample of each product or for unblocked designs where each assessor evaluates a sample(s) of a single product. The hypothesis being tested are, for the most part, general. More powerful techniques, such as ordinal logistic regression, may be appropriate for specific hypotheses; see Table 1.

Requirements

All four of the methods require that the distributions of responses by category be available.

Cochran-Mantel-Haenszel (CMH) Method

The Cochran-Mantel-Haenszel (CMH) method can be used for determining whether there are statistically significant differences in JAR score distributions among two or more products when each product has been evaluated by each assessor. It tests the homogeneity of the JAR scale across products, after controlling (blocking) for the differences among the assessors.

Objectives of the Analysis

The CMH method tests either the null hypothesis that there are no differences in the distributions of JAR scores across the products or the null hypothesis that there is no difference in mean JAR scores across the products.

TABLE 1—Summary of methods for comparing JAR distributions among products.

		2 Products		3 or More Products	
Test format	Method	2 JAR Scale Categories	3 or More JAR Scale Caterories	2 JAR Scale Categories	3 or More JAR Scale Caterories
Each assessor evaluates all products	CMR	X	X	X	X
	Stuart-Maxwell	X	X		
	McNemar	X			
Different assessors evaluate each product	Chi-square	X	X	X	X

[1] *Statistical Consultant, 15 Crammer Lane, Hillsborough, NJ 08844.*

Hypothesis 1: General Association

This form of the CMH tests the null hypothesis that the distrubtion of JAR scores is the same across all products after adjusting for differences between raters. It treats the JAR scale as an unordered (nominal) scale. The alternative hypothesis of general association is that at least one of the products differs on at least one of the JAR scale categories. This form of the test should be used when the researcher wants to determine whether the distributions of JAR scores differ among the products without stating the specific pattern of differences. The null hypothesis will be rejected if the distribution of responses for one product is sufficiently different than the distribution of responses for another product regardless of whether the mean responses for the products are different.

Hypothesis 2: Different Mean Responses

This form of the CMH tests the null hypothesis that the mean JAR scores are constant across the products. The alternative hypothesis is that at least one of the products has a different mean score from the rest. This form of the test is used when the researcher wishes to test the equality of the weighted or unweighted means.

Either one or both of the alternative hypotheses could be of interest to the researcher in a particular study. It is appropriate to test both alternative hypotheses in the same study if the researcher is interested in both hypotheses.

The following examples may help the researcher understand the distinction between the two alternative hypotheses. In both examples, the "Not Enough," "Just Right," and "Too Much" categories of a 3-point JAR scale are coded as 1, 2, and 3, respectively. Example 1 illustrates a situation where there are differences among the products in the number of responses in each of the JAR scale categories, but there are no differences in mean responses among the products.

Example 1

Frequency of responses

	"Not Enough" (1)	"Just Right" (2)	"Too Much" (3)	Mean
Product A	10	80	10	2.0
Product B	15	70	15	2.0
Product C	20	60	20	2.0

In Example 1, the null hypothesis of identical frequency distributions would be rejected. There are statistically significant differences among the products in the distribution of responses across the JAR scale categories ($p < 0.0001$), but there are no significant differences among products in the mean responses. For details on how to perform the CMH test, see the case study examples on the following pages.

Example 2 illustrates a situation where there are statistically significant differences in both the mean responses among the products and in the number of responses in each of the scale categories.

Example 2

Frequency of responses

	"Not Enough" (1)	"Just Right" (2)	"Too Much" (3)	Mean
Product A	5	60	35	2.3
Product B	20	60	20	2.0
Product C	35	60	5	1.7

Requirements for the Analysis

To use the CMH method for analyzing data from JAR scales, the individual JAR scores from each assessor for each product must be available. Each product included in the analysis must have been evaluated by each assessor. Additionally, like its continuous counterpart, the Randomized Complete Block design, the validity of this analysis requires that certain additivity or homogeneity requirements are met.

Details of the Analysis

Several commercially available statistical computer programs such as SAS [5] and JMP can be used to perform the analysis for the CMH methods. There are no simple formulas for hand calculation available for computing the CMH statistics, and the use of a computer program is recommended. The mathematical details that are necessary to explain the formulas that are used in the CMH methods can be found in Refs. [1], [2], and [5].

To conduct the CMH tests, assign a numerical code to each category of the JAR scale. If the scale contains more than three categories, the analyses can be performed one of two ways: 1) using all of the original scale categories, or 2) by combining the categories on each side of the midpoint to create a three-category scale (e.g., "Too Little," "Just Right," "Too Much").

When testing the null hypothesis of no general association, the analysis treats the JAR categories as nominal data values, so any numerical or text codes can be used for the categories as long as each category is assigned a different code.

When testing the null hypothesis of common means across the products, the numerical values assigned to the scale levels are used to order the levels. Additionally, for simple means, the values are used to form the means themselves. Often the values are assigned as ordered integers. For a 3-point JAR scale, two common approaches are to assign codes of {1, 2, 3} or {–1, 0, 1} to the "Not Enough," "Just Right," and "Too Much" categories, respectively. For a 5-point JAR scale, a researcher could use either {1,2,3,4,5} or {–2, –1, 0, 1, 2} as codes for the categories. Optionally, other forms of optimal weights (e.g., ridit weights) may be available in the statistical analysis program. Details of these scoring methods are beyond the scope of this document.

The assessors' responses can be summarized in an r by c contingency table where r = number of products (rows) and c = number of scale categories (columns). The body of the contingency table shows the frequency of responses for each scale category for each product (see below).

		$c = 3$ columns		
		"Not Enough" (1)	"Just Right" (2)	"Too Much" (3)
$r = 3$ *rows*	Product A	5	60	35
	Product B	20	60	20
	Product C	35	60	5

The CMH statistic for testing the hypothesis of general association follows a chi-square distribution with degrees of freedom = (products-1) × (columns-1). The CMH statistic for testing the hypothesis of differences in mean responses between the products follows a chi-square distribution with degrees of freedom = products-1.

Case Study Data Examples

For analysis of the case study data, the 5-point JAR scale was collapsed to three categories by combining the two categories on the "Not Enough" side of the midpoint and by combining the two categories on the "Too Much" side of the midpoint.

The following program statements can be used in the SAS software program to perform the CMH methods:

```
proc freq;
     by attribute;
     tables product * category / norow
nocol nopercent;
     tables assessor * product *
category/ cmh noprint;
     run;
```

The first "tables" statement creates a summary table that shows the frequency of responses in each category for each product. The second "tables" statement performs the CMH tests for general association and for differences in mean responses.

Results

Attribute = JAR Size Frequency

Product	1	2	3	Total
170	24	37	41	102
896	22	40	40	102
914	28	32	42	102

		CMH	
Alternative Hypothesis	DF	Value	*p*-value
Row Mean Scores Differ	2	0.174	0.916
General Association	4	1.691	0.792

Attribute = JAR Color Frequency

Product	1	2	3	Total
170	11	85	6	102
896	1	98	3	102
914	21	81	0	102

		CMH	
Alternative Hypothesis	DF	Value	*p*-value
Row Mean Scores Differ	2	24.53	<0.0001
General Association	4	29.94	<0.0001

Attribute = JAR Amt. Frequency

Product	1	2	3	Total
170	16	63	23	102
896	5	84	13	102
914	54	45	3	102

		CMH	
Alternative Hypothesis	DF	Value	*p*-value
Row Mean Scores Differ	2	63.98	<0.0001
General Association	4	77.77	<0.0001

Attribute = JAR Thin/Thick Frequency

Product	1	2	3	Total
170	13	84	5	102
896	9	90	3	102
914	4	77	21	102

		CMH	
Alternative Hypothesis	DF	Value	*p*-value
Row Mean Scores Differ	2	23.73	<0.0001
General Association	4	29.70	<0.0001

Attribute = JAR Stickiness Frequency

Product	1	2	3	Total
170	4	81	17	102
896	1	87	14	102
914	10	82	9	102

		CMH	
Alternative Hypothesis	DF	Value	*p*-value
Row Mean Scores Differ	2	7.51	0.023
General Association	4	11.13	0.025

The CMH analysis can also be done with the JMP software program by requesting a contingency table analysis within the "Fit Y By X" platform. Variables that represent the products and the JAR attributes should be defined as either nominal or ordinal variables in order for the analysis to produce the correct test.

Conclusions from the Analysis

There are no statistically significant differences among the three products in the distributions of the assessors' scores on the JAR scale for the size attribute. For the color, amount of flavor, thickness, and stickiness attributes there are statistically significant differences between at least two of the products in the distributions of JAR scores and in the mean responses. For the four attributes where significant differences

were found, a recommended follow-up analysis would be to repeat the CMH method for subsets of two products at a time to determine which pairs of products have significantly different distributions of scores. This approach is equivalent to using the Stuart-Maxwell method as a follow-up procedure to the CMH method for determining whether the distributions of scores differ between two products.

Benefits of the Analysis

The CMH method allows the researcher to determine whether there are significant differences in JAR score distributions among two or more products. Other approaches such as the McNemar test and the Stuart-Maxwell test are only suitable for testing for differences in distributions between two products. The CMH method allows the researcher to analyze data from JAR scales having more than three categories. The McNemar test requires that the data be combined into two categories. The CMH method provides more power (i.e., a higher probability that a statistically significant difference is found when one of the alternative hypotheses are true) than the chi-square method when the same assessors evaluate all products by taking advantage of the positive correlation that typically occurs when individual assessor's rate two or more products during the same test session.

Caveats

The CMH method is not available in some statistical analysis computer programs. If the CMH method is not available, an alternative approach is to use the Stuart-Maxwell test to analyze two products at a time. Use of the CMH method is limited to complete block test designs where each assessor evaluates all of the products in the test.

Recommendations

The CMH method is appropriate for determining whether there are differences in JAR score distributions among two or more products when the products are all evaluated by the same group of assessors. If each product is evaluated by a different group of assessors, then this method is not appropriate and a chi-square method or more general technique, such as an ordinal regression should be used.

Stuart-Maxwell Method

The Stuart-Maxwell method can be used to compare the distribution of JAR scores from two products when each assessor evaluates each product. For example, the researcher may want to know whether there is a difference between two products in the proportion of scores in the "Too Much" category or in the "Just Right" category. The Stuart-Maxwell method is a special form of the more general CMH method discussed above [b].

Objectives of the Analysis

The null hypothesis of the Stuart-Maxwell method is that the JAR score distributions for two products are identical. The alternative hypothesis is that there is a difference in the distribution of JAR scores between two products. If there are more than two products, the Stuart-Maxwell method can be used as a follow-up test after the Cochran-Mantel-Haenszel (CMH) method has determined that there are differences in the JAR score distributions among the products. In this situation, the Stuart-Maxwell method is used to determine which pairs of products have significantly different JAR score distributions.

Requirements for the Analysis

To use the Stuart-Maxwell method, both products must have been evaluated by the same assessors. The data must first be arranged in a table that lists the number of assessors that gave the same rating on the JAR scale to Product A and Product B and the number of assessors that gave different ratings to Product A and Product B as shown below.

	Rating on Product B			
Rating on Product A	"Too Little"	"Just Right"	"Too Much"	Row Totals
"Too Little"	n_{11}	n_{12}	n_{13}	$n_{1.}$
"Just Right"	n_{21}	n_{22}	n_{23}	$n_{2.}$
"Too Much"	n_{31}	n_{32}	n_{33}	$n_{3.}$
Column Totals	$n_{.1}$	$n_{.2}$	$n_{.3}$	

where:

n_{11} = number of assessors that gave the rating "Too Little" to both products,
n_{12} = number of assessors that gave a rating of "Too Little" to Product A and a rating of "Just Right" to Product B
$n_{1.}$ = number of assessors that gave the rating "Too Little" to Product A
$n_{.1}$ = number of assessors that gave the rating "Too Little" to Product B

Details of the Analysis

First, calculate the difference in the number of ratings in each scale category (e.g., "Too Little," JAR, "Too Much") between the two products as follows:

$$d_1 = n_{1.} - n_{.1} \quad d_2 = n_{2.} - n_{.2} \quad d_3 = n_{3.} - n_{.3}$$

Next, compute the test statistic:

$$\chi^2 = \frac{0.5(n_{12}+n_{21})d_3^2 + 0.5(n_{13}+n_{31})d_2^2 + 0.5(n_{23}+n_{32})d_1^2}{2\{0.25(n_{12}+n_{21})(n_{13}+n_{31}) + 0.25(n_{12}+n_{21})(n_{23}+n_{32}) + 0.25(n_{13}+n_{31})(n_{23}+n_{32})\}}$$

Then, compare χ^2 to a value from the chi-square table with 2 degrees of freedom (df) at the desired significance level.

Note: The above formula for χ^2 is specific to the situation where the JAR scale contains three categories. If the JAR scale contains more than three categories, computation of the test statistic requires inversion of a matrix. The formula for computing the test statistic in this case is given in the reference for Stuart (1955). Since the Stuart-Maxwell method is a special case of the CMH method, an alternative approach is to use a statistical computer program that performs the CMH method.

Case Study Data Example

Attribute = JAR Amt. Flavor

	Rating on Product 896			Row
Rating on Product 170	"Too Little"	"Just Right"	"Too Much"	Totals
"Too Little"	3	11	2	16
"Just Right"	1	60	2	63
"Too Much"	1	13	9	23
Column Totals	5	84	13	

$$d_1 = 16 - 5 = 11 \quad d_2 = 63 - 84 = -21 \quad d_3 = 23 - 13 = 10$$

$$\chi^2 = \frac{(1/2)(11+1)10^2 + (1/2)(2+1)(-21)^2 + (1/2)(2+13)11^2}{2\{(1/4)(11+1)(2+1) + (1/4)(11+1)(2+13) + (1/4)(2+1)(2+13)\}} = 16.62$$

The test statistic 16.62 is greater than the critical value of 13.82 from a chi-square distribution table with 2 degrees of freedom at the 0.001 significance level. This indicates that there is a statistically significant difference between products 170 and 896 in the JAR scale distributions for amount of flavor.

Conclusion from the Analysis

The conclusion that there is a statistically significant difference in JAR score distributions for amount of flavor between Product 170 and Product 896 does not tell the researcher how the JAR score distributions differ. The researcher can often determine how the distributions differ simply by looking at the table of frequencies. In this example, product 896 received more scores than product 170 in the "Just Right" category (84 versus 63). A follow-up analysis that may be of interest to the researcher is to combine the responses in the "Too Little" and "Too Much" categories and use the McNemar method to determine whether there is a difference in the number of "Just Right" and "Not Just Right" responses between the two products. This follow-up analysis will tell the researcher whether or not a significantly higher proportion of assessors gave ratings of "Just Right" to one product than the other.

Benefits of the Analysis

The Stuart-Maxwell method allows the researcher to determine whether there is a significant difference in JAR score distributions between two products in the situation where the JAR scale contains three or more categories and each assessor evaluated both products. Another approach, the McNemar test, requires that the data be combined into two categories. The Stuart-Maxwell method provides more precision than the chi-square method for testing situations where the same assessors evaluate all products. When the JAR scale contains three categories, formulas are available for computing the test statistic without the use of a computer.

Caveats

The Stuart-Maxwell method is not available by name in most common statistical analysis computer programs. However, since the Stuart-Maxwell method is a special case of the CMH method, any software program that performs the CMH method will provide the Stuart-Maxwell method as well. As with the CMH method, this method also requires certain homogeneity assumptions to be valid.

Recommendations

The Stuart-Maxwell method can be used to compare the JAR score distributions of two products when the products are all evaluated by the same group of assessors. If each product is evaluated by a different group of assessors, then this method is not appropriate and a chi-square method or more general technique such as an ordinal regression should be used.

McNemar Method

The McNemar method can be used for determining whether there are differences in JAR score distributions between two products when data from the JAR scale have been combined into two categories. The McNemar method is appropriate when both products have been evaluated by the same assessors.

Objectives of the Analysis

The null hypothesis tested by the McNemar method is that the proportions of JAR scores in the two categories are equal for the two products being compared. The alternate hypothesis is that the proportions for the two products are different. The McNemar method is typically used to determine whether there are differences in the JAR score distributions between two products when ratings on the JAR scale have

been combined into two categories in one of the following ways:

- "Too Little" and "Too Much" ratings combined to create the categories "Just Right" and "Not Just Right"
- "Too Little" and "Just Right" ratings combined to create the categories "Too Little or Just Right" and "Too Much"
- "Just Right" and "Too Much" ratings combined to create the categories "Too Little" and "Just Right or Too Much"

The McNemar test can be used as a follow-up test after a statistically significant outcome from either the Cochran Mantel Haenszel (CMH) method or the Stuart-Maxwell method in order to determine how the distributions of JAR scores differ between two products.

Requirements for the Analysis

To use the McNemar method, both products must have been evaluated by the same assessors. The data must first be arranged in a table that lists the number of assessors that gave the same rating on the JAR scale to Product A and Product B and the number of assessors that gave different ratings to Product A and Product B as shown below. The example below shows the data arrangement when the scale values have been combined to create the two categories "Just Right" and "Not Just Right."

	Rating on Product B	
Rating on Product A	"Just Right"	"Not Just Right"
"Just Right"	n_{11}	n_{12}
"Not Just Right"	n_{21}	n_{22}

where:

n_{11} = number of assessors that gave the rating "Just Right" to both products

n_{12} = number of assessors that gave the rating "Just Right" to Product A and the rating "Not Just Right" to Product B

n_{21} = numberof assessors that gave the rating "Just Right" to Product B and the rating "Not Just Right" to Product A

n_{22} = number of assessors that gave the rating "Not Just Right" to both products

Details of the Analysis

To determine whether the number of responses in the two categories differs significantly between the two products, first calculate the McNemar test statistic:

$$\chi^2 = \frac{(|n_{12} - n_{21}| - 1)^2}{n_{12} + n_{21}}.$$

Then, compare χ^2 to a table of the chi-square distribution with 1 df at the desired significance level.

Note: When n_{12} and/or n_{21} are small (say, $n_{12}+n_{21}<10$), then the McNemar test statistic χ^2 is not well approximated by the chi-square distribution. A two-tailed exact test based on the cumulative binomial distribution is recommended instead (see SAS code below).

The following program statements can be used in the SAS software program to perform the McNemar method:

```
proc freq;
       by attribute;
    tables product * category /
agree;
       exact mcnem;
       run;
```

The "agree" option provides the McNemar test. The "mcnem" keyword in the "exact" statement provides the exact test based on the cumulative binomial distribution.

Since the McNemar method is a special case of the CMH method, the McNemar test can also be done with the JMP software program by requesting a contingency table analysis within the "Fit Y By X" platform. Variables that represent the products and the JAR attributes should be defined as either nominal or ordinal variables in order for the analysis to produce the correct test. In the output window, refer to the results for the CMH test [7].

The McNemar method can also be performed with the SPSS software program by choosing either of the following two menu paths:

- Analyze / Descriptive Statistics/ Crosstabs/click the "Statistics" button and choose "McNemar" or
- Analyze / Nonparametric Tests / 2 Related Samples / select the box labeled "McNemar"

Case Study Data Example

In the following example, the McNemar test is used as a follow-up test to the Stuart-Maxwell method for the attribute "JAR Amt. Flavor" to determine whether there is a difference between Products 170 and 896 in the distribution of JAR ratings when the ratings are combined into the two categories "Just Right" and "Not Just Right." Please refer to the section of this document that describes the Stuart-Maxwell method for details on the Stuart-Maxwell method.

Attribute = JAR Amt. Flavor

	Rating on Product 896	
Rating on Product 170	"Just Right"	"Not Just Right"
"Just Right"	60	3
"Not Just Right"	24	15

$$\chi^2 = \frac{(|3 - 24| - 1)^2}{3 + 24} = 14.81$$

The test statistic 14.81 is greater than the tabled value of 10.83 from the chi-square distribution with 1 df at the 0.001 significance level.

Conclusion from the Analysis

There is a statistically significant difference between Product 170 and Product 896 in the proportion of assessors that rated the products "Just Right" for Amount of Flavor.

Benefits of the Analysis

The McNemar method allows the researcher to determine whether there is a significant difference in JAR score distributions between two products when JAR scale values have been combined into two categories and each assessor has evaluated both products. The McNemar method provides

more precision than the chi-square method for testing situations where the same assessors evaluate all products. The computation of the McNemar test statistic is easily done without a computer.

Caveats

One disadvantage of the McNemar method is that the analysis accommodates only two scale categories per product. As with the CMH method, this method also requires certain homogeneity assumptions to be valid.

Recommendations

The McNemar method is recommended for use when two products are being compared, the JAR scale have been collapsed to two categores and each assessor has evaluated a sample from both products. If each product is evaluated by a different group of assessors, then this method is not appropriate and a chi-square method or more general technique such as an ordinal regression should be used.

Chi-square Method

The chi-square method is appropriate when each assessor evaluates only one product and the researcher wishes to compare the distribution of JAR scores across two or more products.

Objectives of the Analysis

The chi-square method can be used to test the null hypothesis that there are no differences in the distributions of JAR scores among the products. The alternative hypothesis is that at least one product is different from the others on this JAR scale. For example, one product may have a higher proportion of scores in the "Just Right" category than another product.

Requirements for the Analysis

This analysis method requires that the assessors' ratings are independent. This usually implies that a different group of assessors evaluates each product. Note that it may be possible to structure the testing so that assessors' ratings of multiple products behave as if they are independent (for instance, by separating the evaluations of products by a long enough period of time that assessors will not recall their prior evaluations).

To use the chi-square method it is not necessary to have the assessors' individual responses available. It is only necessary to know the total number of responses in each category of the JAR scale for each product as shown below.

	JAR scale category			Row
	"Too Little"	"Just Right"	"Too Much"	Totals
Product A	n_{11}	n_{12}	n_{13}	$n_{1.}$
Product B	n_{21}	n_{22}	n_{23}	$n_{2.}$
Product C	n_{31}	n_{32}	n_{33}	$n_{3.}$
Column Totals	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{..}$

where n_{11} = number of assessors that gave the rating "Too Little" to Product A $n_{1.}$ = number of assessors that rated Product A $n_{.1}$ = sum of the number of ratings of "Too Little" for all products $n_{..}$ = sum of the number of ratings in all JAR categories for all products

Details of the Analysis

Most statistical software programs have the capability of performing the chi-square test. If the appropriate computer software is not available, the calculations needed to perform the chi-square test can easily be done by hand as follows

1. Compute the expected number of responses in each JAR scale category for each product:

$$e_{ij} = \frac{(n_{i.})(n_{.j})}{n_{..}}$$

e_{ij} = expected number of responses in category j for the ith product

2. Compute the test statistic by using the observed number of responses and expected number in each JAR scale category for each product as shown below. The sum is taken over all products and all scale categories:

$$\chi^2 = \Sigma \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \Sigma \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

3. Compare χ^2 to the critical value from a table of the chi-square distribution at the desired significance level with degrees of freedom equal to (number of products – 1) × (number of scale categories – 1) (for chi-square table, see Appendix B of Ref. [8]).

Case Study Data Examples

Attribute = JAR Amt. Flavor

Number of responses in each category
(expected values in parentheses)

Product	Too Little	Just Right	Too Much	Totals
170	16	63	23	102
	(25)	(64)	(13)	
896	5	84	13	102
	(25)	(64)	(13)	
914	54	45	3	102
	(25)	(64)	(13)	
Totals	75	192	39	306

$$\text{Test statistic} = \frac{(16-25)^2}{25} + \frac{(63-64)^2}{64} + \frac{(23-13)^2}{13} + \cdots + \frac{(23-13)^2}{13}$$

$$= 80.17$$

Next, compare the test statistic to the critical value from a table of the chi-square distribution at the desired level of significance with degrees of freedom equal to (No. of products – 1) × (No. of JAR scale categories used in the analysis – 1). The test statistic 80.17 above is greater than the tabled value of 18.47 from the chi-square distribution with (three products – 1) × (three scale categories – 1) = 4 degrees of freedom at the 0.001 significance level. (for chi-square table, see Appendix B of Ref. [8].

Since this test indicates that there is a statistically significant difference in the JAR score distributions among the three products, the researcher may then want to do a follow-up test to determine whether there is a significant difference in the JAR score distributions between products 170 and 896. The first step is to create a subtable for Products 170 and 896. Then compute the expected number of responses for each product in each JAR scale category.

Number of responses in each category
(expected values in parentheses)

Product	"Too Little"	"Just Right"	"Too Much"	Totals
170	16	63	23	102
	(10.5)	(73.5)	(18)	
896	5	84	13	102
	(10.5)	(73.5)	(18)	
Totals	21	147	36	204

$$\text{Test statistic} = \frac{(16-10.5)^2}{10.5} + \frac{(63-73.5)^2}{73.5} + \frac{(23-18)^2}{18} + \cdots + \frac{(13-18)^2}{18} = 11.54$$

The test statistic 11.54 is greater than the tabled value of 10.60 from the chi-square distribution with 2 df (2 products – 1) × (3 scale categories – 1) at the 0.005 significance level.

Finally, suppose the researcher wants to determine whether the proportion of responses in the "Just Right" category is the same for Products 170 and 896. First, create a subtable for Products 170 and 896 with the responses for the "Too Little" and "Too Much" categories combined (see below). Then compute the expected number of responses for each product in each cateogory.

Number of responses in each category
(expected values in parentheses)

Product	"Just Right"	"Not Just Right"	Totals
170	63	39	102
	(73.5)	(28.5)	
896	84	18	102
	(73.5)	(28.5)	
Totals	147	57	204

$$\text{Test statistic} = \frac{(63-73.5)^2}{73.5} + \frac{(39-28.5)^2}{28.5} + \frac{(84-73.5)^2}{73.5} + \frac{(18-28.5)^2}{28.5} = 10.74$$

The test statistic 10.74 is greater than the tabled value of 7.88 from the chi-square distribution with 1 df (2 products – 1) × (2 scale categories – 1) at the 0.005 significance level.

Conclusions from the Analysis

There is a statistically significant difference in JAR scale distributions for Amount of Flavor among the three products ($p < 0.001$).

Based on the first follow-up analysis, the JAR scale distributions for Products 170 and 896 are significantly different ($p < 0.005$).

Based on the second follow-up analysis where JAR scale categories were combined, there is a statistically significant difference in the proportion of "Just Right" ratings between products 170 and 896.

Benefits of the Analysis

The chi-square method allows the researcher to determine whether there are significant differences in JAR score distributions between any number of products for any number of JAR scale categories in situations where assessors' ratings of the products are independent. When significant differences in JAR scale distributions are found, follow-up analyses can be done using the chi-square method to explore those differences further. The chi-square method is available in most statistical software programs, but the computations needed to carry out the method can easily be done without the use of a computer if the appropriate software is not available.

Caveats

Some researchers use the chi-square method instead of the CMH, Stuart-Maxwell, or McNemar methods to test for differences in JAR score distributions between two or more products, regardless of whether the assessors each evaluate only one product or all of the products. This is generally invalid. The *chi-square test requires that the assessors' ratings of the products be independent*. In studies where the same assessor evaluates more than one product, individual assessor responses on multiple products are often positively correlated. The CMH, Stuart-Maxwell and McNemar methods take this correlation into account, but the chi-square method does not. For this reason, the CMH, Stuart-Maxwell, or McNemar methods are more sensitive than the chi-square method when the responses for each assessor are positively correlated. If assessors' ratings of the products are positively correlated, then the p-values from the chi-square method are higher than p-values from the CMH, Stuart-Maxwell, or McNemar tests. Therefore, when the chi-square test is used in a situation where each assessor evaluates two or more of the products in the study, there is a possibility that differences in the distributions of JAR ratings among products will be declared as non-significant when statistically significant differences really do exist.

Recommendations

The chi-square method is recommended for comparing JAR score distributions among two or more products in situations when different groups of assessors evaluate each product. In situations when the products are all evaluated by the same group of assessors, then the CMH, Stuart-Maxwell, or McNemar methods are recommended instead.

References

[1] Mantel, N. and Haenszel, W. J., "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease," *Nat. Cancer Inst.*, Vol. 22, 1959, pp. 719–748.

[2] Kuritz, S. J., Landis, J. R., and Koch, G. G., "A General Overview of Mantel-Haenszel Methods: Applications and Recent Developments," *Annu. Rev. Public Health*, Vol. 9, 1988, pp. 123–160.

[3] Stuart, A. "A Test for Homogeneity of the Marginal Distributions in a Two-Way Classification," *Biometrika*, Vol. 42, 1955, pp. 412–416.

[4] McNemar, Q., "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages," *Psychometrika*, Vol. 12, 1947, pp. 153–157.

[5] SAS Institute Inc., *SAS/STAT User's Guide*, Version 6, 4th edition, Vol. 1, SAS Institute, Inc., Cary, NC, 1989, 943 pp.

[6] Stone, H. and Sidel, J. L., *Sensory Evaluation Practices*, 2nd edition, Academic Press, New York, 1993, pp. 88–91.

[7] Fleiss, J. L., *Statistical Methods for Rates and Proportions*, 2nd edition, Wiley, New York, 1981.

[8] Agresti, A., *Categorical Data Analysis*, 2nd edition, Wiley, New York, 2002.

Appendix H: A Proportional Odds/Hazards Approach to JAR Data

Rui Xiong[1] and Jean-Francois Meullenet[1]

Introduction and Background

t-tests, ANOVA, and linear regression all assume that the response is measured on an interval scale, so that the differences between adjacent values have the same meaning across the scale. This assumption is often violated in practice, which can lead to inaccurate conclusions. The proportional odds and proportional hazards models are ordinal regression models that are only sensitive to the order of the observations, not the specific values assigned to the categories. They are used to compare the distributions of JAR scores among products and are performed simultaneously. The proportional odds model (POM) is skew-symmetric, so that reversing the order of the scale simply changes the sign of the mean, while the proportional hazards model (PHM) is asymmetric, so that reversing the order of the scale changes both the order and the sign of the estimate.

Requirements

Raw data of respondent/product/attribute combinations are required for the analysis. These techniques are compute-intensive and require specialized programs, such as SAS/STAT, SPSS, or R. Additionally, the JAR ratings are assumed to be independent; however, this practice in often violated.

"How to"

The proportional odds and hazards models are widely used in medicine and life science survey studies. Recently, both models have been applied to the sensory field for preference map [1] and shelf-life studies [2], respectively. These articles or the book by Agresti [3] should be consulted for technical details. The proportional odds and hazards models have the same underlying assumption, but they use different link functions to model ordinal response data. The comparison of POM and PHM is presented in Table 1. The goodness of fit for both POM and PHM is assessed by the likelihood ratio or deviance G^2.

Proportional Odds Model (POM)

The proportional odds model [3] models the odds of being at or below each scale point across the products (the cumulative odds), and determines an average ratio of those cumulative odds between products. Since the model works with odds and ratios of odds, it is traditional to express the model in terms of logits (log-odds). The model is fit using maximum likelihood and produces estimates of the average log-odds for each scale point as well as for each product included in the analysis. By default, one product, the control, is always set to zero.

The SAS implementation of this model includes a test to determine if the same rating scale was used across the products, and, if included, panelists in the study. It is referred to as a test of equal slopes or of parallelism. This is a generalization of a test for homogeneity of variances in a *t*-test. When this is significant, the data do not meet the assumptions for this analysis.

Proportional Hazards Model (PHM)

The proportion hazards model, also known as a Cox regression model, also considers the odds, but looks at the odds of being in each category, given that the observation is not in the categories below it, and again estimates the average ratio of those odds across products. As with the POM, it is traditional to use logarithms and to express the results on that scale. The analysis does not treat the data symmetrically; the results depend on the order in which the scale points are coded. This analysis is most appropriate when the rating can be viewed as the result of a progression, as in life data where the model originated.

The SAS implementation of this uses the same procedure as does the POM and similarly includes a parallelism test.

Example from Case Study Data

The data from the case study were analyzed using PROC Logistic in SAS/STAT. The following code was used to fit a POM to the Flavor attribute:

TABLE 1—Comparison of the proportional odds and hazards models.

	Proportional Odds Model	Proportional Hazards Model
Assumption	Equal slopes across levels of a response variable	Equal slopes across levels of a response variable
Model	$P[Y \leqslant k] = \frac{1}{1+e^{-(\alpha_k+\beta' x)}}$	$P[Y \leqslant k] = 1 - e^{-e^{\alpha_k+\beta' x}}$
Link function	Logit	Complementary log-log

[1] *Department of Food Science, University of Arkansas, Fayetteville, AR 72704.*

TABLE 2—Parameter estimates from the proportional odds model for size.

Parameter	Estimate	Standard Error	Chi-square	*p*-value
Intercept1	−3.048	0.2597	137.769	<0.0001
Intercept2	−1.181	0.1916	37.990	<0.0001
Intercept3	0.359	0.1843	3.799	0.0513
Intercept4	2.564	0.2397	114.426	<0.0001
Sample 458	0.382	0.2544	2.255	0.1332
Sample 523	0.081	0.2543	0.102	0.7495
Sample 896	−0.026	0.2544	0.010	0.9197
Sample 914	0.138	0.2542	0.296	0.5866

```
Proc logistic data=CaseStudy;
    Class Sample (ref="170")/param=ref;
    Model Flavor=Sample/link=logit scale=none aggregate;
    Title "Proportional odds model for Flavor;"
    Contrast "Samples 458" vs "170" Sample 1 0 0 0 /estimate=both;
    Contrast "Samples 523" vs "170" Sample 0 1 0 0 /estimate=both;
    Contrast "Samples 896" vs "170" Sample 0 0 1 0 /estimate=both;
    Contrast "Samples 914" vs "170" Sample 0 0 0 1 /estimate=both;
    Contrast "Samples 523" vs "458" Sample −1 1 0 0 /estimate=both;
    Contrast "Samples 896" vs "458" Sample −1 0 1 0 /estimate=both;
    Contrast "Samples 914" vs "458" Sample −1 0 0 1 /estimate=both;
    Contrast "Samples 896" vs "523" Sample 0 −1 1 0 /estimate=both;
    Contrast "Samples 914" vs "523" Sample 0 −1 0 1 /estimate=both;
    Contrast "Samples 914" vs "896" Sample 0 0 −1 1 /estimate=both;
Run;quit;
```

This code both fits the model (the "model" statement) and performs pairwise comparisons of the products (the "Contrast" statements). A similar program was used for the other attributes compared in the results sections.

The code for the PHM is quite similar to the code given above, with the only difference being the "link=" specification in the model statement. Note that the link specification becomes "link=cloglog," highlighted below:

```
Proc logistic data=CaseStudy;
    Class Sample (ref="170")/param=ref;
    Model Flavor=Sample/link=cloglog scale=none aggregate;
    Title "Proportional hazards model for Flavor;"
    Contrast "Samples 458" vs "170" Sample 1 0 0 0 /estimate=both;
    Contrast "Samples 523" vs "170" Sample 0 1 0 0 /estimate=both;
    Contrast "Samples 896" vs "170" Sample 0 0 1 0 /estimate=both;
    Contrast "Samples 914" vs "170" Sample 0 0 0 1 /estimate=both;
    Contrast "Samples 523" vs "458" Sample −1 1 0 0 /estimate=both;
    Contrast "Samples 896" vs "458" Sample −1 0 1 0 /estimate=both;
    Contrast "Samples 914" vs "458" Sample −1 0 0 1 /estimate=both;
    Contrast "Samples 896" vs "523" Sample 0 −1 1 0 /estimate=both;
    Contrast "Samples 914" vs "523" Sample 0 −1 0 1 /estimate=both;
    Contrast "Samples 914" vs "896" Sample 0 0 −1 1 /estimate=both;
Run;quit;
```

Results and Conclusions

For the JAR attribute Size, the chi-square (χ^2) for testing the equal slopes assumption was 7.6 with p-value of 0.814, which was not significant with respect to a chi-square distribution with 12 degrees of freedom (DF) at a significance level (α) of 0.05. This suggested that the parallelism assumption was satisfied. The likelihood ratio (deviance) G^2 was 7.132 (DF = 12) with $p = 0.849$, indicating that the proportional odds model adequately fitted the data. The parameterization used in the SAS system is one that leaves out the parameter for the baseline (Sample 170 in this case) with which each sample is compared. Hence, a positive parameter estimate (β) in Table 2 means that Sample 170 was "larger" in size than the compared sample, while a negative estimate

TABLE 3—Estimates of parameters and odds ratios from the proportional odds model for Color.

Effect	Estimate	Standard Error	Odds Ratio	Wald Chi-square	*p*-value
Intercept1	–6.629	1.0450		40.246	<0.0001
Intercept2	–2.431	0.3266		55.393	<0.0001
Intercept3	3.458	0.3926		77.568	<0.0001
Sample 458 vs 170	0.575	0.4208	1.777	1.867	0.1718
Sample 523 vs 170	0.091	0.4408	1.095	0.043	0.8363
Sample 896 vs 170	–0.713	0.4653	0.490	2.347	0.1255
Sample 914 vs 170	1.124	0.4008	3.078	7.869	0.0050
Sample 523 vs 458	–0.484	0.4156	0.6164	1.355	0.2444
Sample 896 vs 458	–1.288	0.4547	0.276	8.020	0.0046
Sample 914 vs 458	0.549	0.3639	1.732	2.279	0.1311
Sample 896 vs 523	–0.804	0.4642	0.448	2.999	0.0833
Sample 914 vs 523	1.033	0.3947	2.810	6.850	0.0089
Sample 914 vs 896	1.837	0.4415	6.278	17.311	<0.0001

means that Sample 170 was "smaller" in size. The *p*-value for the chi-square statistic is used to test whether the difference between the compared sample and the baseline sample is significant. Since all the *p*-values were much greater than $\alpha = 0.05$ (Table 2), all the Samples were not significantly different in size from sample 170. Overall, the effect of products was not significant ($\chi^2 = 3.273$, DF = 4, *p*-value = 0.513) at $\alpha = 0.05$. This suggested that all the products had a similar distribution of Size scores.

For the JAR attribute color, the parallelism (equal slopes) assumption was not met for POM ($\chi^2 = 15.900$, DF = 8, *p*-value = 0.044) at $\alpha = 0.05$, but was met at $\alpha = 0.01$. This typically can occur when one or more of the products is more variable than the remaining products. The overall effect of products was then significant ($\chi^2 = 19.944$, DF = 4, *p*-value = 0.0005), implying that some of the products have different mean log-odds. The products which were significantly different in color can be identified using the included contrasts. We used a Bonferroni correction to account for the 10 multiple tests, testing each pairwise comparison at $\alpha = 0.05/10 = 0.005$.

The parameter estimates and odds ratios between all pairs of the samples were obtained from contrasts and are presented in Table 3. The *p*-values in Table 3 were used to test if a pair of samples was significantly different at $\alpha' = 0.005$. For example, the *p*-value for the pair of Samples 458 and 170 was $0.1718 > \alpha' = 0.005$, indicating that the two samples were not significantly different in color (i.e., meaning the JAR score distributions were similar not that the products were identical in color). The *p*-values for the pairs of Samples 896 versus 458 and 914 versus 896 were 0.0046 and <0.0001, respectively, which indicated that Sample 896 was significantly different in color from Samples 458 and 914. As mentioned above, the signs of parameter estimates can be used to determine the directional difference between two products. Sample 896 was overall significantly darker in color than Sample 458 because of the negative parameter estimate of –1.288, while Sample 914 was significantly lighter in color than Sample 896 because of the positive estimate of 1.837. As result, sample 896 was significantly darker in color than Samples 458 and 914, and other pairs of samples were not significantly different in color. The contrasting method provided for the POM in the SAS LOGISTIC procedure is another advantage over the two-stage test procedure like the chi-square/McNemar tests. The interpretation of parameters is usually done using odds ratios. For example, the odds ratio of 6.278 ($=e^{1.837}$, 1.873 was the parameter estimate) for Samples 914 versus 896 (Table 3) means that the odds of consumer rating sample 914 as "Too Light" in color was 6.278 times the odds for Sample 896, so consumers rated sample 914 lighter in color than Sample 896.

When the parallelism test is significant, this means that there is differences between the codes beyond a simple mean shift. This suggests that the analyst consider alternative models to determine if the same conclusions hold. In this case the parallelism assumption was not significant for PHM ($\chi^2 = 14.014$, DF = 8, *p*-value = 0.081) at $\alpha = 0.05$. The overall effect of products was significant ($\chi^2 = 16.875$, DF = 4, *p*-value = 0.002) at the significance level of 0.05, suggesting that some products have different distributions for Color JAR scores. The parameter estimates for PHM are provided in Table 4. Like POM, a positive parameter estimate for PHM (Table 4) means that Sample 170 was "Darker" in color than the compared sample, while a negative estimate means that sample 170 was "Lighter" in color. The *p*-values show that Samples 914 versus 170, 914 versus 523, and 914 versus 896 were significantly different from each other ($\alpha' = 0.005$), respectively. Sample 914 had "Lighter Color" JAR scores than Samples 170, 523, and 896. By comparing Tables 3 and 4, the results

TABLE 4—Parameter estimates from the proportional hazards model for Color.

Effect	Estimate	Standard Error	Wald Chi-square	*p*-value
Intercept1	–6.654	1.0138	43.078	<0.0001
Intercept2	–2.539	0.2106	145.381	<0.0001
Intercept3	1.099	0.1384	62.957	<0.0001
Sample 458 vs 170	0.477	0.2323	4.213	0.0401
Sample 523 vs 170	0.059	0.1960	0.092	0.7618
Sample 896 vs 170	0.011	0.1937	0.003	0.9569
Sample 914 vs 170	1.076	0.2943	13.377	0.0003
Sample 523 vs 458	–0.417	0.2339	3.185	0.0743
Sample 896 vs 458	–0.466	0.2326	4.021	0.0449
Sample 914 vs 458	0.599	0.2979	4.050	0.0442
Sample 896 vs 523	–0.049	0.1964	0.062	0.8032
Sample 914 vs 523	1.017	0.2944	11.928	0.0006
Sample 914 vs 896	1.066	0.294	13.115	0.0003

from both POM and PHM were different. In this case, we trusted the results from PHM because the parallelism assumption was met at $\alpha = 0.05$ for PHM but not for POM. A disadvantage of PHM is that it does not provide odds ratios for the interpretation of parameters

For the JAR attribute flavor, the parallelism assumption was satisfied for POM ($\chi^2 = 20.425$, DF = 12, p-value = 0.06) at $\alpha = 0.05$. The overall effect of products was significant ($\chi^2 = 105.198$, DF = 4, p-value = 0.0001), indicating that the products were not from the same population for flavor. The p-values show that Sample 914 had significantly lower JAR flavor scores than all other samples because of the positive estimates, while Sample 458 had significantly higher JAR flavor than Samples 170, 896, and 914.

For the JAR attribute Thin/Thick, the equal slopes assumption was met for POM ($\chi^2 = 13.171$, DF = 8, p-value = 0.106) at $\alpha = 0.05$. The overall effect of products was significant ($\chi^2 = 27.096$, DF = 4, p-value < 0.0001) at the significance level of 0.05, suggesting that not all products had similar distributions of their respective thin/thick scores. The p-values together with the signs of the parameter estimates show that only sample 914 had significantly higher JAR thickness scores than all other Samples and other samples were not significantly different from each other.

For the JAR attribute Stickiness, the parallelism assumption was met for POM ($\chi^2 = 10.129$, DF = 12, p-value = 0.605) at $\alpha = 0.05$, but the overall effect of products was not significant ($\chi^2 = 8.511$, DF = 4, p-value = 0.075) at the significance level of 0.05. There was no sufficient evidence to conclude that all the products did not come from the same distribution of stickiness scores.

Conclusions from the Analysis

For both size and stickiness, there was no evidence that the distributions of the JAR scores for the various products were different.

For Color, Flavor, and Thin/Thick attributes, there were significant differences among the samples. Sample 914 had significantly lower JAR Color scores than Samples 170, 523, and 896. Sample 914 had significantly lower JAR flavor scores than other samples, while Sample 458 had significantly higher flavor scores than Samples 170 and 914. For the JAR attribute Thin/Thick, only Sample 914 had significantly higher JAR scores than all other samples.

Pros and Cons

The primary benefit of these models is that normal distributions of data are not required. A secondary benefit in the SAS implementation is the built-in test of the equal slopes (parallelism) assumptions. If the parallelism assumption is met, the overall product effect can be assessed; if this is significant, differences between product pairs can be assessed.

When the parallelism assumptions are not met, the analysis can be compromised. This is on the level of failing the homogeneity test in an ANOVA. When this occurs, the analyst should either try an alternate model or identify and correct the offending codes. When the parallelism assumption for POM fails, it is recommended to use the proportional hazards model. When neither model is appropriate, the general multinomial logistic model should be considered.

Recommendations

These analyses are recommended as a means to determine whether similar JAR distributions exist between products when the data are not normally distributed.

References

[1] Meullenet, J.-F., Xiong, R., Hankins, J. A. R., Dias, P., Zivanovic, P., Monsoor, M. A., Bellman-Homer, T., Liu, Z., and Fromm, H., "Preference Modeling of Commercial Toasted White Corn Tortilla Chips Using Proportional Odds Model," *Food Quality Preference* Vol. 14, No. 7, 2003, pp. 603–614.

[2] Gimenez, A. M., Gambaro, A., Varela, P., Garitta, L., and Hough, G., "Use of Survival Analysis Methodology to Estimate Shelf-life of "alfajor"," *Proceeding of 2003 Pangborn Sensory Evaluation Meeting*, Boston, MA.

[3] Agresti, A., *Categorical Data Analysis*, Wiley, New York, 1990.

Appendix I: Student's *t*-Test—Analysis of Variance of Two Samples

Merry Jo Parker[1]

Introduction and Objectives

The Student's *t*-test is a statistical method for comparing the mean JAR scores of two product samples. The results of the *t*-test determine whether or not the means of two samples of data are significantly different.

Requirements

To conduct a *t*-test analysis, the distribution of data from each panelist for each product is needed. Data are required to be normally distributed with homogeneous variance.

"How to"

When each panelist has rated both samples on the same JAR scale, the test is a typical paired *t*-test, discussed in most introductory statistics texts. When different groups of respondents rate the two products, the appropriate test is the two-sample *t*-test. It is recommended that the data be tested for normality before applying this test. If that test fails, a sign-test or Wilcoxin-Mann-Whitney test should be used in place of the paired *t*-test and two-sample *t*-test, respectively. The Student's *t* test is meant to be used when there are only two samples. If there are more, then the analyst should perform a Randomized Complete Block (RCB) Analysis of Variance (paired data) or a one-way ANOVA (independent samples), followed by pairwise comparisons between the codes. In the RCB analysis, the subjects would be the blocks, and the products would be the treatments.

Example

In the example below, Samples 170 and 194 are compared on the Size, Color, Flavor, Thick/Thin, and Stickiness scales. Table 1 displays the marginal counts and means for both samples on each attribute. Note that the individual differences are required. In this example, tests of normality are not included.

Results and Conclusions

The *t*-tests suggest that there are significant differences in color, amount of flavor, viscosity, and stickiness between Sample 914 and Sample 170. There were no significant differences among the samples in size. The mean score ratings for 914 suggest that it is too light in color, too weak in flavor, and too thick. For stickiness, the mean score ratings for 170 is higher than 914, suggesting that 170 may be too sticky. (See Table 2.)

Pros and Cons

The scales of size, stickiness, color, and thickness are representative of why caution should be exercised when using the Student's *t* test to analyze JAR scales. Student's *t* test analysis assumes that the data are normal in distribution and are homogeneous in variance. The distribution of scores for size is not normal; it is bimodal, indicating a varied range of opinions on the ideal size of the product resulting in no clear direction for change. Likewise, Student's *t*-test analysis indicated a significant difference between Sample 914 and 170 for stickiness, thickness, flavor, and color. If the researcher is only using Student's *t* test to analyze the JAR scores, the conclusion would be that Sample 914 should have a darker color and thinner consistency. The JAR distribution of scores indicates that Samples 170 and 914 both have high "Just About Right" ratings, i.e., 75 % and higher, for each of these attributes. The question is, even though these samples are statistically significantly different, i.e., the JAR scores for Sample 914 are high, should these attributes really be changed?

Recommendation

T test analysis can be an effective method for evaluating the mean differences between two samples using JAR scales. However, it should always be used in combination with an evaluation of the score distributions. A bimodal distribution may indicate subgroups within the population that is being tested, or it may indicate panelist inability/confusion with a scale.

TABLE 1—Marginal data and means.

	Size		Color		Flavor Amt		Thick/Thin		Stickiness	
Size	170	914	170	914	170	914	170	914	170	914
1= "Too Low"	3	9	0	0	2	6	0	0	0	2
2= "Somewhat Too Low"	21	19	11	21	14	49	12	4	5	8
3= "Just About Right"	37	32	85	81	63	44	85	77	80	83
4= "Somewhat Too High"	33	37	6	0	21	3	5	21	16	9
5= "Too High"	8	5	0	0	2	0	0	0	1	0
Means	3.22	3.10	2.95	2.79	3.07	2.43	2.93	3.17	3.13	2.97

[1] *Food Perspectives, 2880 Vicksburg Lane, Plymouth, MN 55447.*

TABLE 2—Summary of paired *t*-test calculations and results.

	Σd	Σd^2	**Mean** *d*	*S*	$t_{(df=101)}$	*p*-value
Size	12	170	0.12	1.29	0.92	<0.05
Color	16	28	0.16	0.50	3.15	<0.01
Flavor amt.	64	128	0.63	0.93	6.79	<0.01
Thick/Thin	–25	37	0.25	0.55	4.48	<0.01
Stickiness	17	49	0.17	0.68	2.45	<0.05

Appendix J: Analysis of Variance (ANOVA)

Merry Jo Parker[1]

Introduction and Objectives

Analysis of Variance (ANOVA) is a statistical method for studying differences between the mean scores of samples. ANOVA takes into account variance from different sources. When used to analyze "Just About Right" (JAR) scales, the source of variance is most often treatments and judges, so a two-way ANOVA is used.

Requirements

To conduct an ANOVA analysis, the distribution of data from each panelist for each product is needed. Data should be normally distributed with homogeneous variance.

"How to"

When the data are from dependent samples (e.g., each panelist judges two or more of the products on the attributes) the data should be analyzed using a Randomized Complete Blocks (RCB) ANOVA. When the data are from independent samples (e.g., each panelist judges only one sample) the data can be analyzed using a one-way ANOVA. Additional covariates (day of testing, order of presentation, etc.) can also be included in the analysis.

The computational details are beyond the scope of this document and will not be covered here. Please consult a statistical textbook for details. Likewise, the data preparation tends to be dependent on the particular statistical package being used for the analysis and will not be considered here. Generally these analyses require individual level data.

Example

Five JAR attributes for samples 170, 896, and 914 have been analyzed. The marginal data and analysis summaries are presented in the Appendix, while the conclusions are given below.

Results and Conclusions

The ANOVA results indicate that there were significant differences in color, amount of flavor, thickness, and stickiness between Sample 914 and the other two samples (170 and 896). There were no significant differences among the samples in size. Samples 170 and 896 were similar to each other for all five attributes. The mean score ratings for 914 suggest that it is too light in color, too weak in flavor, and too thick. For stickiness, the mean score ratings for 170 and 896 are higher than 914, suggesting that they may be too sticky.

Pros and Cons

The scales of size, stickiness, color, flavor, and thickness are representative of why caution should be exercised when using ANOVA to analyze JAR scales. ANOVA analysis assumes that the data are normal in distribution and are homogeneous in variance. The distribution of scores for size is not normal; it is bimodal, indicating a range of opinions on the ideal size of the product, resulting in no clear direction for change. Likewise, ANOVA analysis indicated a significant difference between Sample 914 and both 170 and 896, which were at parity, for stickiness, thickness, flavor, and color. If the researcher is only using ANOVA to analyze the JAR scores, the conclusion would be that the Overall Liking for Sample 914 could be improved if it were to have a darker color, stronger flavor, and be thinner. Note, however, that all these products had high (>75 %) "Just About Right" scores for these attributes. The business question then becomes, "Is it worth the cost to improve Product 914?"

Recommendation

ANOVA analysis can be an effective method for evaluating mean JAR scale differences; however, the distribution of the JAR responses should always be evaluated prior to interpreting the ANOVA. A bimodal distribution may indicate subgroups within the population that is being tested, or it may indicate panelist inability/confusion with a scale.Appendix

Raw Data (condensed) followed by ANOVA tables.

Size	170	914	896
1 = "Much Too Small"	3	9	3
2 = "Too Small"	21	19	19
3 = "Just About Right"	37	32	41
4 = "Too Large"	33	37	30
5 = "Much Too Large"	8	5	9
Means	3.22	3.10	3.23

	D.F.	Sum of Squares	Mean of Squares	*F*-value	*p*-value
Samples	2	1.026	0.513	0.66	0.5178
Judges	101	139.114	1.377	1.77	0.0003
Error	202	156.974	0.777		
Total	305	297.114	0.974		
Std. Error (SEM)	0.087				

Tukey's HSD 5 % = 0.293*

No Significant Differences

*Tukey's HSD is the difference needed between the means of 170, 914, and 896 for a sample to be significantly different from another sample for this attribute.

Color	170	914	896
1 = "Much Too Light"	0	0	0
2 = "Too Light"	11	21	1
3 = "Just About Right"	85	81	98
4 = "Too Dark"	6	0	3

[1] *Food Perspectives, 2880 Vicksburg Lane, Plymouth, MN 55447.*

Color	170	914	896
5 = "Much Too Dark"	0	0	0
Means	2.95	2.79	3.02

	D.F.	Sum of Squares	Mean of Squares	*F*-value	*p*-value
Samples	2	2.725	1.363	13.8	0.000
Judges	101	17.451	0.173	1.75	0.0004
Error	202	19.941	0.099		
Total	305	40.118	0.132		
Std. Error (SEM)	0.031				

Tukey's HSD 5 % = 0.105*

Significant Differences = 170 versus 914 and 896 versus 914

*Tukey's HSD is the difference needed between the means of 170, 914, and 896 for a sample to be significantly different from another sample for this attribute.

Amount of Flavor	170	914	896
1 = "Much Too Weak"	2	6	0
2 = "Too Weak"	14	49	6
3 = "Just About Right"	63	44	83
4 = "Too Strong"	21	3	12
5 = "Much Too Strong"	2	0	1
Means	3.07	2.43	3.08

	D.F.	Sum of Squares	Mean of Squares	*F*-value	*p*-value
Samples	2	28.046	14.023	43.83	0.000
Judges	101	50.291	0.498	1.56	0.0042
Error	202	202	64.621	0.320	
Total	305	305	142.958		
Std. Error (SEM)	0.056				

Tukey's HSD 5 % = 0.188*

Significant Differences = 170 versus 914 and 896 versus 914

*Tukey's HSD is the difference needed between the means of 170, 914, and 896 for a sample to be significantly different from another sample for this attribute.

Thinness/Thickness	170	914	896
1 = "Much Too Thin"	0	0	1
2 = "Too Thin"	12	4	8
3 = "Just About Right"	85	77	90
4 = "Too Thick"	5	21	3
5 = "Much Too Thick"	0	0	0
Means	2.93	3.17	2.93

	D.F.	Sum of Squares	Mean of Squares	*F*-value	*p*-value
Samples	2	3.765	1.882	13.79	0.000
Judges	101	25.637	0.254	1.86	0.0001
Error	202	27.569	0.136		
Total	305	56.971	0.187		
Std. Error (SEM)	0.036				

Tukey's HSD 5 % = 0.123*

Significant Differences = 170 versus 914 and 896 versus 914

*Tukey's HSD is the difference needed between the means of 170, 914, and 896 for a
sample to be significantly different from another sample for this attribute.

Stickiness	170	914	896
1 = "Not Nearly Sticky Enough"	0	2	0
2 = "Not Sticky Enough"	5	8	1
3 = "Just About Right"	80	83	88
4 = "Too Sticky"	16	9	12
5 = "Much Too Sticky"	1	0	1
Means	3.13	2.97	3.13

	D.F.	Sum of Squares	Mean of Squares	*F*-value	*p*-value
Samples	2	1.673	0.837	4.33	0.0144
Judges	101	24.605	0.244	1.26	0.0833
Error	202	38.993	0.193		
Total	305	65.271	0.214		
Std. Error (SEM)	0.043				

Tukey's HSD 5 % = 0.146*

Significant Differences = 170 versus 914 and 896 versus 914

*Tukey's HSD is the difference needed between the means of 170, 914, and 896 for a
sample to be significantly different from another sample for this attribute.

Appendix K: Thurstonian Ideal Point Modeling

Jeannine Delwiche[1]

Introduction and Objectives

Thurstonian ideal point modeling allows one to compare the JAR ratings of multiple products to a theoretical ideal product. It compares the probabilistic distribution of a product against the probabilistic distribution of the ideal.

Background

When thinking about the ratings of an ideal product, researchers tend to conceptualize an ideal product as always receiving a rating of "Just Right" (Fig. 1).

However, a respondent's product perceptions as well as his definition of the ideal product may vary over time. Even in the absence of product variation, JAR scale ratings for a truly "ideal" product would therefore approximate a normal distribution (Fig. 2).

Considering JAR ratings as distributions rather than absolute points can lead to the following situations: A product could have a distribution similar to that of the ideal product, but with a mean that deviates significantly from the ideal (as in Fig. 3). Chi-square analysis reveals that the two distributions below are significantly different.

On the other hand, a highly variable product could have a mean "Just Right" rating, and yet not be ideal because of "heavy tails" (Fig. 4—notice the mean is greatly depressed). Chi-square analysis again reveals a significant difference between the two distributions.

Thurstonian ideal point modeling allows one to compare multiple products to a theoretical ideal product. It compares the probabilistic distribution of a product against the probabilistic distribution of the ideal. IFPrograms™ provides the estimation of scale means relative to the ideal mean for each scale. These means are in units of d', measured from the ideal point. In addition, the program gives relative scale boundaries, ideal product proportions for each category for each scale, and the variance-covariance matrix of the scale means.

It is necessary to elaborate on what is meant by the "estimation of scale means" and "relative scale boundaries." While rating scales are generally assumed to have equal interval spacing, (Fig. 5, top), respondents often use the scales as though they were unequally spaced (Fig. 5, bottom), specifically, the end categories of the scale are used less often than the other points. The ratings, therefore, are more ordinal than interval in nature, which is a violation of parametric statistics [1]. Thurstonian ideal point modeling is able to account for these psychological effects, and converts the rating values that are based upon a number system without equal intervals to true scale values that are based upon a number system with equal intervals. These scale values are given in terms of d' and can be determined not only for the ideal distribution, but also for the rated product(s).

Requirements

The only requirement for analysis is that data be categorical. To minimize data distortion due to transformation, it is recommended that ratings be collected on categorical scales.

"How to"

To conduct this analysis, the response frequencies for each category by product and scale are determined. One typically looks at one scale at a time for more than one product, but it is also possible to look at several scales for one product. One inputs the frequency data (how often each category of the scale was selected for each product for a given scale), and a few other details (number of scales, number of products, and number of scale categories).

Results and Conclusions

The output for "size" is shown below in Fig. 6.

First listed is the "Relative Boundaries" of the scale, which indicates the actual size of the intervals subjects are using. Beneath this is the "Ratings Means," which is simply the mean value of the ratings. This is followed by the "Scale Means," which are, as mentioned earlier, means in units of d', measured from the ideal point. Next is the "Variance-covariance matrix for scale means," and the values on the diagonal are the variance associated with each product (for that scale). From the scale means (in d') and their associated variance, one can use another function of the IFPrograms™, "Comparing d' values," to see if products differ from one another significantly. The results from this analysis are summarized in Table 1.

The final line gives the "Ideal (or Reference) Proportions for Each Category." These values can be subsequently used in chi-square analyses, comparing each product distribution relative to the ideal product distribution. For the "Expected Values," one uses the ideal proportions multiplied by the number of observations. The output from these analyses is summarized in Table 2.

The first step is to compare the d' values of the samples (Table 1). Samples 170 and 896 are not significantly different from one another, based upon chi-Square analysis performed by IFPrograms™ of d' and their corresponding variance. However, compared to samples 170 and 896, sample 914 is significantly lower in amount of flavor and stickiness, and significantly higher in amount of salt and thickness. However, the samples show no significant difference in size.

The next step is to compare the JAR distributions of each sample to the JAR distribution for the "ideal" sample (Table 2). None of the samples differs significantly from the ideal size; only sample 914 differs significantly from the ideal for amount of flavor and salt, both samples 896 and 914 differ significantly from the ideal for thin/thickness, and both

[1] *Firmenich, Inc., PO Box 5880, Princeton, NJ 08543.*

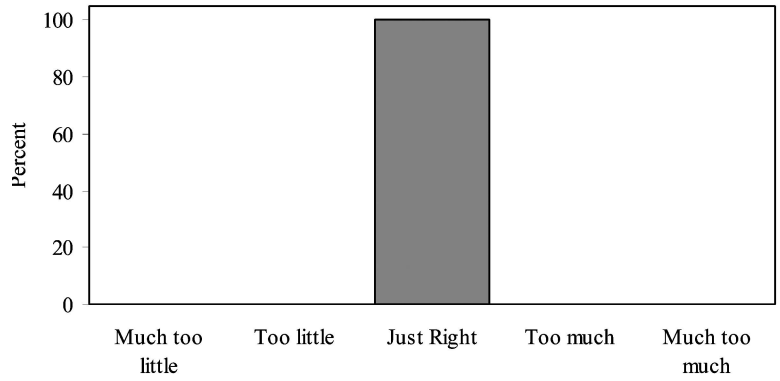

Fig. 1—JAR distribution for a constant ideal product.

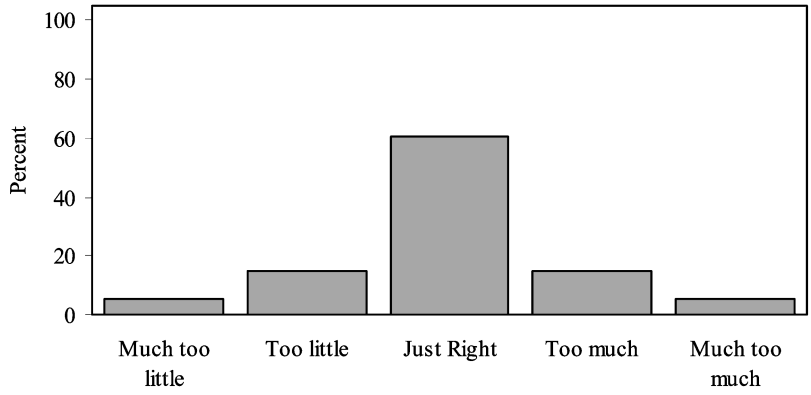

Fig. 2—JAR distribution for an ideal product distribution.

samples 170 and 896 differ significantly from the ideal for stickiness. Overall, it is sample 914 that deviates the most from the ideal, tending towards too large a size, and too much flavor, salt, and thickness. Sample 170 is closest to the ideal, tending towards too much flavor and having somewhat too much stickiness, but not differing from the ideal size, amount of salt, and thickness. Sample 896 shows intermediate results, not differing from the ideal in size and amount of salt, but tending towards too much flavor and having too much stickiness and not enough thickness.

Pros and Cons

The benefit of this analysis is that by using probabilistic modeling, the variant nature of the ideal product is accounted for. The determination of the ideal distribution provides a way to determine reasonably the expected values of the ideal product category frequencies and the ability to perform chi-square analyses. It allows differentiation between samples that do not differ from the distribution of the ideal

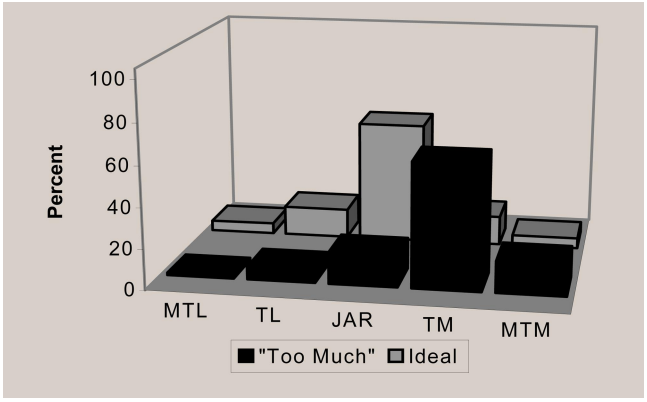

Fig. 3—Comparisons of JAR distribution for ideal vs too much.

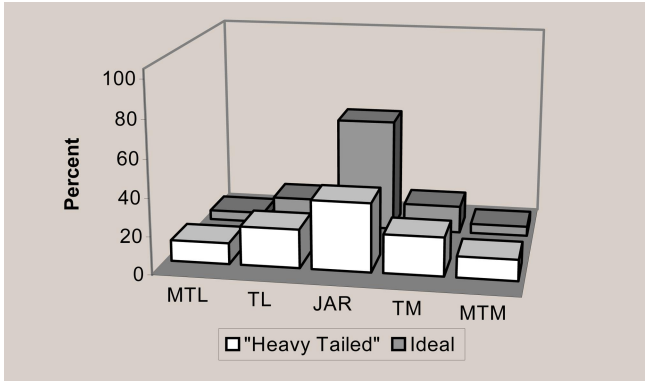

Fig. 4—Comparisons of JAR distribution for ideal vs heavy tailed.

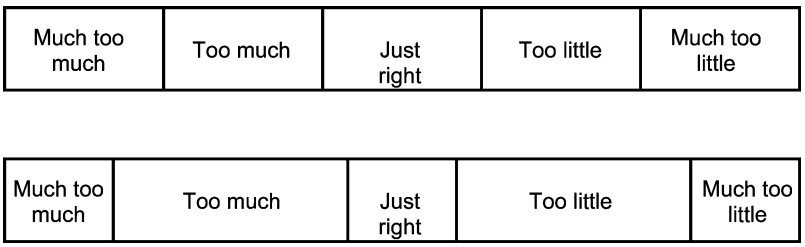

Fig. 5—Equal and unequal interval scale boundaries.

```
* Analysis of Relative-to-Reference, Rated 2-AC or Just About Right Scale Data

      Scale:   1

    Relative Boundaries
         2.238
          .669

  Reference, Alternative or Ideal Mean: 0.000

     Rating Means
        3.216
        3.235
        3.098

     Scale Means
         .323
         .354
         .133

  Variance-covariance matrix for scale means

  .022  .000  .000
  .000  .022  .000
  .000  .000  .022

Ideal (or Reference) Proportions for Each Category
.057 .261 .364 .261 .057
```

Fig. 6—Output of IFProgram™ JAR scale analysis (Thurstonian ideal point modeling) for "size" attribute.

TABLE 1—Comparison of d' values of samples' JAR ratings.

Attribute	170	896	914	Chi-square	*p*-value
Size	0.323[a] (0.022)	0.354[a] (0.022)	0.133[a] (0.022)	1.30	0.52
Amount of Flavor	0.192[a] (0.027)	0.257[a] (0.029)	–1.473[b] (0.033)	61.23	<0.01
Amount of Salt	–0.057[a] (0.022)	0.001[a] (0.023)	0.491[b] (0.023)	7.96	0.02
Thin/Thickness	–0.373[a] (0.037)	–0.331[a] (0.039)	0.750[b] (0.040)	20.63	<0.01
Stickiness	0.588[a] (0.042)	0.613[a] (0.039)	–0.100[b] (0.038)	8.45	0.01

In each row, means with the same superscript are not significantly different from one another ($p < 0.05$)

TABLE 2—Comparison of JAR distributions to "Ideal" sample JAR distributions.

Size				
Product vs Ideal	**Chi-sq=4.90** *p<0.298*	**Chi-sq=6.23** *p<0.182*	**Chi-sq=8.80** *p<0.066*	
	170	**896**	**914**	**Ideal**
Too little	3	3	9	5.8
Somewhat too little	21	19	19	26.6
Just Right	37	40	32	37.1
Somewhat too much	33	31	37	26.6
Too much	8	9	5	5.8
Amount of Flavor				
Product vs Ideal	Chi-sq=9.07 *p<0.059*	Chi-sq=9.06 *p<0.060*	Chi-sq=139.54 ***p<1×10⁻¹***	
	170[a]	**896**[a]	**914**[a]	**Ideal**
Too little	2	0	6	0.7
Somewhat too little	14	5	48	14.2
Just Right	63	84	45	72.1
Somewhat too much	21	12	3	14.2
Too much	2	1	0	0.7
Amount of Salt				
Product vs Ideal	Chi-sq=3.34 *p<0.502*	Chi-sq=6.25 *p<0.181*	Chi-sq=16.33 ***p<0.003***	
	170	**896**	**914**[a]	**Ideal**
Too little	16	15	14	17.2
Somewhat too little	26	25	10	21.6
Just Right	27	19	25	24.3
Somewhat too much	15	30	22	21.6
Too much	18	13	30	17.2
Thin/Thick				
Product vs Ideal	Chi-sq=4.94 *p<0.294*	Chi-sq=10.99 ***p<0.027***	Chi-sq=26.29 ***p<0.00002***	
	170	**896**	**914**	**Ideal**
Too little	0	1	0	0.1
Somewhat too little	13	8	4	7.7
Just Right	84	90	77	86.5
Somewhat too much	5	3	21	7.7
Too much	0	0	0	0.1
Stickiness				
Product vs Ideal	Chi-sq=13.00 ***p<0.011***	Chi-sq=11.00 ***p<0.027***	Chi-sq=7.19 *p<0.126*	
	170[a]	**896**[a]	**914**[a]	**Ideal**
Too little	0	0	2	0.4
Somewhat too little	4	1	8	7.4
Just Right	81	87	82	86.2
Somewhat too much	16	13	9	7.4
Too much	1	1	0	0.4

[a]Unequal sums of observed & expected frequencies. Significant *p*-values are in **bold**.

product from those that only have a similar mean. One drawback of this technique is that Thurstonian ideal point modeling cannot be conducted by hand or with the use of a simple spreadsheet. Another is the somewhat difficult framework involved in presenting this approach.

Recommendation

This method is recommended for JAR scale analysis whenever one wishes to compare a product(s) to the ideal product for targeted consumer segments. It is especially effective when there is extensive historical JAR data on a product and consumer segment(s) that can be used to determine the ideal distribution.

References

[1] O'Mahony, M., *Sensory Evaluation of Food*, Marcel Dekker, Inc., New York, 1986.

Appendix L: Penalty Analysis or Mean Drop Analysis

Mary Schraidt[1]

Introduction and Objectives

Penalty analysis or Mean drop analysis is a method for determining if respondents' "Just About Right" ratings for a specific attribute are associated with a drop in some hedonic or choice measure, most commonly Overall Liking. Penalty/Mean drop analysis is conducted on "Just About Right" data to determine if those who do not find a particular attribute "Just About Right" rate it lower for Overall Liking on that attribute than those who find the same attribute JAR. Penalty/Mean drop analysis is not a formal method for determining drivers of liking, but is an effective tool for linking attribute performance to Overall Liking.

Requirements

In order to conduct penalty analysis, the respondent's individual Overall Liking rating and ratings on the JAR attributes of interest are required; the analysis is typically performed for each product × attribute combination. Typically the data are collapsed into three categories "Too High," "Just About Right," and "Too Low," irrespective of the number of scale points.

"How to"

The following example illustrates the use of penalty analysis of a single JAR rating on Overall Liking. First, ratings are grouped into "above JAR," "at JAR," and "below JAR." Then the mean Overall Liking rating is calculated for each group. The following table presents hypothetical results:

Overall Flavor Strength	Percentage of Respondents	Overall Liking Mean
"Too Weak"	21%	6.0
"Just About Right"	55%	7.6
"Too Strong"	24%	4.8

The penalties (mean drops) are calculated as the differences between the mean liking of each non-JAR group and the mean of the JAR group.

"Too Weak"	6.0	"Too Strong"	4.8
"Just About Right"	7.6	"Just About Right"	7.6
Drop	−1.6	Drop	−2.8

These values (−1.6/−2.8) are plotted versus the percentage giving each response (21 % and 24 %, respectively). Note that for the "Overall Liking Mean" in the above table, it is recommended to use the Overall Liking mean of those respondents that rated the attribute "Just About Right" and not the Overall Liking sample mean. Using the overall liking sample mean would result in "double counting" the impact on some of the respondents. (See Fig. 1.)

A minimum percentage skew for "Not Just Right" is often employed as a means of eliminating smaller, less impactful attributes from consideration. This cutoff may depend on the consumer base size, but is typically around 20 %. When the base size is larger, percentages less than 20 % may well be reliable and can be plotted. Some researchers suggest a minimum criterion for the overall base size × percentage of the skew as 20.

The following is a guideline for interpreting the magnitude of a particular penalty for Overall Liking.

Attributes which a large percentage of consumers are critical of and which have large penalties can be found in the upper right quadrant of a plot, providing a quick summary of the most concerning diagnostic problems for that product.

Examples

Sample 170: This sample receives a very slightly concerning penalty for Too Much Flavor.

Implication—slightly reduce flavor. The too large skew was a very slightly concerning penalty as it received a −0.02.

Although there is a skew for being Too Small, the penalty analysis shows this imbalance is positive. (See Fig. 2.)

Sample 458: This sample receives slightly concerning penalties for Too Much Flavor and Too Sticky.

Implication—slightly reduce flavor and stickiness.

Although there are skews for being Too Large and Too Small, the penalty analysis shows these imbalances are positive. (See Fig. 3.)

Sample 914: This sample receives concerning penalties for Too Thick and Not Enough Flavor, and a very slightly concerning penalty for Not Enough Color.

Implication—reduce thickness and increase flavor.

Although there are skews for being Too Large and Too Small, the penalty analysis shows these imbalances are positive. (See Fig. 4.)

Sample 896: This sample has no concerning penalties. (see Fig. 5.)

Implication—no further refinement for this product.

Sample 523: This sample receives a slightly concerning penalty for Too Much Flavor. (see Fig. 6.)

Implication—slightly reduce flavor.

Although there are skews for being Too Large and Too Small, the penalty analysis shows these imbalances are positive.

Additions to Penalty Analysis

The total penalties may also be included along with penalty analysis. This involves multiplying the percent skew by the penalty for each JAR attribute. A simple ranking of these total penalties may help the researcher prioritize which attributes to consider adjusting.

[1] *Peryam and Kroll, 6323 N. Avondale, Chicago, IL.*

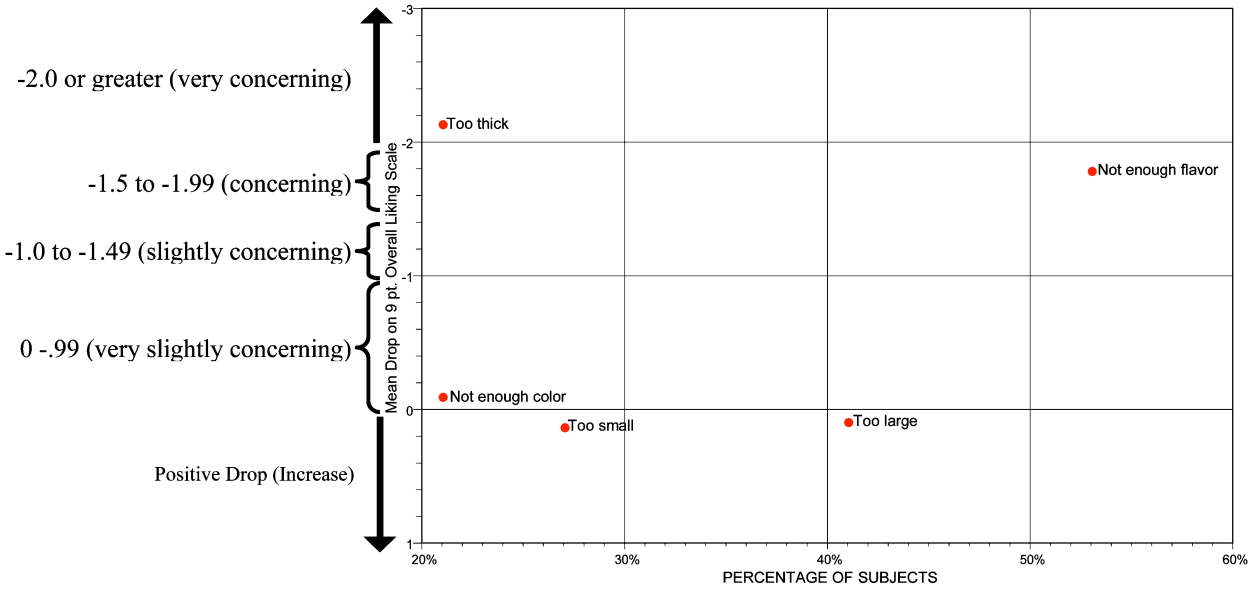

Fig. 1—Penalty plot with negative and positive penalties.

Pros and Cons

Benefits of penalty analysis include easily interpretable data or graphs that link specific product attributes in need of adjustment with the impact their being not "just right." Penalty analysis also separates attributes into those that appear to have impacted Overall Liking from those that have generated "complaints," those attributes that consumers say are not just right, but whose current level has in reality, not impacted liking.

Caveats associated with penalty analysis include unclear action to be taken in the case of equal bimodal data and associated penalties (such as equal penalties for "opposite" findings of "too salty" and "not salty enough"). Penalty analysis does not provide the level of adjustment that needs to be undertaken to correct an attribute, thus the guidance is approximate. Another caveat associated with penalty analysis is that the penalties ignore the potential impact on future attribute adjustment among respondents originally rating the product "just right" for the specified attributes. Finally, although the penalties and subsequent product improvement

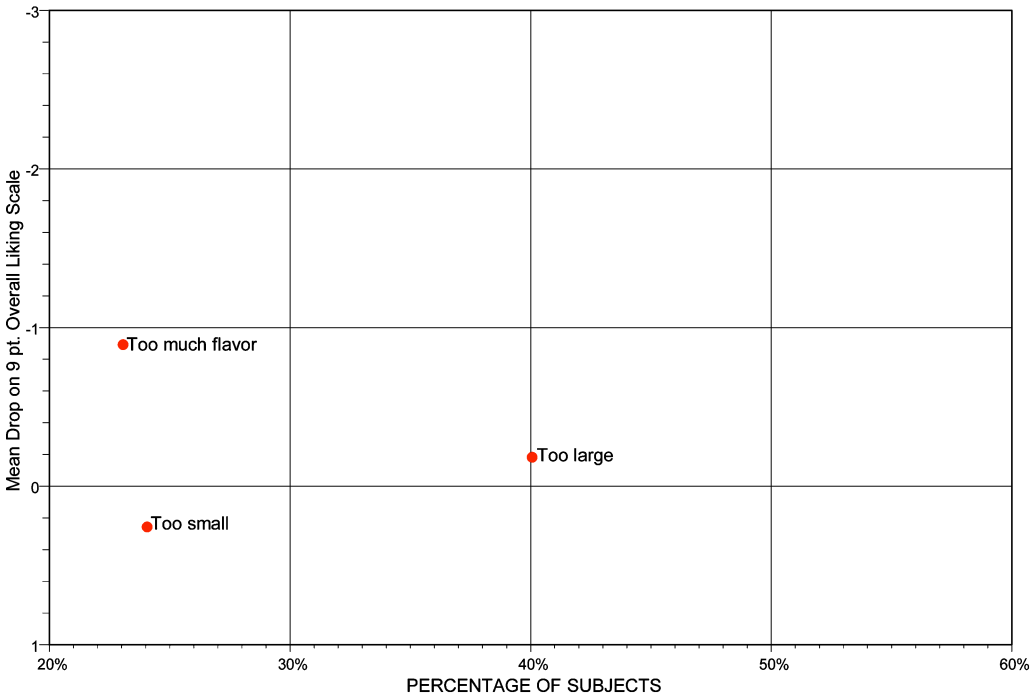

Fig. 2—ASTM mean drop analysis-total. #170.

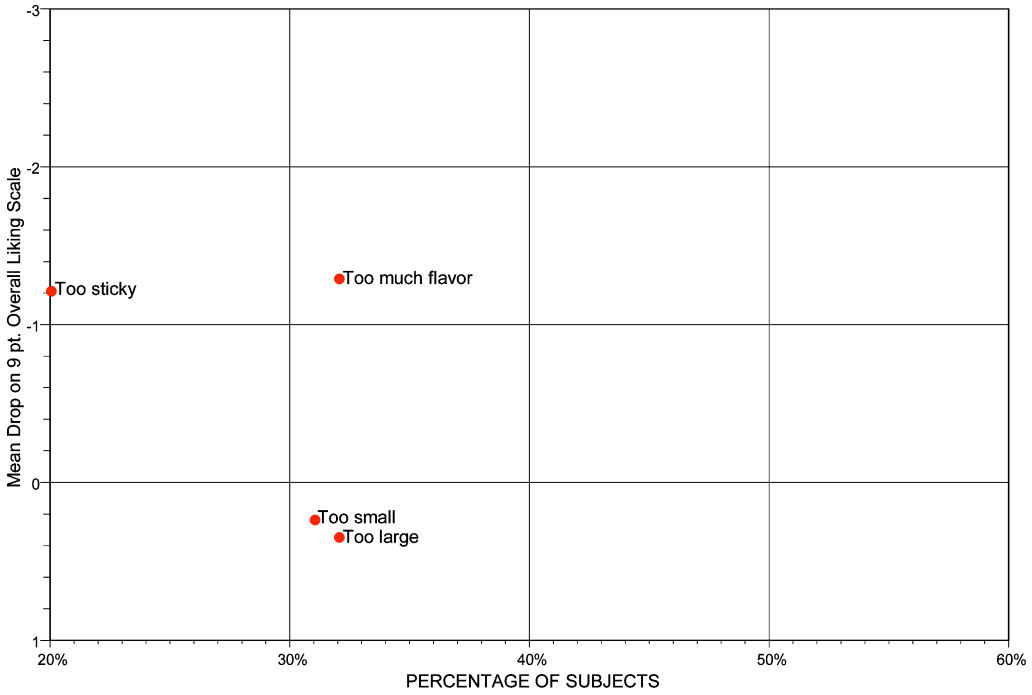

Fig. 3—ASTM mean drop analysis-total. #458.

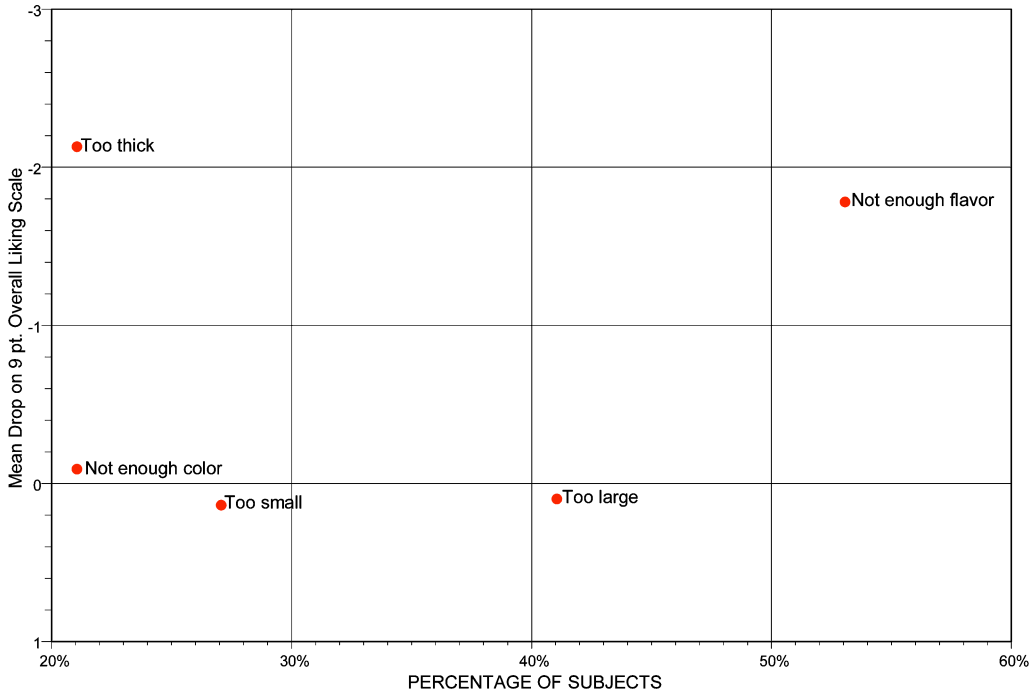

Fig. 4—ASTM mean drop analysis-total. #914.

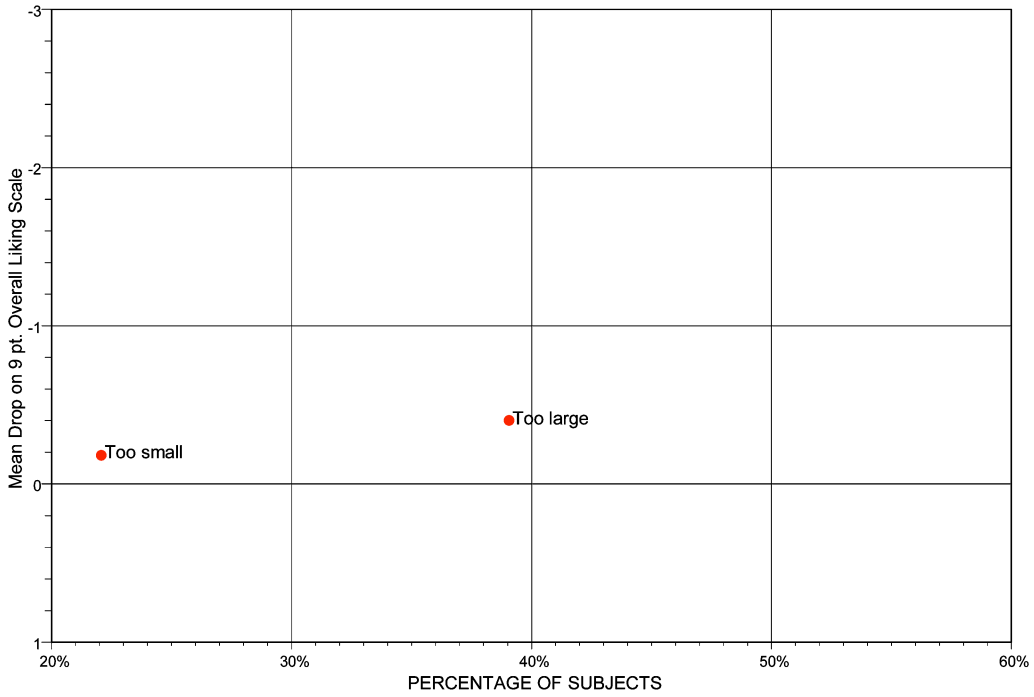

Fig. 5—ASTM mean drop analysis-total. #896.

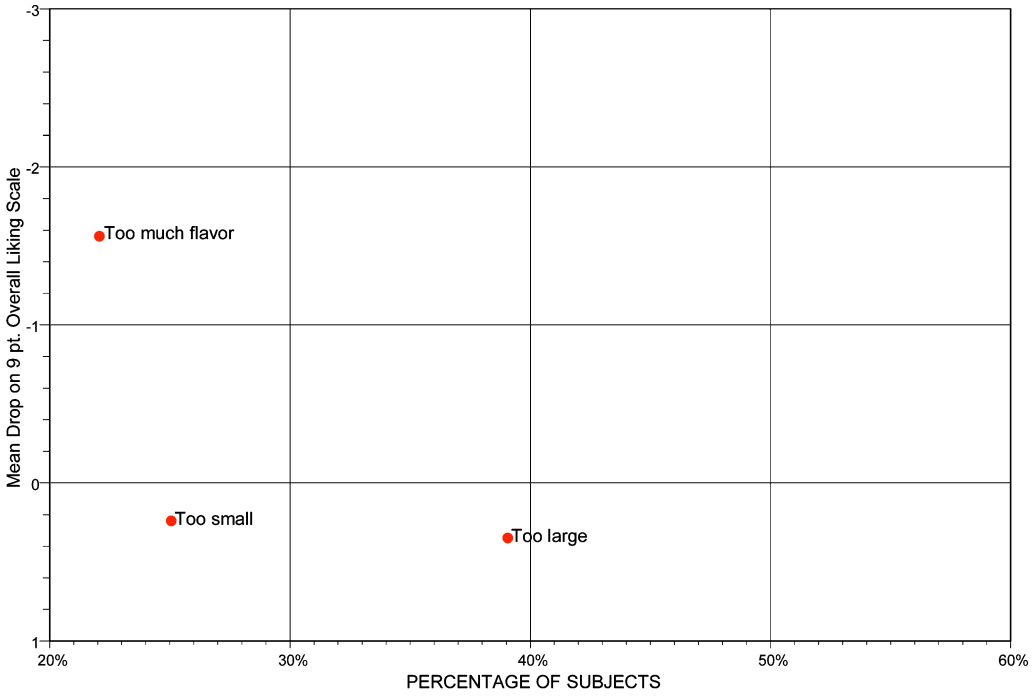

Fig. 6—ASTM mean drop analysis-total. #523.

recommendations are considered individually, product attributes may not change in isolation; altering the level of some attributes may change the perception of other attributes not under consideration.

Recommendation

Penalty analysis is recommended when the researcher wants to understand which attribute skews were associated with lower Overall Liking and in what direction to adjust them.

Appendix M: Using Grand Mean versus Mean of the Proportion of Respondents Who Scored the Product JAR

Dave Plaehn,[1] Gregory Stucky, David Lundahl, and John Horne

Introduction

The object of traditional penalty analysis is to try to determine the effect "Too Much" or "Too Little" of a product attribute has on overall product liking. These product attributes have been called "Just-About-Right" or JAR variables, they are discrete and the middle value corresponds to a consumer response of "Just About Right." There are two different approaches in determining the so-called penalties. Both are examined here.

Conclusion

Calculate the penalty from the JAR subgroup rather than the Grand Mean. There is a mathematical evidence that shows that calculating from the Grand Mean may cause the researcher to make erroneous conclusions in some situations.

Reasons

The objective of penalty analysis is to identify those product attributes that are contributing to a lower product liking score (or purchase intent score). The critical assumption of the JAR scale is that respondents should score the attribute JAR when the attribute is at a point where improving it won't improve the product liking score. With that assumption in place the liking mean score for the respondents who scored JAR will be higher than the liking score for the respondents who do not score the product JAR. Thus it is expected that the mean score of the respondents in JAR will be higher than the mean score of all respondents combined (Grand Mean).

Penalty analysis results show the researcher the relative amount that the product score is being reduced by the respondents who think an attribute is not just about right. Therefore if you calculate the penalty based on the grand mean you will be using a target that is lower than the actual product potential. Additionally when there is high skewness in mean scores, it is possible that the grand mean will be lower than the mean of one of the two subgroups (scale ends). In these cases using the grand mean will show the researcher a positive penalty score which would lead them to an erroneous conclusion.

Figure 1 shows when using the grand mean when the properties of respondents are in a typical distribution, that it may be likely to conclude that having more respondents in "too low" would improve the product score. Although it would improve the grand mean, it would only improve it to the mean of the subgroup. At that point a calculation of the weighted penalty would begin to show negative penalty. Thus you would end up optimizing towards the mean of the "too low" subgroup rather than the mean of the JAR subgroup, thus not adhering to the assumptions of the scale.

In most cases where the scale is being used "normally" by consumers, the two methods will give extremely similar results. However, as more skewness in subgroup means occur and the Grand Mean scores become increasingly lower than the JAR mean, the Grand Mean method has a greater potential to show results that would make the researcher draw an erroneous conclusion.

In those situations then where the Grand Mean is higher than the JAR mean, calculating the penalty from the grand mean may be a viable option, however in these cases there is clear evidence that the JAR scale is being misinterpreted or misused by the respondents (see section on caveats). Thus in these situations, although statistically one could say the grand mean is a more appropriate option, from a psychology and scale usage standpoint, the validity of the data for "typical" interpretation is very low.

Means of Subgroups			Proportions of Respondents			Grand Mean		Weighted Penalty for Grand Mean Method			Weighted Penalty for JAR Mean Method	
lo	jar	hi	lo	jar	hi			lo	hi		lo	hi
4.0	6.0	2.0	0.25	0.25	0.50	3.50		0.500	-3.000		-0.500	-2.000
4.0	6.0	2.0	0.40	0.20	0.40	3.60		0.800	-3.200		-0.800	-1.600
4.0	6.0	2.0	0.17	0.33	0.50	3.67		0.167	-2.500		-0.333	-2.000
4.0	6.0	2.0	0.50	0.17	0.33	3.67		1.000	-3.333		-1.000	-1.333
4.0	6.0	2.0	0.33	0.33	0.33	4.00		0.000	-2.000		-0.667	-1.333
4.0	6.0	2.0	0.40	0.40	0.20	4.40		-0.400	-1.200		-0.800	-0.800
4.0	6.0	2.0	0.25	0.50	0.25	4.50		-0.250	-1.250		-0.500	-1.000
4.0	6.0	2.0	0.33	0.50	0.17	4.67		-0.444	-0.889		-0.667	-0.667

Fig. 1—Most common means distributions with possible proportion distributions. Comparison of Grand Mean and JAR mean weighted penalties.

[1] *InsightsNow, Inc., Corvallis, OR 97333.*

Mathematics

This section details the exact mathematics to allow those who want to conduct a detailed review of their methods.

Let N be the total number of respondents, $\mathbf{Y}$ be a $N \times 1$ respondent "Liking" vector, and $\mathbf{X}$ be a $N \times 1$ vector of JAR responses. Assume $\mathbf{X}$ has c categories, $1, 2, \ldots, c$, where c is odd, and the middle (JAR) level $(c/2+1)$ is the "Just-About-Right" level. Let $\mathbf{n}_i$ be the number of occurrences of the response i in $\mathbf{X}$. Thus, $N = \Sigma_i \mathbf{n}_i$. Let n_{lo} be the number of people giving the JAR response below the JAR level, n_{hi} be the number of people giving the JAR response above the JAR level and $n_{\text{JAR}} = \mathbf{n}_{(c+1)/2}$. Then, $n_{\text{lo}} = \Sigma_{1 \leq i < (c+1)/2} \mathbf{n}_i$ and $n_{\text{hi}} = \Sigma_{(c+1)/2 < i \leq c} \mathbf{n}_i$. Let μ be the mean of $\mathbf{Y}$, the so-called "grand mean." Let $\boldsymbol{\beta}_i$ be the mean liking for those respondents having a JAR response i ($\mathbf{X} = i$) for $i = 1, 2, \ldots, c$. In a similar manner, define β_{lo}, β_{hi}, and β_{JAR}. Then note that the grand mean, μ, is a weighted average of sub-means. Specifically,

$$\mu = \frac{1}{N} \sum_{i=1}^{c} \mathbf{n}_i * \boldsymbol{\beta}_i \tag{1a}$$

and

$$\mu = \frac{1}{N} (n_{\text{lo}} * \beta_{\text{lo}} + n_{\text{JAR}} * \beta_{\text{JAR}} + n_{\text{hi}} * \beta_{\text{hi}}) \tag{1b}$$

Let $\mathbf{p}_i = \mathbf{n}_i / N$, for $i = 1, 2, \ldots, c$, and $p_{\text{lo}} = n_{\text{lo}}/N$, $p_{\text{hi}} = n_{\text{hi}}/N$ *and* $p_{\text{JAR}} = n_{\text{JAR}}/N$. Let $\mathbf{p}$ be the vector whose elements are $\mathbf{p}_i$ and $\boldsymbol{\beta}$ be the vector whose elements are $\boldsymbol{\beta}_i$, then the above equations can be rewritten as

$$\mu = \sum_{i=1}^{c} \mathbf{p}_i * \boldsymbol{\beta}_i = \mathbf{p}' \cdot \boldsymbol{\beta} \tag{2a}$$

(where $\cdot$ stands for vector dot product or matrix multiplication) and

$$\mu = p_{\text{lo}} * \beta_{\text{lo}} + p_{\text{JAR}} * \beta_{\text{JAR}} + p_{\text{hi}} * \beta_{\text{hi}} \tag{2b}$$

There are two different ways of determining penalties. In one method the penalties are calculated by subtracting the grand mean (μ) from the "group" means. In the other approach the JAR mean (η_{JAR}) is subtracted from the group means. Let $\mathbf{penGrand}_i = \boldsymbol{\beta}_i - \mu$ and $\mathbf{penJAR}_i = \boldsymbol{\beta}_i - \eta_{\text{JAR}}$, for all i, be the respective types of penalties for the two approaches associated with the $\mathbf{X}$ response i. Similarly, define $penGrand_{\text{lo}}$, $penGrand_{\text{JAR}}$, $penGrand_{\text{hi}}$, $penJAR_{\text{lo}}$, $penJAR_{\text{JAR}}$, and $penJAR_{\text{hi}}$. Let **penGrand** be the vector whose elements are $\mathbf{penGrand}_i$. Similarly, define **penJAR**. If μ is subtracted from the above equations, using the fact that $\Sigma_i \mathbf{p}_i = 1$ and $p_{\text{lo}} + p_{\text{JAR}} + p_{\text{hi}} = 1$, then

$$\mu - \mu = 0 = \left(\sum_{i=1}^{c} \mathbf{p}_i * \boldsymbol{\beta}_i \right) - \mu = \sum_{i=1}^{c} \mathbf{p}_i * (\boldsymbol{\beta}_i - \mu)$$

$$= \sum_{i=1}^{c} \mathbf{p}_i * \mathbf{penGrand}_i = \mathbf{p}' \cdot \mathbf{penGrand} \tag{3a}$$

and

$$\mu - \mu = 0 = p_{\text{lo}} * penGrand_{\text{lo}} + p_{\text{JAR}} * penGrand_{\text{JAR}} + p_{\text{hi}} * penGrand_{\text{hi}} \tag{3b}$$

On the other hand, subtracting η_{JAR} from both sides of Eqs (2*a*) and (2*b*), and noting that $penJAR_{\text{JAR}} = \mathbf{penJAR}_{(c+1)/2} = 0$, gives

$$\mu - \eta_{\text{JAR}} = \left(\sum_{i=1}^{c} \mathbf{p}_i * \boldsymbol{\eta}_i \right) - \eta_{\text{JAR}} = \sum_{i=1}^{c} \mathbf{p}_i * (\boldsymbol{\eta}_i - \eta_{\text{JAR}})$$

$$= \sum_{i=1}^{c} \mathbf{p}_i * \mathbf{penJAR}_i = \sum_{i \neq (c/2+1)} \mathbf{p}_i * \mathbf{penJAR}_i$$

$$= \mathbf{p}' \cdot \mathbf{penJAR} \tag{4a}$$

and

$$\mu - \eta_{\text{JAR}} = p_{\text{lo}} * penJAR_{\text{lo}} + p_{\text{JAR}} * penJAR_{\text{JAR}} + p_{\text{hi}} * penJAR_{\text{hi}}$$

$$= p_{\text{lo}} * penJAR_{\text{lo}} + p_{\text{hi}} * penJAR_{\text{hi}} \tag{4b}$$

Let the "weighted penalties" for the latter approach be defined as $\mathbf{pen\_wtJAR}_i = \mathbf{p}_i^* \mathbf{penJAR}_i$. Similarly define pen_wtJAR_{lo} and pen_wtJAR_{hi}. Then from Eqs (4*a*) and (4*b*),

$$\mu = \eta_{\text{JAR}} + \sum_{i \neq (c/2+1)} \mathbf{pen\_wtJAR}_i \tag{5a}$$

and

$$\mu = \eta_{\text{JAR}} + pen\_wtJAR_{\text{lo}} + pen\_wtJAR_{\text{hi}} \tag{5b}$$

To get equations similar to those of Eqs (5*a*) and (5*b*) for the case where the penalties are determined by subtracting the grand mean, it is necessary to define the weighted penalties as $\mathbf{pen\_wtGrand}_i = \mathbf{p}_i * \mathbf{penGrand}_i / p_{\text{JAR}}$. Similarly define $pen\_wtGrand_{\text{lo}}$ and $pen\_wtGrand_{\text{hi}}$. Then, from Eqs (3*a*) and (3*b*).

$$0 = \sum_{i=1}^{c} \mathbf{p}_i * \mathbf{penGrand}_i \Rightarrow - \mathbf{penGrand}_{(c/2+1)}$$

$$= \sum_{i \neq (c/2+1)} \mathbf{p}_i * \mathbf{penGrand}_i \Rightarrow \mathbf{p}_{(c/2+1)} * (\mu - \boldsymbol{\eta}_{(c/2+1)})$$

$$= \sum_{i \neq (c/2+1)} \mathbf{p}_i * \mathbf{penGrand}_i \Rightarrow p_{\text{JAR}} * (\mu - \eta_{\text{JAR}})$$

$$= \sum_{i \neq (c/2+1)} \mathbf{p}_i * \mathbf{penGrand}_i \Rightarrow \mu = \eta_{\text{JAR}}$$

$$+ \frac{1}{p_{\text{JAR}}} \sum_{i \neq (c/2+1)} \mathbf{p}_i * \mathbf{penGrand}_i \Rightarrow \mu = \eta_{\text{JAR}}$$

$$+ \sum_{i \neq (c/2+1)} \mathbf{pen\_wtGrand}_i \tag{6a}$$

Similarly,

$$\mu = \eta_{\text{JAR}} + (pen\_wtGrand_{\text{lo}} + pen\_wtGrand_{\text{hi}}) \tag{6b}$$

Comparing Eqs (6*a*) and (6*b*) with Eqs (5*a*) and (5*b*), it must be that

$$\sum_{i \neq (c/2+1)} \mathbf{pen\_wtJAR}_i = \sum_{i \neq (c/2+1)} \mathbf{pen\_wtGrand}_i \tag{7a}$$

and

$$pen\_wtJAR_{lo} + pen\_wtJAR_{hi} = pen\_wtGrand_{lo} + pen\_wtGrand_{hi} \quad (7b)$$

For Further Consideration

Based on the assumptions of JAR scale use, the following "rules" should be carefully considered.

If the % of respondents in a single tail is higher than the % of respondents in the JAR, a penalty analysis will not give "predictable" values. For example if 70 % of the respondents said a product was too hard, then the action a company would take would be to make it softer to such a substantial degree that the respondents who thought that the extremely hard product was just right, would be expected to greatly change their opinion.

If there is bimodal distribution of the % of respondents, there is clearly segmentation and it is possible that the product will not succeed unless altered to one extreme or the other.

If the mean scores are highly skewed or are bimodal the attribute is suspect that the interpretation and use of the scale should be brought into question. When this occurs penalties of zero or positive penalties will occur.

Appendix N: A Regression-Based Approach for Testing Significance of JAR Variable Penalties

Dave Plaehn[1] and John Horne[1]

Introduction and Objectives

Traditional penalty analysis attempts to relate JAR variable responses to overall liking or some other "reference" variable. No variance estimates are calculated around the penalties or mean drops, and as a result, significance testing is not done. Consequently, this method does not give any gauge of reliability or importance of the results. Focus is often placed on large penalties that are also associated with a large proportion of respondents; 20 % of respondents on a given side of a "Just About Right" point (e.g., "Too Much" or "Too Little") is frequently used as a minimum standard of importance.

We propose a regression-based approach to better understand which penalties are important for a given product. This approach recodes JAR variable scores into indicator (dummy) variables in order to address the non-linear nature of the typical JAR variable scale (the middle category as "Just About Right" and the other categories as some degree of "Too Much" or "Too Little"). Regression coefficients resulting from the indicator variables are analogous to mean drops of the reference variable. Significance testing can be done on these coefficients parametrically by using the standard error estimates from the regression model itself; semi-parametrically by using standard error estimates from methods such as jackknife and bootstrap; or non-parametrically, for example, by forming "confidence intervals" from the distribution of a large number of bootstrap samples. All of these methods are presented below. While some parts of this approach can be used to test multiple products and attributes simultaneously, we will consider only one JAR variable/product combination at a time. Along the way we will prove that the regression coefficients from ordinary least-squares (OLS) regression are identical to the traditional penalties assuming the so-called JAR mean is used to determine the penalties.

Requirements for the Analysis

The analysis requires the individual raw liking (or other reference variable) scores and the individual JAR variable scores for each variable/product combination. JAR variable scores must be transformed to indicator variables. A statistical package that implements regression models is required. A package that also implements cross-validation, jackknife, and bootstrap is useful.

"How to"

Notation

Let vectors and matrices be represented by bold lower and upper case letters, respectively. If $\mathbf{a}$ is a vector, denote the ith element of $\mathbf{a}$ by $\mathbf{a}_i$ or $\mathbf{a}(i)$. For a matrix, $\mathbf{A}$, let the element of the ith row and jth column be represented by $\mathbf{A}(i,j)$ or $\mathbf{A}_{ij}$. Let $\mathbf{A}(:,j)$ or $\mathbf{A}_j$ be the jth column of $\mathbf{A}$ and let $\mathbf{A}(i,:)$ be the ith row of $\mathbf{A}$. Assume all vectors are column vectors and that the inner or dot product of two vectors $\mathbf{a}$ and $\mathbf{b}$ is given by $\mathbf{a}'\mathbf{b}$, where "$'$" represents vector or matrix transpose, and it is assumed that $\mathbf{a}$ and $\mathbf{b}$ have the same number of elements. Similarly denote the regular matrix product by "adjacency." Let "$\equiv$" mean "is defined as."

JAR Variable Score Recoding

To create a "sensible" regression model between a JAR variable and a reference variable such as liking, the JAR variable must be somehow transformed. A simple approach is to change the JAR variable categories, or a combination of those categories, into dummy or indicator variables. Formally, let $\mathbf{y}$ be a column vector of the reference variable and let $\mathbf{x}$ be an associated vector of JAR variable responses. Assume $\mathbf{x}$ has c categories where c is odd and that the category $(c+1)/2$ is the "just about right" category. Transform $\mathbf{x}$ according to

$$\tilde{\mathbf{X}}(i,j) = 1 \Leftrightarrow \mathbf{x}(i) = j \quad \text{else } \tilde{\mathbf{X}}(i,j) = 0 \tag{1}$$

The columns of $\tilde{\mathbf{X}}$ are linearly dependent (if you know all but 1 of the columns of $\tilde{\mathbf{X}}$ you can calculate the remaining column) and, consequently, $\tilde{\mathbf{X}}$ cannot be used for ordinary least-squares regression (OLS). To remedy this, remove the "Just About Right" column from $\tilde{\mathbf{X}}$ and call the new array $\mathbf{X}$:

$$\mathbf{X}(:,j) = \tilde{\mathbf{X}}(:,j), \quad 1 \leqslant j < (c+1)/2$$

$$\mathbf{X}(:,j) = \tilde{\mathbf{X}}(:,j+1), \quad (c+1)/2 < j \leqslant c \tag{2}$$

The effect of removing the "Just About Right" column is that the regression model intercept becomes an estimate of the mean of the reference variable for those respondents giving the "Just About Right" response (the "JAR mean").

Make the Model

One can now create the regression model. The regression equation is given by

$$\mathbf{y} = \beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \beta_0 + \sum_{j=1}^{c-1} \mathbf{X}(:,j)\boldsymbol{\beta}_j + \boldsymbol{\varepsilon} \tag{3}$$

where β_0 is the model intercept, $\boldsymbol{\beta}$ is a vector of regression coefficients and $\boldsymbol{\varepsilon}$ is a vector of model errors. Taking the mean of Eq. (3) gives

[1] *InsightsNow, Inc., Corvallis, OR 97333.*

$$\mu_{\mathbf{y}} = \beta_0 + \sum_{j}^{c-1} \mathbf{p}_j \boldsymbol{\beta}_j \tag{4}$$

where $\mu_{\mathbf{y}}$ is the mean of $\mathbf{y}$ and $\mathbf{p}_j$ is the mean of $\mathbf{X}(:,j)$, the proportion of respondents giving the JAR variable response j or $j+1$ as the case may be (see Eq. (2)). It is further assumed that this is a "least squares" type of model and that, consequently, $\sum_i \boldsymbol{\varepsilon}_i = 0$. As mentioned above, β_0 is the model estimate of the so-called JAR mean, the reference variable mean of those respondents rating the product "Just About Right." The other coefficients, $\boldsymbol{\beta}_j$, are typically negative and are thought of as "penalties" (mean drops). In fact, an equation analogous to Eq. (4) holds for traditional or count-based penalty analysis. Let μ_j be the reference variable mean for those respondents giving the JAR variable response j and let $\tilde{\mathbf{p}}_j$ be the proportion of respondents giving that response. So,

$$\tilde{\mathbf{p}}_j = \text{mean}(\tilde{\mathbf{X}}(:,j)) \quad \text{and} \quad \mu_j = \mathbf{y}'^{*}\tilde{\mathbf{X}}(:,j)/\sum_i \tilde{\mathbf{X}}(i,j) \tag{5}$$

First note that $\mu_{\mathbf{y}}$ is just the weighted average of the μ_j, i.e.,

$$\mu_{\mathbf{y}} = \sum_{j=1}^{c} \tilde{\mathbf{p}}_j \mu_j \tag{6}$$

Let $\mu_{\text{JAR}} \equiv \mu_{(c+1)/2}$ and $\tilde{\boldsymbol{\beta}}_j \equiv \mu_j - \mu_{\text{JAR}}$, the "penalty" from the JAR mean. By subtracting μ_{JAR} from both sides of Eq. (6), making substitutions, etc., one can show

$$\mu_{\mathbf{y}} = \mu_{\text{JAR}} + \sum_{j \neq (c+1)/2} \tilde{\mathbf{p}}_j \tilde{\boldsymbol{\beta}}_j \tag{7}$$

Note that the elements of $\mathbf{p}$ and $\tilde{\mathbf{p}}$ are the same except $\mathbf{p}$ has no element corresponding to the "Just About Right" category (see Eq. (2)). If OLS is the regression approach then it can be shown that $\beta_0 = \mu_{\text{JAR}}$ and $\beta_j = \tilde{\beta}_j$ if $1 \leq j < (c+1)2$, $\beta_j = \tilde{\beta}_{j+1}$ if $(c+1)/2 < j \leq c$ (see Appendix).

Significance Testing for Regression Coefficients

Conducting a "model-based" penalty analysis as given above allows for significance testing of the regression coefficients and, consequently, gives a measure of the reliability and variability (via confidence intervals) of the results. Four different approaches to significance testing are provided below. They are divided into three groups: parametric, semi-parametric, and non-parametric. The hypothesis being tested in each case is whether or not a given coefficient differs from 0 (H_0).

Parametric Method: Standard Error Estimates from a Regression Model

If the errors $\boldsymbol{\varepsilon}_i$ are independent and normally distributed with variance σ^2, then the model coefficients are normally distributed and may be tested for statistical significance using the standard error estimate (STDERR(·)) from OLS regression and a t-test. These have the following form:

$$\text{STDERR}(\boldsymbol{\beta}_j) = \sqrt{\text{MSE}(\mathbf{X}'\mathbf{X})^{-1}_{jj}} \tag{8}$$

where MSE = mean squared error $= (1/N)\sum_i \boldsymbol{\varepsilon}_i^2$ is an estimate of σ^2 and the number of respondents is N and

$$t = \frac{\boldsymbol{\beta}_j}{\text{STDERR}(\boldsymbol{\beta}_j)}. \tag{9}$$

Because the columns of $\mathbf{X}$ are orthogonal ($\mathbf{X}_i'\mathbf{X}_j = 0$ for $i \neq j$) an "ANOVA-like" approach could also be taken since the model sums of squares can be partitioned among the dummy variables.

Semi-parametric Method 1: Standard Error Estimates from Jackknife

Another approach to calculating coefficient standard error is to use model cross-validation [1] in conjunction with the jackknife procedure. Cross-validation is a model validation technique that attempts to estimate how well a model will predict "new" samples. The data are partitioned into M segments of equal or nearly equal size according to some user-defined scheme. For each segment, a regression sub-model is calculated based on all the data except that of the given segment. Consequently, M sub-models are made providing M sets of regression coefficients.

The variance that these coefficients form around the applicable coefficient from the same model on the original dataset is used to form the standard errors. However, because the number of data rows in each sub-model is smaller than the total number of data rows, a correction factor is used that adjusts the variance up based on the number of cross-validation segments. The standard error estimate for $\boldsymbol{\beta}_j$ using this approach, assuming equal partition sizes [2], is given by

$$\text{STDERR}(\beta_j) = \sqrt{\frac{M-1}{M}\sum_{m=1}^{M} (\beta_{j(m)} - \beta_j)^2} \tag{10}$$

where: β_j is the coefficient for the jth indicator variable from the original data, $\beta_{j(m)}$ is the coefficient for the jth variable from the mth cross-validation sub-model, and M is the number of sub-models. Hypothesis testing for this approach uses the same t-statistic described in Eq. (9) above. Only the method for arriving at the standard error estimate differs.

Semi-parametric Method 2: Standard Error Estimates from Bootstrap

Still another approach for estimating standard errors around regression coefficients is bootstrap re-sampling. The bootstrap method re-samples the original data *with replacement*. All bootstrap samples will have the same size as the original data. By generating a large number of bootstrap samples, each observation is about as likely to contribute to the final variance as is each other observation.

The estimated standard error for the bootstrap approach is as follows:

$$\text{STDERR}(\boldsymbol{\beta}_j) = \sqrt{\frac{\sum_{b=1}^{B} (\boldsymbol{\beta}_{j(b)} - \boldsymbol{\beta}_j)^2}{B-1}} \tag{11}$$

where: β_j is the coefficient for the jth indicator variable from the original data, $\beta_{j(b)}$ is the coefficient for the jth variable from the bth bootstrap sample, and B is the number of bootstrap models or samples. Again, this is an alternative method for estimating standard errors and once we have those, we can proceed with hypothesis testing using the t-statistic in Eq. (9).

Nonparametric Method: Confidence Intervals from Bootstrap Sample Distributions

All of the above approaches have a parametric component. They rely on the assumption that the individual errors, $\boldsymbol{\varepsilon}_i$ from the model are independent and normally distributed. When this assumption is violated, non-parametric approaches should be substituted. An example of a non-parametric approach to significance testing of regression coefficients is the percentile bootstrap (see, for example, Wilcox, [3]). In this approach, confidence intervals around the coefficients from the regression model are determined by rank-ordering the bootstrap coefficients and finding the appropriate percentiles in these distributions. These intervals are bounded by the $B(\alpha/2)+1$ and the $B-B(\alpha/2)$ rank-ordered bootstrap regression coefficients. As with other confidence interval approaches, if this interval does not include zero, the coefficient is concluded to be significantly different from zero. Because this approach does not rely on any other distributions of random variables (e.g., t) to do hypothesis testing, it operates completely independent of any assumptions of normality.

Examples from Case Study Data

There were five products and five JAR variables in the case study data set. JAR variables were transformed to two indicator variables each, with "Too Little" (categories "1" and "2") represented in one of the indicators and "Too Much" (categories "4" and "5") represented in the other.

OLS was used as the regression approach on each JAR variable independently. Overall Liking was the reference variable. No respondents rated the color of Product #914 as "Too Dark." As a result, there was no variation in the indicator variable for this side of this JAR variable and only 49 regression coefficients were calculated from 25 (5 products $\times$ 5 JAR variables) separate models. Of these coefficients, 16 were associated with JAR variables where more than 20 % of respondents rated a particular product as having "Too Much" or "Too Little" of the respective attribute. Diagnostics from traditional penalty analysis and all four of the above-described methods are shown in Table 1 for these 16 coefficients. Identical conclusions were drawn from all four methods. Penalties associated with "Flavor Too Strong" in Products #458 and #523 along with "Flavor Too Weak" and "Too Much Thickness" in Product #914 were consistently significant across methods ($p<0.05$, boldface in Table 1). "Flavor Too Strong" in Product #170 was likewise significant at $p<0.1$ across all four methods (90 % bootstrap confidence interval –1.75 to –0.01).

Standard errors, for the three methods that utilized them, were likewise in similar ranges and followed nearly identical distributions across all 49 coefficients tested (Fig. 1). Most of the variance that did exist between the standard errors calculated by the various methods was found among those JAR attributes with very few respondents on one side or the other. The outlier standard error from the jackknife approach in Fig. 1(*A*) was associated with a single respondent rating Product #896 as too light in color. There was no evidence from these analyses that one or another of these methods leads to systematically larger or smaller standard errors.

Conclusions

Benefits and Risks of the Analysis

The model-based approach presented here can be seen as a natural extension of traditional penalty analysis. As the case study shows, the penalties and penalty-weights are identical when OLS is used as the regression approach (see Appendix). The benefit of the model-based approach is that it provides the analyst with significance testing and confidence intervals for the penalties.

A further benefit of the dummy variable approach in combination with analyzing one JAR-variable-product combination at a time is that the columns of the dummy variable array ($\mathbf{X}$) are orthogonal. Consequently, there are no issues of collinearity and ill-conditioning. Thus, one need not be concerned with using OLS regression, which is commonly available in statistical packages.

Each of the four methods used to test the significance of JAR variable penalties has some associated benefits and risks in their own right. The parametric approach is the simplest to use from a computational standpoint, but is not appropriate if the individual errors from the model are not normally distributed. The jackknife and bootstrap approaches are computationally similar, although the bootstrap may require more computing resources as it generates more samples. A possible benefit of the bootstrap approach is that because of the larger number of samples, the coefficients may be more likely to follow normal distributions than the smaller number of coefficients generated from the jackknife approach. Further, when full cross-validation is used in the jackknife approach, the likelihood that a single sub-sample will be replicated multiple times is quite high. This leads to a more non-continuous distribution of coefficients. Both of these characteristics can be seen in Fig. 2. Both sets of coefficients do not differ significantly from normal, but the bootstrap coefficients conform to the normal distribution better than the much smaller number of jackknife coefficients. The jackknife coefficients also follow a non-continuous distribution. While there are 102 coefficients generated in the example shown in Fig. 2, there are only 15 *unique* coefficients (i.e., only 15 *unique* sub-samples were generated from the leave-one-out cross-validation approach). The problem of having a non-continuous distribution of jackknife coefficients can be remedied by increasing the size of the cross-validation segments (i.e., leave-d-out, where $d>1$). However, increasing the size of the cross-validation segments also reduces the number of sub-models, and consequently reduces the number of coefficients that the jackknifed estimates of standard errors are based upon.

When neither the jackknife nor the bootstrap coefficients follow a normal distribution, the non-parametric approach should be used as it avoids any distributional assumptions. The percentile bootstrap approach presented here is but a single example of the nonparametric ap-

TABLE 1—Diagnostics from model-based penalty analysis on case study data.

	Traditional Penalty			Model-based (OLS)					5% Boot confidence	
Attribute	% resp	JAR Mean	Mean Drop	β_0	β_j	Normal deviate SE (p-value)	Jackknife SE (p-value)	Bootstrap SE (p-value)	LL	UL
Product 170										
Too Small Size	23.5%	5.541	0.251	5.541	0.251	0.580 (0.666)	0.579 (0.665)	0.571 (0.661)	−0.859	1.374
Too Large Size	40.2%	5.541	−0.199	5.541	−0.199	0.502 0.692	0.516 (0.700)	0.509 (0.696)	−1.198	0.810
Too Strong Flavor	22.5%	5.952	−0.909	5.952	−0.909	0.522 (0.085)	0.513 (0.079)	0.504 (0.074)	−1.910	0.067
Product 458										
Too Small Size	31.4%	5.297	0.234	5.297	0.234	0.421 (0.580)	0.443 (0.599)	0.439 (0.595)	−0.647	1.076
Too Large Size	32.4%	5.297	0.339	5.297	0.339	0.418 (0.419)	0.402 (0.401)	0.394 (0.391)	−0.442	1.106
Too Strong Flavor	32.4%	5.984	−1.348	5.984	−1.348	**0.349 (<0.001)**	**0.362 (<0.001)**	**0.347 (<0.001)**	**−2.028**	**−0.668**
Product 523										
Too Small Size	24.5%	5.811	0.229	5.811	0.229	0.458 (0.618)	0.487 (0.639)	0.476 (0.631)	−0.710	1.151
Too Large Size	39.2%	5.811	0.339	5.811	0.339	0.403 (0.403)	0.393 (0.390)	0.384 (0.379)	−0.415	1.097
Too Strong Flavor	21.6%	6.333	−1.561	6.333	−1.561	**0.400 (<0.001)**	**0.402 (<0.001)**	**0.395 (<0.001)**	**−2.335**	**−0.786**
Product 896										
Too Small Size	21.6%	6.500	−0.182	6.500	−0.182	0.512 (0.723)	0.469 (0.699)	0.457 (0.692)	−1.082	0.718
Too Large Size	39.2%	6.500	−0.400	6.500	−0.400	0.431 (0.356)	0.446 (0.372)	0.441 (0.367)	−1.267	0.459
Product 914										
Too Small Size	27.5%	6.438	0.134	6.438	0.134	0.546 (0.807)	0.548 (0.807)	0.543 (0.806)	−0.927	1.219
Too Large Size	41.2%	6.438	0.086	6.438	0.086	0.495 (0.862)	0.525 (0.870)	0.520 (0.869)	−0.922	1.111
Too Light Color	20.6%	6.531	−0.102	6.531	−0.102	0.514 (0.843)	0.567 (0.857)	0.556 (0.854)	−1.247	0.931
Too Weak Flavor	52.9%	7.556	−1.796	7.556	−1.796	**0.378 (<0.001)**	**0.367 (<0.001)**	**0.364 (<0.001)**	**−2.491**	**−1.082**
Too Much Thickness	20.8%	7.091	−2.139	7.091	−2.139	**0.448 (<0.001)**	**0.433 (<0.001)**	**0.422 (<0.001)**	**−2.964**	**−1.314**

Key

Traditional Penalty	%resp (percentage of respondents rating a product on a given side of JAR scale) JAR Mean (mean OAL of respondents who rated a product as JAR on a given attribute) Mean Drop (difference between JAR mean and mean OAL of respondents who related a product on a given side of JAR scale)
Model-based OLS	β_0 (intercept, analogous to JAR mean); β_j (tested coefficient (analogous to mean drop)
Normal deviate	SE (standard error estimate from OLS model); p-value (based on two tailed t-test, approx. 100 degrees of freedom)
Jackknife	SE (standard error estimate from jackknife with full crossvalidation); p-value (based on two-tailed t-test, approx. 100 degrees of freedom)
Booststrap	SE (standard error estimate from bootstrap, B=10,000); p-value (based on two-tailed t-test, approx. 100 degrees of freedom)
5% boot confidence	LL (lower bound of 95 % CI, associated with 251st rank ordered bootstrap sample) UL (upper bound of 95 % CI, associated with 9,750th rank ordered bootstrap sample)

proaches available. Other approaches are described by MacKinnon [4] and Wilcox [3].

How Many Indicator Variables?

Some questions remain regarding the number of indicator variables (J) that should be formed from a single JAR variable. As described above, when a JAR variable has c categories, J has a maximum of $c-1$. Overall model error will often be lower (i.e., the model will have higher predictive ability) as J approaches $c-1$. However, there may be some benefits to interpretation if fewer indicator variables are used. "Traditional" penalty analysis divides "non-just-About-Right" responses into two categories. If the incidence in either category is less than a certain threshold (often 20 % of respondents), the penalty associated with that category is deemed unimportant. Similar conclusions can be drawn, from this regression-based approach if each JAR variable is transformed into two indicators, rather than four or more. Additionally, indicator variables within JAR variables and respondents must be mutually exclusive from one another. The more indicator variables a single JAR variable is transformed into, the more sparse the data and the greater the opportunity to conclude that potentially important attributes are not statistically significant. Therefore, even though more indicators are better if the overall model error is the only consideration, fewer indicators may be better from an interpretative standpoint.

Limits of the Analysis

With the exception of parametric approach, there are no easy-to-use mechanisms to conduct the analytic methods described here. The analyst must either have access to advanced statistical or mathematical software and know how to use it to produce the appropriate jackknife or bootstrap estimates, or have access to a programmer versed in these methods. Additionally, if there are many products and/or JAR variables, the method could be time consuming, unless a program was made to "loop" through the various combinations.

Lastly, the method considers only one JAR variable per product at a time, as opposed to "in concert." Relative importance of JAR variables can thus only be assessed indirectly.

Appendix: Proof of the Equivalence of Traditional Penalties (Relative to the JAR Mean) and OLS Regression Coefficients

As stated above, it can be shown that Eqs. (4) and (7) are equivalent when OLS is the regression method. Let $\mathbf{X}$ and $\mathbf{y}$

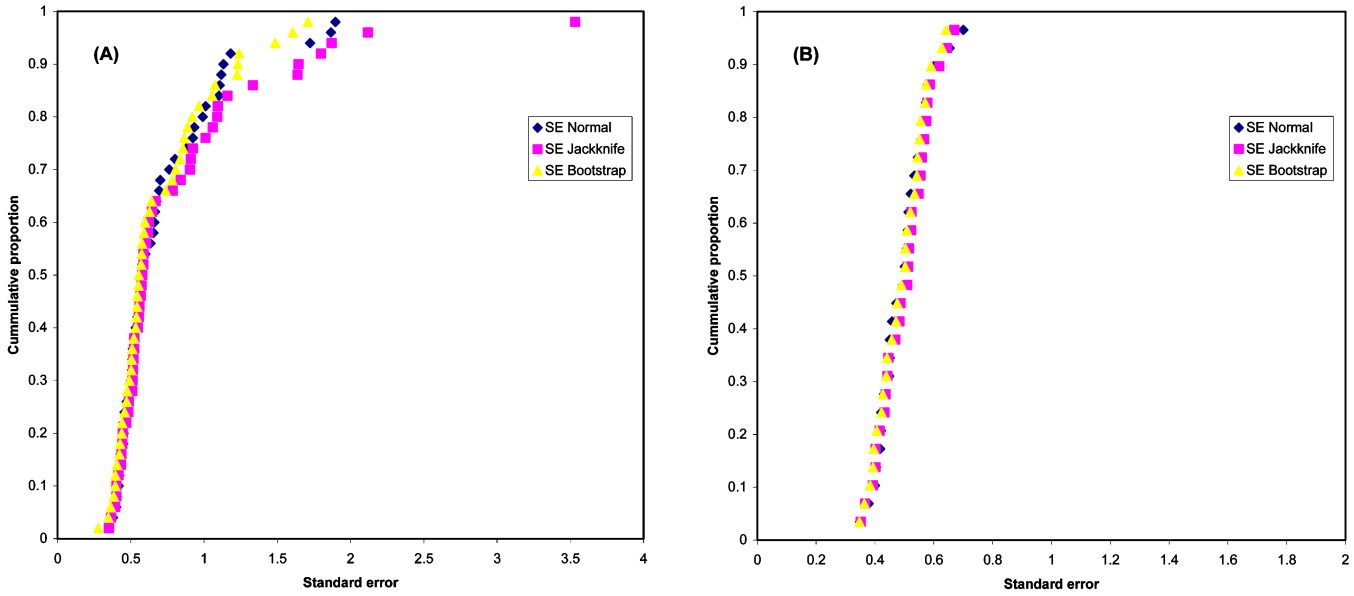

Fig. 1—Cumulative distributions of standard errors calculated by the standard normal deviate jackknife and bootstrapping methods. (A) Cumulative distributions of all 49 standard errors from the analysis of 5 products and 5 JAR variables; (B) Cumulative distributions of 28 standard errors where the percentage of respondents exceeded 10% on one side of a JAR variable.

be as above. It is customary in OLS to add a column of ones to the regressor matrix so as to "capture" the model intercept. So let $\mathbf{X}_1 = [\mathbf{1}_N \mathrm{X}]$ where is a $N \times 1$ column vector of 1 and N is the number of respondents. Then the OLS model coefficients are given by

$$\begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} \tag{12}$$

Let $n_j = \Sigma_{i=1}^{N} \tilde{\mathbf{X}}(i,j)$ be the number of respondents giving the JAR variable response j and let $n_{(c+1)/2} = n_{\mathrm{JAR}}$, where c is the number of JAR variable categories. It follows then that $N = \Sigma_{j=1}^{c} n_j$. Breaking Eq. (12) into "piece" one can show for $c = 5$ (5-point JAR variable scale):

$$\mathbf{X}_1'\mathbf{X}_1 = \begin{bmatrix} N & n_1 & n_2 & n_4 & n_5 \\ n_1 & n_1 & 0 & 0 & 0 \\ n_2 & 0 & n_2 & 0 & 0 \\ n_4 & 0 & 0 & n_4 & 0 \\ n_5 & 0 & 0 & 0 & n_5 \end{bmatrix} \quad \text{and} \quad \mathbf{X}_1'\mathbf{y} = \begin{bmatrix} N\mu_{\mathbf{y}} \\ n_1\mu_1 \\ n_2\mu_2 \\ n_4\mu_4 \\ n_5\mu_5 \end{bmatrix} \tag{13}$$

In general, $\mathbf{X}_1'\mathbf{X}_1$ is symmetric, with the first column and the diagonal both equal to

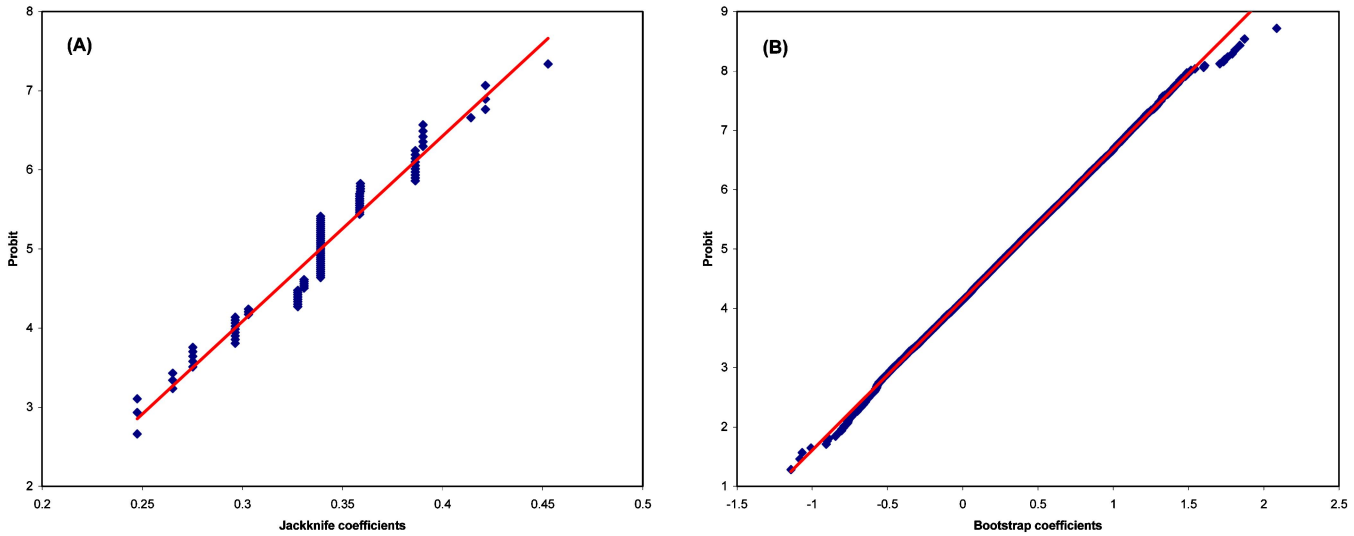

Fig. 2—Normal probability plots of coefficients for a selected indicator variable in the case study data. (A) coefficients from jackknife samples (M-102); (B) coefficients from bootstrap samples (B=10,000).

$$[N\ n_1\ n_2\ \cdots n_{(c+1)/2-1}\ n_{(c+1)/2+1}\ \cdots\ n_c]'$$

and with the remaining entries (excepting, of course, the first row) being 0, due to the orthogonality of $\mathbf{X}$. $\mathbf{X}_1'\mathbf{y}$ extends, similarly. It turns out then that

$$(X_1'X_1)^{-1} = \frac{1}{n_{\text{JAR}}}\begin{bmatrix} 1 & -1 & -1 & -1 & -1 \\ -1 & \frac{n_1+n_{\text{JAR}}}{n_1} & 1 & 1 & 1 \\ -1 & 1 & \frac{n_2+n_{\text{JAR}}}{n_2} & 1 & 1 \\ -1 & 1 & 1 & \frac{n_4+n_{\text{JAR}}}{n_4} & 1 \\ -1 & 1 & 1 & 1 & \frac{n_5+n_{\text{JAR}}}{n_5} \end{bmatrix} \tag{14}$$

In general, $(\mathbf{X}_1'\mathbf{X}_1)^{-1}$ is symmetric and, setting $\mathbf{A} \equiv (\mathbf{X}_1'\mathbf{X}_1)^{-1}$

$$\mathbf{A}_{i1} = \frac{-1}{n_{\text{JAR}}} \quad \text{for } 1 < i \leq c$$

$$\mathbf{A}_{11} = \mathbf{A}_{ij} = \frac{1}{n_{\text{JAR}}} \quad \text{for } 1 < i,j < \frac{c+1}{2}$$

$$A_{jj} = \frac{N - \Sigma_{i\neq j,(c+1)/2} n_i}{n_j} = \frac{n_j + n_{\text{JAR}}}{n_j} \quad \text{for } 1 < j \leq c, j \neq \frac{c+1}{2} \tag{15}$$

Calculating the coefficients, then

$$\begin{aligned} \beta_0 = \mathbf{A}(1,:)\mathbf{X}_1'\mathbf{y} &= \frac{1}{n_{\text{JAR}}}\left(N\mu_{\mathbf{y}} - \sum_{i\neq(c+1)/2} n_i\mu_i\right) \\ &= \frac{1}{n_{\text{JAR}}}\left(N\frac{\Sigma_i n_i\mu_i}{N} - \sum_{i\neq(c+1)/2} n_i\mu_i\right) \\ &= \mu_{\text{JAR}} \end{aligned} \tag{16}$$

and for $1 \leq j < (c+1)/2$

$$\begin{aligned} \boldsymbol{\beta}_j = \mathbf{A}(j,:)\mathbf{X}_1'\mathbf{y} &= \frac{1}{n_{\text{JAR}}}\left(-N\mu_{\mathbf{y}} + \sum_{i\neq j,(c+1)/2} n_i\mu_i \right. \\ &\qquad \left. + n_j\mu_j\left(\frac{n_j+n_{\text{JAR}}}{n_j}\right)\right) \\ &= \frac{1}{n_{\text{JAR}}}(-n_{\text{JAR}}\mu_{\text{JAR}} - n_j\mu_j \\ &\qquad + \mu_j(n_{\text{JAR}} + n_j)) \\ &= \mu_j - \mu_{\text{JAR}}. \end{aligned} \tag{17}$$

Similarly, if $j > (c+1)/2$, then

$$\boldsymbol{\beta}_{j-1} = \mu_j - \mu_{\text{JAR}}. \tag{18}$$

References

[1] Stone, M., "Cross-validatory Choice and Assessment of Statistical Prediction," *J. Roy. Stat. Soc., B*, Vol. 36, pp. 111–133.

[2] Busing, F., Meijer, E., and Van der Leeden, R., "Delete-m Jackknife for Unequal m," *Stat. Comput.*, Vol. 9, 1999, pp. 3–8.

[3] Wilcox, R., *Applying Contemporary Statistical Techniques*, Academic Press, San Diego, 2003.

[4] MacKinnon, J. G., "Bootstrap Methods in Econometrics," *Econ. Rec.*, Vol. 82, 2006, pp. s2–s18.

Appendix O: Bootstrapping Penalty Analysis

Rui Xiong[1] and Jean-Francois Meullenet[1]

Introduction and Objectives

Bootstrapping penalty analysis is a method that allows you to perform statistical testing on the results of a penalty analysis, using a technique called bootstrapping to estimate variances and covariances. Recall that the purpose of penalty analysis is to identify attributes that appear to have a strong impact on Overall Liking, on a product-by-product basis.. However, there are no significance procedures to test the significance of each attribute's effect on Overall Liking. The bootstrap method, is a well-established computer-based Monte Carlo technique for determining an estimate for the standard errors, confidence intervals, biases, and prediction errors.

Requirements for the Analysis

This technique requires the raw data on the JAR and overall attributes for each product. It additionally requires specialized statistical software, which can be written in statistical packages such as SAS and R.

"How to"

A bootstrap analysis generates hundreds of samples from the original data, performs the penalty analysis on each sample, and then aggregates the results of the individual penalty analyses to assess the variability of the estimates in the original complete data. Each bootstrap sample is a random sample of size *N*, *with replacement*, from the original sample of size *N*. In large samples, that mean each observation will only occur in about 2/3 of the bootstrap samples.

The process of estimating the standard error of a mean drop is illustrated in Fig. 1. The data set in Fig. 1 contains ten pairs of overall liking scores and JAR flavor scores. Mean drops for "Too Little" and "Too Much" categories were calculated according to penalty analysis for each bootstrap sample.

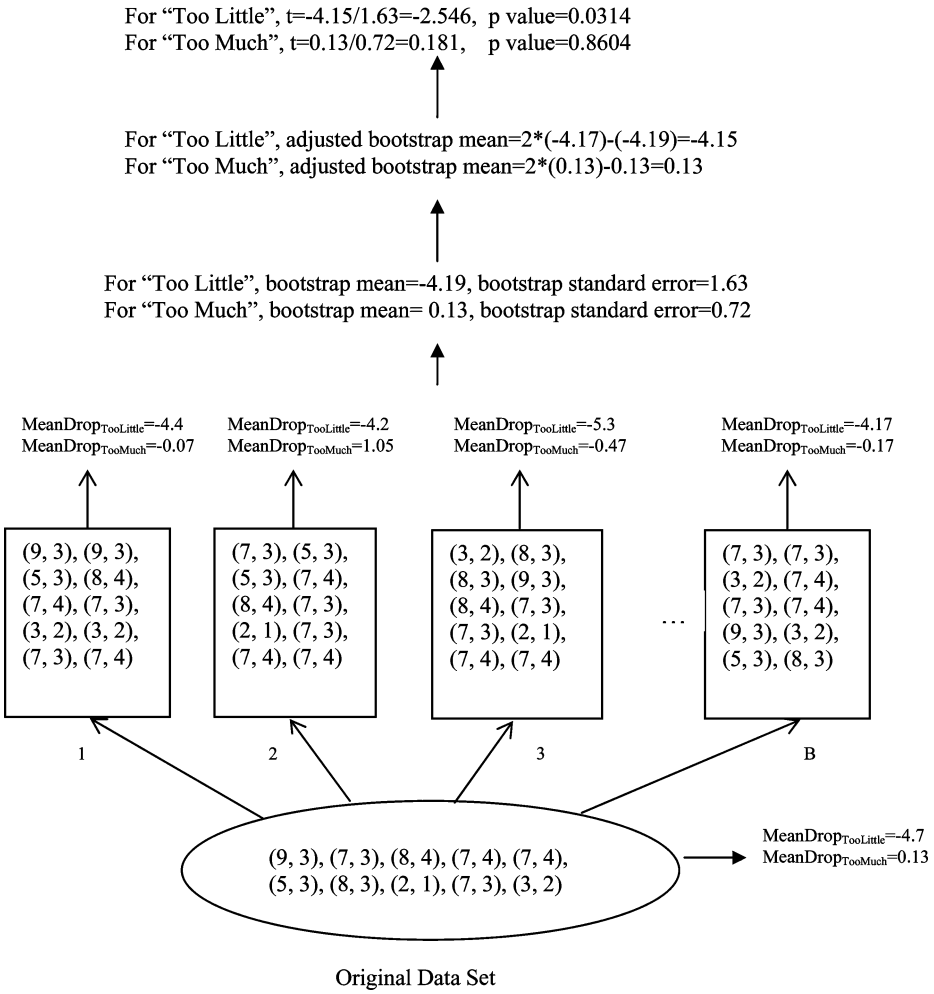

Fig. 1—Scheme for bootstrapping penalty analysis with bootstrap replications of B=10,000.

[1] *Department of Food Science, University of Arkansas, Fayetteville, AR 72704.*

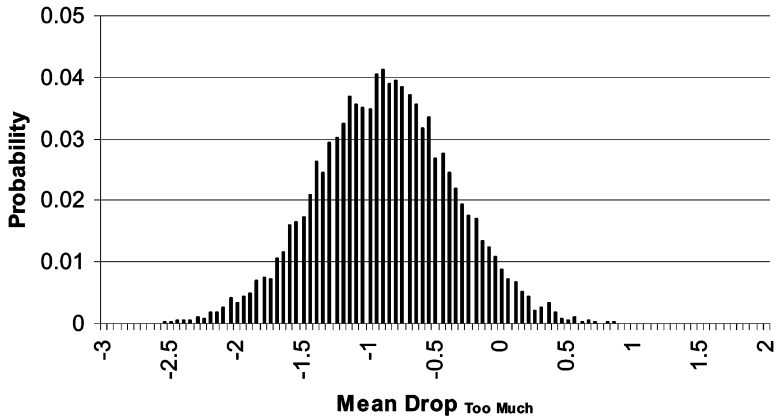

Fig. 2—Histogram of the mean drops for "Too Much" flavor for sample 170 with bootstrap replications of 10,000.

The boostrap estimate of variance of the mean drop is estimated by the variance of the mean drop across the bootstrap samples. When there are B bootstrap samples, the mean drop and its standard error are computed using the following formula:

$$\text{Bootstrap estimate of mean: } \bar{s}_b = \frac{\Sigma_{i=1}^{B} s_i^*}{B}$$

and

$$\text{Bootstrap estimate of standard error: } \hat{\text{se}}_b = \sqrt{\frac{\Sigma(s_i^* - \bar{s}_b)^2}{B-1}}$$

where s_i^* is the mean drop for the ith bootstrap sample, B is the number of bootstrap samples or replications, $\bar{s}_b$ is the bootstrap estimate of the mean, and $\hat{\text{se}}_b$ is the standard error of the mean. The bootstrap estimate of the mean is somewhat biased. The bias can be removed using the following adjustment:

$$\bar{s}_b' = 2s_n - \bar{s}_b$$

where s_n is the estimate of the mean from the original data set, and $\bar{s}_b'$ is the adjusted bootstrap estimate of the mean. The number of bootstrap samples (B) is at least 100 and often several thousands. In practice, B is either chosen based

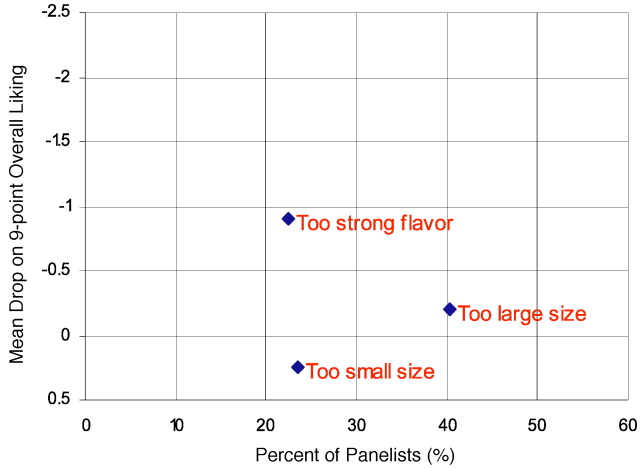

Fig. 3—Penalty analysis for sample 170.

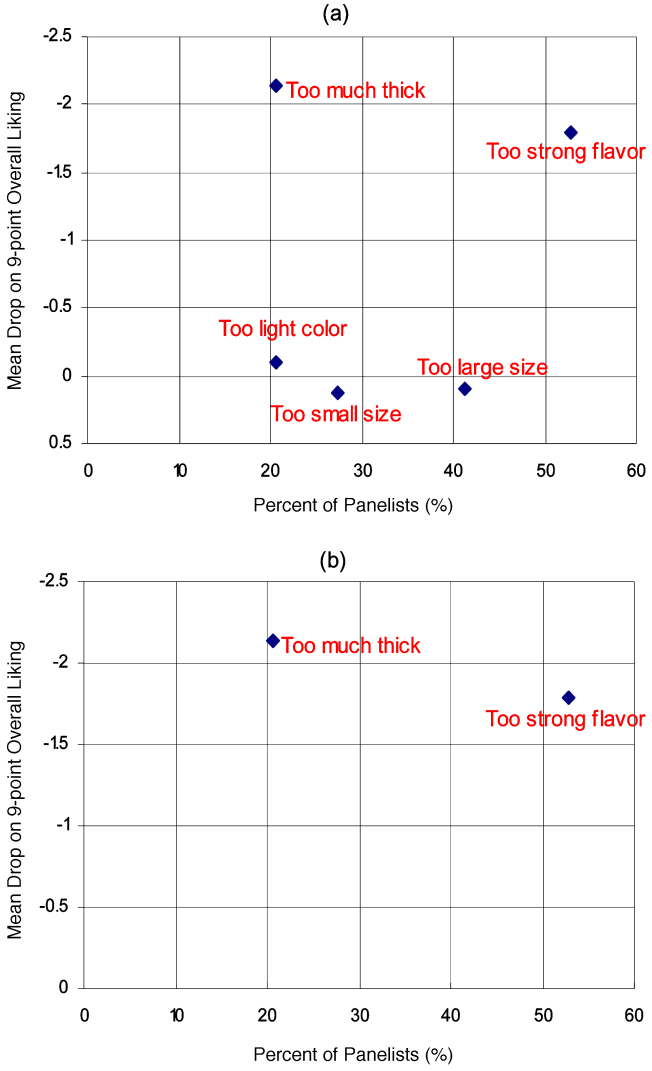

Fig. 4—Comparison of penalty and bootstrapping penalty analyses.

TABLE 1—Penalty and bootstrapping penalty analyses for Sample 170.

	Penalty Analysis		Bootstrapping Penalty Analysis				
	% of Panelists	Mean Drop	% of Panelists	Adj. Mean Drop	Standard Error	*t*-value	*p*-value[a]
"Too small" Size	23.53	0.25	23.76	0.25	0.57	0.44	0.6586
"Too large" Size	40.20	–0.20	40.59	–0.2	0.51	–0.39	0.6994
"Too strong" Flavor	22.55	–0.91	22.77	–0.92	0.5	–1.82	0.0711

[a]Two-tailed test was used for *p*-value (n=101)

TABLE 2—Penalty and bootstrapping penalty analyses for Sample 458.

	Penalty Analysis		Bootstrapping Penalty Analysis				
Attribute	% of Panelists	Mean Drop	% of Panelists	Adj. Mean Drop	Standard Error	*t*-value	*p*-value[a]
"Too small" Size	31.37	0.23	31.37	0.24	0.44	0.54	0.5895
"Too large" Size	32.35	0.34	32.35	0.34	0.39	0.87	0.3853
"Too strong" Flavor	32.35	–1.35	32.35	–1.34	0.35	–3.87	0.0002

[a]Two-tailed test was used for *p*-value (n=101)

TABLE 3—Penalty and bootstrapping penalty analyses for Sample 523.

	Penalty Analysis		Bootstrapping Penalty Analysis				
Attribute	% of Panelists	Mean Drop	% of Panelists	Adj. Mean Drop	Standard Error	*t*-value	*p*-value[a]
"Too small" Size	24.51	0.23	24.51	0.23	0.48	0.48	0.6357
"Too large" Size	39.22	0.34	39.22	0.34	0.39	0.87	0.3876
"Too strong" Flavor	21.57	–1.56	21.57	–1.57	0.39	–3.98	0.0001

[a]Two-tailed test was used for *p*-value (n=102)

TABLE 4—Penalty and bootstrapping penalty analyses for Sample 896.

	Penalty Analysis		Bootstrapping Penalty Analysis				
Attribute	% of Panelists	Mean Drop	% of Panelists	Adj. Mean Drop	Standard Error	*t*-value	*p*-value[a]
"Too small" Size	21.57	–0.18	21.57	–0.19	0.46	–0.4	0.6891
"Too large" Size	39.22	–0.40	39.22	–0.40	0.44	–0.91	0.3629

[a]Two-tailed test was used for *p*-value (n=102)

TABLE 5—Penalty and bootstrapping penalty analyses for Sample 914.

	Penalty Analysis		Bootstrapping Penalty Analysis				
Attribute	% of Panelists	Mean Drop	% of Panelists	Adj. Mean Drop	Standard Error	*t*-value	*p*-value[a]
"Too small" Size	27.45	0.13	27.45	0.14	0.54	0.26	0.79294
"Too large" Size	41.18	0.09	41.18	0.09	0.52	0.16	0.86934
"Too light" Color	20.59	–0.10	20.59	–0.11	0.55	–0.2	0.83873
"Too weak" Flavor	52.94	–1.80	52.94	–1.79	0.37	–4.87	<0.00001
"Too much" Thickness	20.59	–2.14	20.59	–2.14	0.42	–5.05	<0.00001

[a]Two-tailed test was used for *p*-value (n=100)

on prior experience, or a moderate B is selected and B is increased until the distribution of mean drops stabilizes.

Example using Case Study Data

This example uses the mean drops in overall liking for the JAR attributes which had more than 20 % in the non-Jar category in that code. 10,000 bootstrap samples were selected, although 1000 would have been sufficient.

Results and Conclusions

In this case study, the bootstrap distributions for each mean drop were approximately normally distributed, so a *t*-test was used to test the significance of each of the mean drops.. Figure 2 display an example histogram of the bootstrap values for the "Too Much" mean drop on the flavor attribute for Sample 170 The results from bootstrapping penalty analysis

TABLE 6—Comparison of five samples.

Sample	Size	Color	Flavor	Thin/Thick	Stickiness	Observed OAL Mean
170	□	□	□	□	□	5.52
458	□	□	×	□	□	5.47
523	□	□	×	□	□	6.00
896	□	□	□	□	□	6.30
914	□	□	×	×	□	6.56

□ stands for being JAR or no significant effect on OAL
× stands for not being JAR and having a significant effect on OAL

are presented in Table 1. Figure 3 summaries the penalty analysis of three mean drops for code 170. It is apparent from Table 1 that the mean drops were not statistically significant at a significance level of 0.05, so the data are insufficient to conclude that these drops have a strong effect on on Overall Liking (OAL).

Table 2 summarizes the penalty and bootstrapping penalty analyses for Sample 458. Similar to Sample 170, more than 20 % of panelists rated size and flavor as not being JAR. Only the "Too Strong" flavor rating had a decreased mean drop. The other attributes had mean increases. Since they are less than JAR, one would have expected a decrease in Overall Liking associated with these attributes not being JAR. The statistical tests confirm that only "Flavor Too Strong" had a significant impact on Overall Liking, while the other two are not significant.

Table 3 presents the analyses for Sample 523, *P*-values for "Too Small" and "Too Large" size were 0.64 and 0.39, respectively, indicating that the mean drops did not significantly differ from zero. As with sample 458, Overall Liking decreases significantly ($p < 0.001$) when the assessors judged the flavor to be "Too Strong."

Table 4 summarizes the analysis for Sample 896. None of the attributes have a significant effect on Overall Liking.

Finally Table 5 presents the analyses for Sample 914. A "Too Weak" flavor rating was associated with a significant mean drop as was the "Too Thick" rating ($p < 0.0001$).

Table 6 summarizes the individual product analyses reported above.. Size, color, and stickiness were either at a JAR level or simply had no significantly negative impact on Overall Liking for all samples. The inappropriateness of Flavor and Thin/Thick reported for Samples 896 and 914 had a significant impact on OAL.

Conclusions from the Analysis

- According to the bootstrapping penalty analysis, none of the five attributes (Size, Color, Flavor, Thin/Thick, Stickiness) had a significant impact on liking scores. The results imply that the JAR attributes evaluated did not capture the sensory weaknesses of this product.
- For Sample 458, a significant proportion of consumers found the flavor to be too strong and this was detrimental to OAL for these consumers.
- For Sample 523, the flavor was also found by some consumers to be too strong and this had a significant negative impact on OAL.
- None of the JAR attribute scores were found to significantly affect the OAL scores for Sample 896.
- The Thin/Thick and Flavor attributes were not found to be JAR for Sample 914. Being too thick and too weak in flavor had a significant negative impact on OAL. This sample was, however, the most liked product.

Pros and Cons

Bootstrapping penalty analysis determines the influence of JAR attributes on overall liking by identifying the attributes for which the mean drops on overall liking are significant. The precision of a bootstrap analysis estimate depends on how many times the data the original data are randomly bootstrapped.

This technique requires specialized software or programming skill and may require a statistical consultant.

Recommendations

This method is recommended as an enhancement to penalty analysis for JAR scales when the analyst wishes to have the capability to test the significance of an observed mean drop.

References

[1] Efron, B., *The Jackknife, the Bootstrap, and Other Resampling Procedures*, SIAM Publications, Philadelphia, 1982.

[2] Tukey, J. W., "Bias and Confidence in Not Quite Large Samples," (Abstract). Ann. Math. Stat., Vol. 29, 1982, p. 614.

Appendix P: Opportunity Analysis

Tony Gualtieri[1]

Introduction and Objectives

Typical penalty analysis (see Appendix M) assesses the association between JAR scale attributes and Overall Liking. It does not account for the relationship between attribute liking and Overall Liking. Opportunity analysis remedies that lack with an additional computation to address the potential effect of changing an attribute on consumers who already like the product. The analysis provides simple summaries of the relationship between overall and attribute liking for that product. This summary scatter plot illustrates the possible risks (defined as a decrease in the proportion of product likers) and opportunities (defined as an increase in the proportion of product likers) associated with changing a product's attribute liking.

Requirements

This analysis requires individual respondents' scores for each JAR scale, attribute liking scale, and Overall Liking. The analysis can be completed only when both JAR data and attribute liking data are available for the attributes of interest.

"How to"

Dichotomize the liking scales into two groups: "likers" and "dislikers." As the results are most useful when the two groups are of equal size, the break-point on a 9-point hedonic scale may fall close to 7, and will differ based on product category. The same division is used to define attribute liking.

Each hedonic attribute (e.g., "liking of salt") receives a risk and opportunity score. "Risk" is the percentage of product likers who also like the attribute (the risk being that they may like the product *because* they like the attribute) and "Opportunity" is the percentage of product dislikers who also dislike the attribute (the opportunity being that they may dislike the product because of their dislike of the attribute). These responses are defined as:

$$Risk = \frac{\text{Count of respondents who are both product and attribute likers}}{\text{Count of all product likers}} \times 100$$

$$Opportunity = \frac{\text{Count of respondents who are both product and attribute dislikers}}{\text{Count of all product dislikers}} \times 100$$

A change in a high-risk attribute may result in decreased liking, while an improvement in a high opportunity attribute may increase liking.

The Venn diagram in Fig. 1 illustrates the conceptual framework of opportunity analysis. The solid circle represents the set of product likers and the dotted circle represents the set of the specific attribute acceptors. The intersection and compliment of these two sets define four classes of respondents:

A. Product Likers / Attribute Dislikers
B. Product and Attribute Likers
C. Product Dislikers / Attribute Likers
D. Product and Attribute Dislikers

Improving the penalized attribute directionally based on indications from penalty analysis may induce members of class D, the "opportunity group," to become members of class B. That is, the people in class D may dislike the product because of the attribute and fixing the attribute could cause them to like the product. At the same time, class B, the "risk group," may no longer like the product if the attribute, which they currently like, is changed. They could become members of class D. Note that classes A and C are assumed to be unaffected by changes in this particular attribute because their product acceptance does not appear to be based on attribute liking. A statistical comparison between risk scores can be made based on McNemar's test (see Appendix G); however, because of small sample sizes and multiple-comparison issues, only large differences tend to be significant.

Example from Case Study Data

This example summarizes an opportunity analysis of the Products 170, 896, and 914. Table 1 displays the mean ratings for each product and indicates the statistical grouping of the product within an attribute. Table 2 displays the results of an accompanying penalty analysis. Figures 2–4 depict the Risk and Opportunity scores for each product.

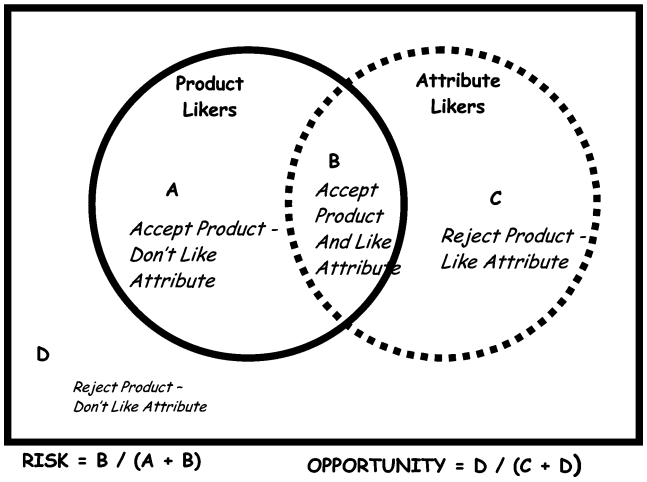

Fig. 1—Opportunity analysis.

[1] *Kellogg's, One Kellogg Square, P.O. Box 3599, Battle Creek, MI 49016.*

TABLE 1—Hedonic liking.

Product	Overall Liking	Size Liking	Color Liking	Flavor Liking	Texture Liking
170	5.5 b	7.1 b	7.1 a	5.6 c	6.5 b
896	6.3 a	7.5 a	7.4 a	6.1 b	7.0 a
914	6.5 a	5.5 c	5.8 b	6.5 a	6.5 b

Products within a column with different letters are significantly different ($p < 0.20$).

Results and Conclusions

Product 170 has inconclusive results (see Fig. 2). Flavor, the only attribute to receive a penalty, is important both to product likers and dislikers and is therefore considered both high risk and high opportunity. Since this product is least liked, the risk group is probably less important than the opportunity group. This implies that reducing the flavor strength is a promising strategy.

For Product 896, flavor has the highest opportunity and the lowest risk (Fig. 3). If improvement is needed, this is the area to focus on, even though it does not have a penalty for flavor. It may be that consumers fault the character of the flavor rather than its intensity.

Product 914 has highest opportunity and lowest risk for appearance attributes; flavor is both high risk and high opportunity, and texture is high risk and low opportunity (Fig. 4). Based on penalty analysis results, reducing the thickness of this product will likely improve liking. There is nothing in the opportunity analysis to suggest that altering the thickness is risky, as long as thickness is related to size and not to texture. On the other hand, making the flavor stronger *may* improve liking—the penalty analysis certainly suggests it; however, caution is advised: the position of this attribute in the opportunity analysis chart implies high risk along with the high opportunity. Unlike Product 170, the risk group here is important since the product received the highest Overall Liking score and a change in flavor could result in a lower proportion of product acceptors.

Summary of Diagnostics

Product 170: Reduce flavor strength (risk is negated by opportunity).

Product 896: Don't change unless needed; flavor is safer to change than appearance or size.

Product 914: Don't change unless needed: a flavor change is risky and probably not warranted; change the appearance instead.

TABLE 2—Penalty analysis summary.

Product	Penalty	Mean Drop in Overall Liking	Percent in Not Right Group	Penalty Score
170	"Flavor Too Strong"	0.91	23	20
896	*No Major Penalties*	N/A	N/A	N/A
914	"Flavor Too Mild"	1.80	53	95
	"Too Thick"	2.14	21	45

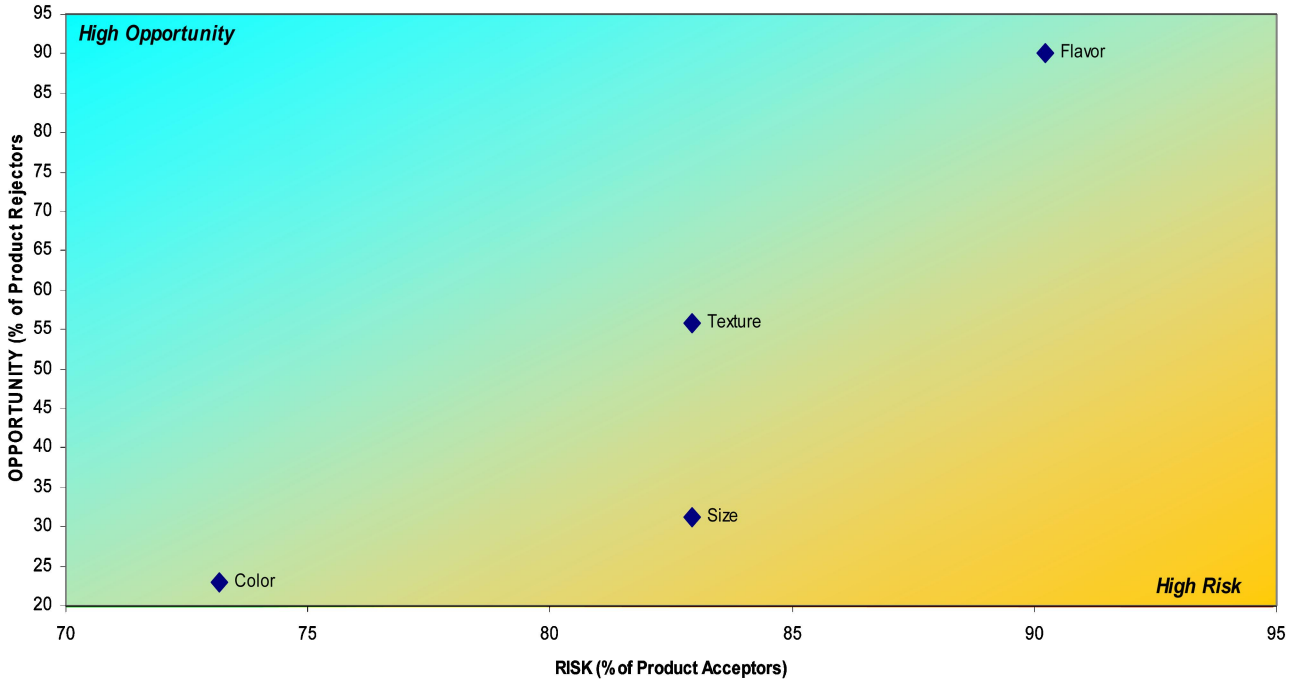

Fig. 2—Opportunity Analysis. ASTM sample data-Product 170. (Likers 40 %-top 3 box.)

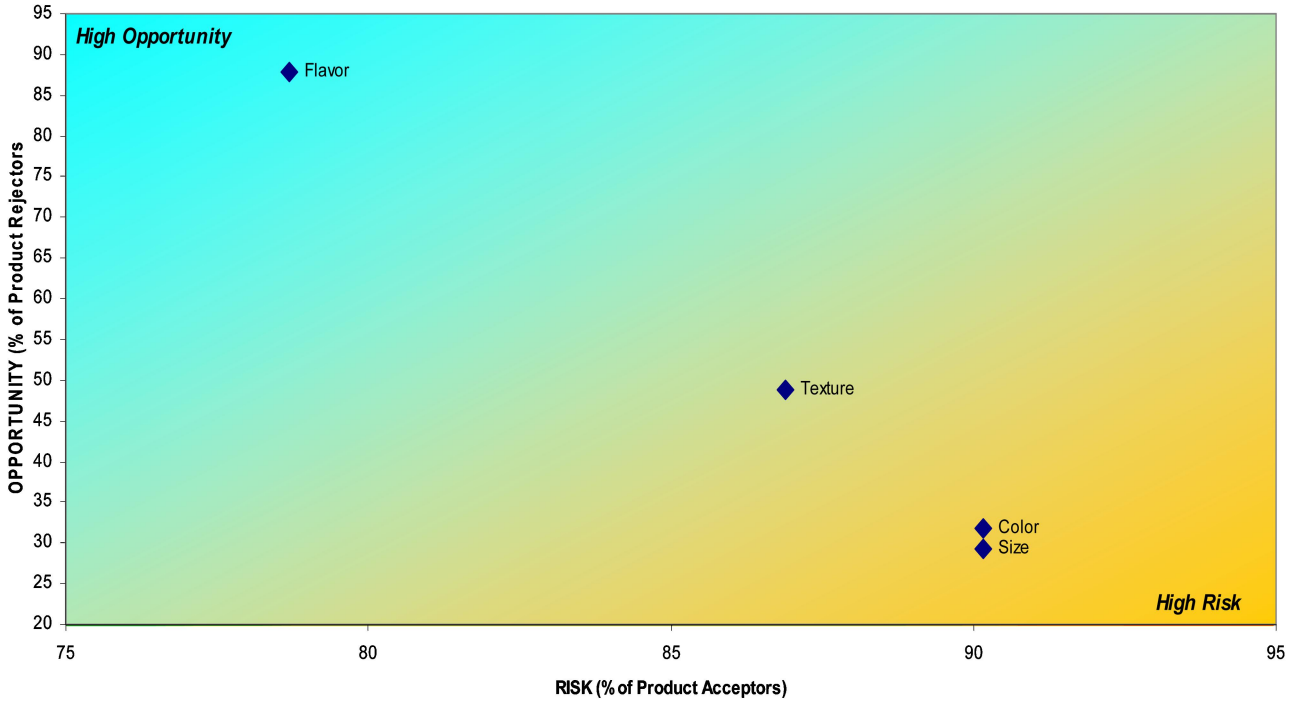

Fig. 3—Opportunity analysis. ASTM sample data-Product 896. (Likers 60 %-top 3 box.)

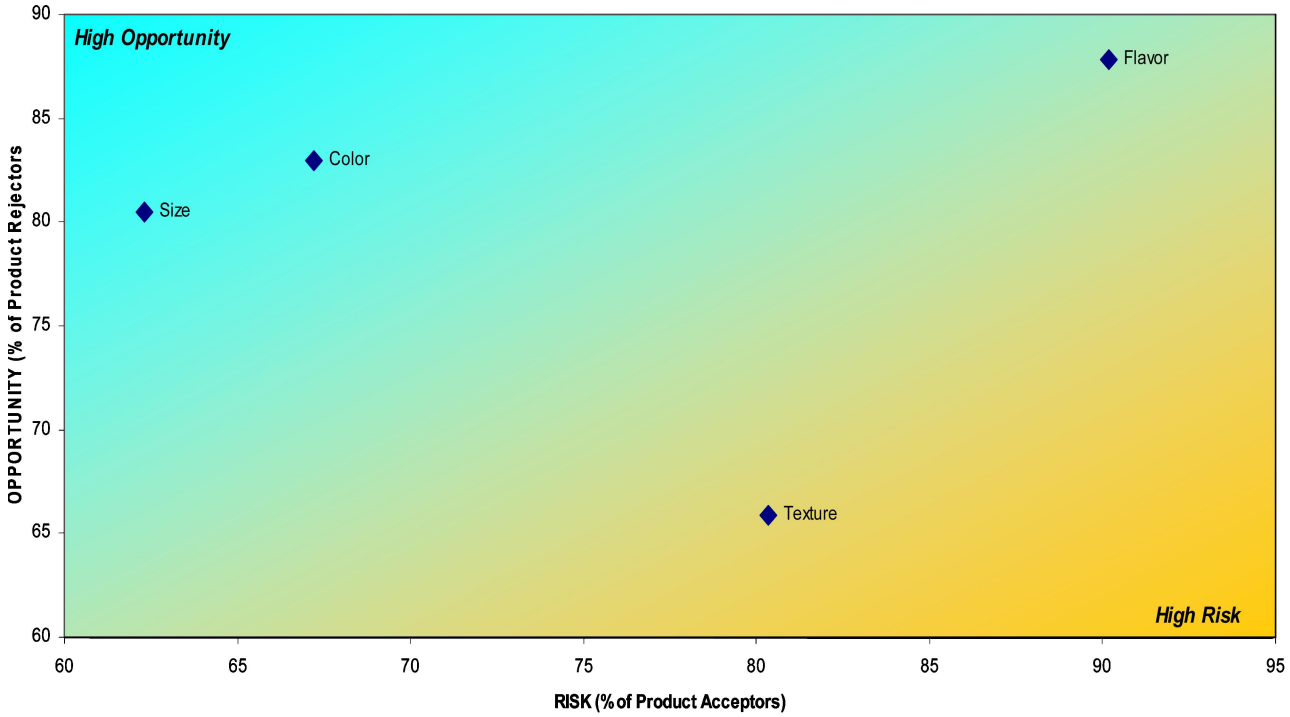

Fig. 4—Opportunity analysis. ASTM sample data-Product 914. (Likers 60 %-top 3 box.)

Pros and Cons

The analysis gives researchers a view of the relationships between attributes and their relative impact on Overall Liking. Opportunity analysis alone does not provide direction for improvement. Instead, it assesses the consequences of making attribute changes on overall liking. Therefore, it should always be run in conjunction with penalty analysis. One disadvantage of this method is that for every attribute JAR scale included in the questionnaire, a scale for liking of the attribute must also be included. This increases the questionnaire length and negates a key benefit of JAR scale usage (see Appendix J), that attribute intensity and hedonic judgment can be combined into one scale.

Recommendation

Opportunity analysis is recommended in combination with penalty analysis when the researcher is amenable to including attribute liking scales as well as JAR scales in the questionnaire.

Appendix Q: PRIMO Analysis

Efim Shvartsburg[1]

Introduction and Objectives

PRIMO (Product Improvement Opportunity) analysis is a method whose initial step is penalty analysis (see Appendix C), followed by the application of Bayesian theory to identify desired changes in JAR attribute ratings in order to maximize the probability of potential product improvement, according to a given criterion. The criterion can be either categorical, such as Top or Top Two Box scale ratings, or continuous, such as mean Overall Liking. For each attribute, this analysis provides a confidence level that the product can be significantly improved with respect to the chosen criterion, by altering the current level of the attribute. The resulting output is a list of JAR attributes and their respective confidence levels, in descending order, corresponding to the potential product improvement. As a result of PRIMO analysis, researchers gain an understanding on how to alter the product on the attributes that promise a high confidence level for improvement. This procedure puts PRIMO Analysis in the class of Bayesian choice models.

Requirements

To conduct PRIMO analysis, raw data for the criterion to be improved (for example, overall liking) as well as the JAR attributes are required.

"How to"

The analysis includes three steps:

1) estimating parameters of the Bayesian model for each JAR attribute for the product
2) finding the optimal attribute decision for each model
3) rank ordering the optimal decision confidence levels for each model

To estimate parameters of the Bayesian model, the following

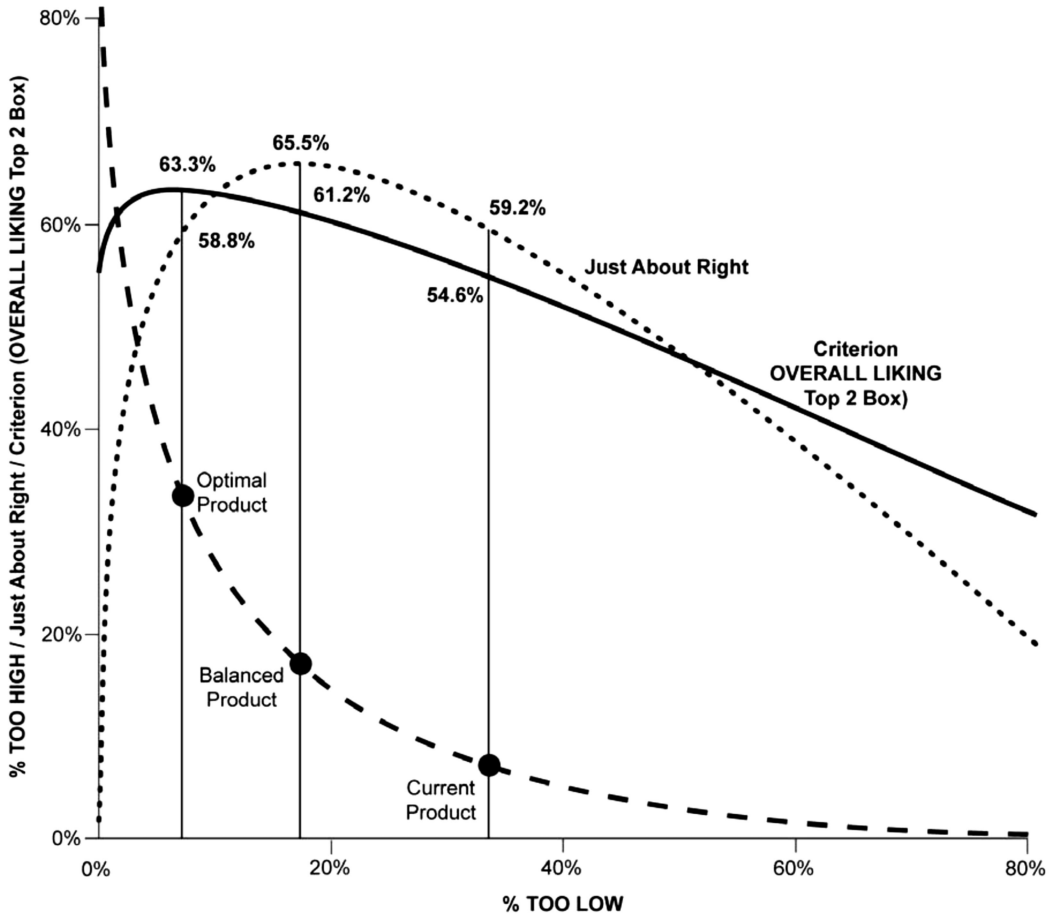

Fig. 1—PRIMO analysis for SALTINESS attribute. The dashed line represents the trajectory of possible product movements in decision space. The solid line represents criterion change along the product trajectory. The dotted line represents "Just About Right" rating change along product trajectory.

[1] *PERT Survey Research, 522 Cottage Grove Road, Bloomfield, CT 06002.*

two assumptions about the consumer population are utilized:

1) For each JAR attribute, there exists an underlying psychophysical continuous measure (not necessarily explicitly specified or known to experimenter).
2) Bayesian utility for this hypothetical measure has a normal distribution.

In other words, each consumer has a utility for a corresponding JAR attribute on this underlying psychophysical continuum. The value of the consumer's utility for each possible level of an attribute depends on his or her perception of the attribute, and is independent of the tested product. When the attribute level is changed, the perception of what is "Just Right" or "Too Low" or "Too High" for the consumer remains the same. Additionally, the proportions of consumers who find some level of an attribute to be "Too Low" through "Just Right" through "Too High" are normally distributed.

After the parameters of a Bayesian model and the product positioning in consumer space are simultaneously estimated, ratings for any possible product repositioning can be predicted. This allows finding the optimal product positioning on any attribute for the specified criterion.

As an example, Fig. 1 illustrates the Bayesian decision space for a hypothetical product for a Saltiness attribute [1].

It is important to emphasize that decision space for each JAR attribute is one-dimensional, so each product positioning uniquely defines all related measures: "% Too Low," "% Too High," "% Just About Right," and expected criterion value. Any of these measures can be used as the independent variable that uniquely defines all other variables. In Fig. 1, "% Too Low" is shown as the independent variable that defines all other measures of product performance: "% Too High," "% Just About Right," and expected criterion value.

Due to the fact that the penalty for being "Too Low" (51.8 %) is much greater than the penalty for being "Too High" (18.8 %), optimal product positioning is skewed toward "Too High" dimension. The projected optimal (in this case, maximizing Top Two Box Overall Liking rating) product would have 58.8% "Just About Right" rating, 34.3% "Too High" and only 6.9 % "Too Low." The maximum possible value of the criterion (Top Two Box Overall Liking rating) is 63.3 %.

The balanced product is the one that has equal proportions of "Too High" and "Too Low" ratings and maximum possible value of "Just About Right" rating. In this case, the projected balanced product has 65.5 % "Just About Right" rating and equal 17.3 % of "Too High" and "Too Low" ratings. At the same time, the balanced product has suboptimal criterion level of 61.2 % only.

The optimal product after consequential adjustment of Saltiness attribute would represent potential improvement over the current product with 98 % confidence. This is a confidence level for a one-tailed hypothesis that the projected criterion level (63.3 %) exceeds the current criterion level (54.6 %) based on the current sample of 152 respondents.

Results and Conclusions

Application of PRIMO Analysis, using a Mean Overall Liking criterion, to the data in the case study yielded the following results and recommendations:

PRIMO ANALYSIS FOR PRODUCT 170
CRITERION—MEAN OVERALL RATING

OPPORTUNITY RANK	IMPROVEMENT CONFIDENCE LEVEL	ATTRIBUTES	POST-IMPROVEMENT MEAN	DISTRIBUTION AFTER IMPROVEMENT			IMPROVEMENT BY REDUCING
				% IN JUST RIGHT GROUP	% IN TOO HIGH GROUP	% IN TOO LOW GROUP	
1	(70%)	JAR-Size	5.58	100.0	0.0	0.0	TOO HIGH GROUP
2	(68%)	JAR-Stickiness	5.58	82.5	7.1	10.4	TOO HIGH GROUP
3	(52%)	JAR-Amt. Flavor	5.52	60.3	27.3	12.4	TOO LOW GROUP
4	(51%)	JAR-Thin/Thick	5.52	83.5	6.6	9.9	TOO LOW GROUP
5	(50%)	JAR-Color	5.52	83.6	6.5	9.9	TOO LOW GROUP

PRIMO ANALYSIS FOR PRODUCT 458
CRITERION—MEAN OVERALL RATING

OPPORTUNITY RANK	IMPROVEMENT CONFIDENCE LEVEL	ATTRIBUTES	POST-IMPROVEMENT MEAN	DISTRIBUTION AFTER IMPROVEMENT			IMPROVEMENT BY REDUCING
				% IN JUST RIGHT GROUP	% IN TOO HIGH GROUP	% IN TOO LOW GROUP	
1	(86%)	JAR-Amt. Flavor	5.60	68.6	13.4	18.0	TOO HIGH GROUP
2	(67%)	JAR-Stickiness	5.53	75.4	9.6	15.0	TOO HIGH GROUP
3	(55%)	JAR-Color	5.47	79.5	0.4	20.1	TOO HIGH GROUP
4	(52%)	JAR-Thin/Thick	5.47	75.9	11.6	12.5	TOO LOW GROUP
5	(50%)	JAR-Size	5.30	36.3	32.4	31.4	TOO HIGH GROUP

PRIMO ANALYSIS FOR PRODUCT 914

CRITERION—MEAN OVERALL RATING

OPPORTUNITY RANK	IMPROVEMENT CONFIDENCE LEVEL	ATTRIBUTES	POST-IMPROVEMENT MEAN	DISTRIBUTION AFTER IMPROVEMENT			IMPROVEMENT BY REDUCING
				% IN JUST RIGHT GROUP	% IN TOO HIGH GROUP	% IN TOO LOW GROUP	
1	(97%)	JAR-Amt. Flavor	6.70	61.0	10.8	28.2	TOO LOW GROUP
2	(64%)	JAR-Thin/Thick	6.55	79.2	14.2	6.6	TOO HIGH GROUP
3	(57%)	JAR-Stickiness	6.58	80.1	13.6	6.2	TOO LOW GROUP
4	(56%)	JAR-Color	6.53	100.0	0.0	0.0	TOO LOW GROUP
5	(50%)	JAR-Size	6.44	31.4	41.2	27.5	TOO HIGH GROUP

PRIMO ANALYSIS FOR PRODUCT 896

CRITERION—MEAN OVERALL RATING

OPPORTUNITY RANK	IMPROVEMENT CONFIDENCE LEVEL	ATTRIBUTES	POST-IMPROVEMENT MEAN	DISTRIBUTION AFTER IMPROVEMENT			IMPROVEMENT BY REDUCING
				% IN JUST RIGHT GROUP	% IN TOO HIGH GROUP	% IN TOO LOW GROUP	
1	(75%)	JAR-Amt. Flavor	6.36	100.0	0.0	0.0	TOO HIGH GROUP
2	(68%)	JAR-Size	6.35	31.4	10.1	58.4	TOO HIGH GROUP
3	(56%)	JAR-Stickiness	6.25	88.6	9.8	1.6	TOO HIGH GROUP
4	(51%)	JAR-Thin/Thick	6.31	89.2	4.1	6.7	TOO LOW GROUP
5	(50%)	JAR-Color	6.23	96.1	2.9	1.0	TOO HIGH GROUP

PRIMO ANALYSIS FOR PRODUCT 458

CRITERION—MEAN OVERALL RATING

OPPORTUNITY RANK	IMPROVEMENT CONFIDENCE LEVEL	ATTRIBUTES	POST-IMPROVEMENT MEAN	DISTRIBUTION AFTER IMPROVEMENT			IMPROVEMENT BY REDUCING
				% IN JUST RIGHT GROUP	% IN TOO HIGH GROUP	% IN TOO LOW GROUP	
1	(100%)	JAR-Amt. Flavor	6.33	100.0	0.0	0.0	TOO HIGH GROUP
2	(75%)	JAR-Stickiness	6.08	87.5	6.6	5.9	TOO HIGH GROUP
3	(59%)	JAR-Thin/Thick	6.03	85.5	5.6	8.9	TOO LOW GROUP
4	(56%)	JAR-Color	6.02	79.0	2.3	18.7	TOO HIGH GROUP
5	(50%)	JAR-Size	5.81	36.3	39.2	24.5	TOO HIGH GROUP

Product 914 is recommended as the best prototype. It is further recommended to increase the strength of flavor. The resulting product could achieve a Mean Overall Liking score of 6.70 that is better than the current Overall Liking level with 97 % confidence. Although Product 523 can be improved with more than 99.5 % confidence by reducing strength of flavor, the resulting product will not achieve an Overall Liking score comparable with Product 914.

Pros and Cons

PRIMO analysis allows researchers to find optimal product improvement decisions in regard to any product performance criterion for which there exists a JAR attribute. It provides researchers with quantifiable recommendations regarding the attributes that can be improved, the direction of improvement for each attribute, the potential improvement in terms of criterion of product performance, and the confidence level that statistically significant product improvement could be achieved.

The universal nature of the PRIMO analysis Bayesian choice model allows comparisons and optimal choice not only between JAR attributes, but also between various products and criteria. PRIMO analysis is especially useful when dealing with several product prototypes, aiding in the identification of the prototype with the most potential for improvement. Also, PRIMO analysis provides a natural segue into risk analysis by providing confidence levels of successful product improvement.

The limitation of PRIMO analysis is an unspoken assumption that attributes are mutually independent and that altering one product attribute would not affect other attributes' ratings. This assumption is frequently violated in practice. PRIMO analysis evaluates one attribute at a time. The version of the analysis that simultaneously analyzes all pairs of attributes requires a very large sample size for high confidence inferences, which makes it too expensive to implement for practical purposes.

Also, PRIMO analysis requires a special software program, which can be developed using SAS or Excel macros or purchased from the author.

Recommendation

PRIMO analysis is recommended for JAR scale analysis whenever the researcher wishes to determine the attributes of a product having the most potential for improvement.

References

[1] The theory behind the decision space is explained in *Decision Space: Multidimensional Utility Analysis* (Cambridge Studies in Probability, Induction and Decision Theory), Cambridge University Press, 2007.

Appendix R: Chi-Square

Lynn Templeton[1]

Introduction and Objectives

The chi-square approach looks at the relationship between "Just Right" and hedonic ratings for each product, using only the data from the "Not Just Right" sections of the scale. The objective of using a chi-square analysis is to determine if there are significant differences in liking between respondents that have rated the product differently on the JAR scales, i.e., to test the null hypothesis that there is no association between the JAR group and overall liking. For example, one could use the chi-square approach to determine whether respondents that rated the product as "Not Sweet Enough" differed from those respondents who rated the product as "Too Sweet."

Requirements

To use the chi-square analysis, you must have Overall Liking and JAR ratings for each respondent.

"How to"

Since this technique compares only the non-JAR groups for each attribute, the JAR values are discarded, and the remaining data are transformed for each judge. The JAR frequency data are first collapsed into two categories: 1. "Not Enough" and 2. "Too Much," omitting the "Just Right" scores. Likewise, the hedonic frequency data are collapsed into two categories: 1. "Dislike" and 2. "Like," omitting the "Neither Like nor Dislike" responses.

Then a record for each judge is created, showing into which JAR and hedonic categories he or she falls.

Example

JAR frequency: 1. "Not enough" and 2. "Too Much"
Hedonic frequency: 1. "Dislike" and 2. "Like"

Judge	JAR	Hedonic
1	1	1
2	1	2
3	2	2
4	1	1
5	1	2
⋮	⋮	⋮
N	1	1

TABLE 1—Example 1.

	Like	Dislike	Row Totals
Frequencies, "Not Enough"	20	7	27
Frequencies, "Too Much"	33	9	42
Column totals	53	16	69

TABLE 2—Example 2.

	Like	Dislike	Row Totals
Frequencies, "Not Enough"	23	22	45
Frequencies, "Too Much"	16	2	18
Column totals	39	24	63

The following step is to construct a 2 by 2 contingency table of these four categories. That table might look like Table 1.

The formula to calculate the chi-square is as follows:

$$\chi^2 = \sum_i^n \sum_j^p \frac{[X_{ij} - \mathrm{E}(X_{ij})]^2}{\mathrm{E}(X_{ij})} = \sum_i^n \sum_j^p \frac{\left(X_{ij} - \frac{X_{i+}X_{+j}}{X_{++}}\right)}{\frac{X_{i+}X_{+j}}{X_{++}}}$$

where

χ^2 = chi-square value,
X_{ij} = observed value,
$\mathrm{E}(X_{ij})$ = expected values,
X_{11} = is the element in the first row and first column of the table,
X_{1+} and X_{+1} = are the corresponding marginal sums, and
X_{++} = is the global sum

The chi-square statistic should be compared against the appropriate critical value from the chi-square table. For the test described here, that critical value for a test at the 95% level of confidence is 3.84 [1]. If the test statistic is above that critical value, the difference between the two groups is significant. Otherwise there is insufficient evidence to reject that hypothesis. The degrees of freedom are calculated as follows: $(n-1)\ (p-1)$, with n = number of rows and p = number of columns. For, Table 1, the degrees of freedom is $(2-1)\ (2-1) = 1$.

$$\chi^2 = \frac{(20 - [(27 * 53)/69)]^2}{(27 * 53)/69} + \frac{(7 - [(27 * 16)/69)]^2}{(27 * 16)/69} + \frac{(33 - [(42 * 53)/69)]^2}{(42 * 53)/69} + \frac{(9 - [(42 * 16)/69)]^2}{(42 * 16)/69} = 0.19$$

Conclusion: Because $0.19 < 3.84$, which is the critical value at 95% confidence with 1 degree of freedom (df), we conclude that there is no relationship between Overall Liking and the JAR attribute ratings. For Table 1, the computed chi-square value is 0.19(df = 1).

For, Table 2, the computed chi-square value is of 7.78(df = 1). The critical value of Chi-Square at 5% level of significance for 1 degree of freedom is 3.84 [1].

Conclusion: Because the computed chi-square value of $7.78 > 3.84$, at 95% level of confidence and 1 degree of freedom, we can conclude that overall liking and the JAR attribute rating are not independent. Examining the table

[1] *S.C. Johnson & Son, Inc., 1525 Howard Street, Racine, WI 53403-2236.*

TABLE 3—Chi-square, Yates' corrected chi-square, phi-square, and Fisher's exact tests summary results for Size and Color, for samples 170, 458, 523, 896, and 914 of the ASTM data set for JAR scales, including frequencies.

		Overall Liking				Overall Liking		
Sample S#170	Size	Like	Dislike	Row Totals	Color	Like	Dislike	Row Totals
	Frequencies, "Not Enough"	14	10	24	Frequencies, "Not Enough"	3	5	8
	Percent of total	22.95%	16.39%	39.34%	Percent of total	21.43%	35.71%	57.14%
	Frequencies, "Too Much"	23	14	37	Frequencies, "Too Much"	2	4	6
	Percent of total	37.71%	22.95%	60.66%	Percent of total	14.29%	28.57%	42.86%
	Column totals	37	24	61	Column totals	5	9	14
	Percent of total	60.66%	39.34%		Percent of total	35.71%	64.29%	
	Chi-square (df=1)	0.09	$p=0.7649$		Chi-square (df=1)	0.03	$p=0.8721$	
	Yates corrected Chi-square	0	$p=0.9754$		Yates corrected Chi-square	0.16	$p=0.6873$	
	Phi-square	0.00147			Phi-square	0.00185		
	Fisher exact p, one-tailed		$p=0.4861$		Fisher exact p, one-tailed		$p=0.6573$	
	two-tailed		$p=0.7940$		two-tailed		$p=1.0000$	
S#458		Like	Dislike	Row Totals		Like	Dislike	Row Totals
	Frequencies, "Not Enough"	19	10	29	Frequencies, "Not Enough"	7	4	11
	Percent of total	32.76%	17.24%	50.00%	Percent of total	58.33%	33.33%	91.67%
	Frequencies, "Too Much"	20	9	29	Frequencies, "Too Much"	0	1	1
	Percent of total	34.48%	15.52%	50.00%	Percent of total	0.00%	8.33%	8.33%
	Column totals	39	19	58	Column totals	7	5	12
	Percent of total	67.24%	32.76%		Percent of total	58.33%	41.67%	
	Chi-square (df=1)	0.08	$p=0.7797$		Chi-square (df=1)	1.53	$p=0.2165$	
	Yates corrected Chi-square	0	$p=1.0000$		Yates corrected Chi-square	0.03	$p=0.8599$	
	Phi-square	0.00135			Phi-square	0.12727		
	Fisher exact p, one-tailed		$p=0.5000$		Fisher exact p, one-tailed		$p=0.4167$	
	two-tailed		$p=1.0000$		two-tailed		$p=0.4167$	
S#523		Like	Dislike	Row Totals		Like	Dislike	Row Totals
	Frequencies, "Not Enough"	17	5	22	Frequencies, "Not Enough"	7	3	10
	Percent of total	28.81%	8.48%	37.29%	Percent of total	46.67%	20.00%	66.67%
	Frequencies, "Too Much"	28	9	37	Frequencies, "Too Much"	2	3	5
	Percent of total	47.46%	15.25%	62.71%	Percent of total	13.33%	20.00%	33.33%
	Column totals	45	14	59	Column totals	9	6	15
	Percent of total	76.27%	23.73%		Percent of total	60.00%	40.00%	
	Chi-square (df=1)	0.02	$p=0.8891$		Chi-square (df=1)	1.25	$p=0.2636$	
	Yates corrected Chi-square	0.03	$p=0.8595$		Yates corrected Chi-square	0.31	$p=0.5762$	
	Phi-square	0.00033			Phi-square	0.08333		
	Fisher exact p, one-tailed		$p=0.5755$		Fisher exact p, one-tailed		$p=0.2867$	
	two-tailed		$p=1.0000$		two-tailed		$p=0.3287$	
S#896		Like	Dislike	Row Totals		Like	Dislike	Row Totals
	Frequencies, "Not Enough"	17	4	21	Frequencies, "Not Enough"	1	0	1
	Percent of total	28.81%	6.78%	35.59%	Percent of total	25.00%	0.00%	25.00%
	Frequencies, "Too much"	28	10	38	Frequencies, "Too Much"	1	2	3
	Percent of total	47.46%	16.95%	64.41%	Percent of total	25.00%	50.00%	75.00%
	Column totals	45	14	59	Column totals	2	2	4
	Percent of total	76.27%	23.73%		Percent of total	50.00%	50.00%	
	Chi-square (df=1)	0.39	$p=0.5298$		Chi-square (df=1)	1.33	$p=0.2482$	

TABLE 3— (*Continued.*)

Sample S#170	Size	Overall Liking: Like	Dislike	Row Totals	Color	Overall Liking: Like	Dislike	Row Totals
	Yates corrected Chi-square	0.1	$p=0.7575$		Yates corrected Chi-square	0	$p=1.0000$	
	Phi-square	0.00669			Phi-square	0.33333		
	Fisher exact p, one-tailed		$p=0.3851$		Fisher exact p, one-tailed		$p=0.5000$	
	two-tailed		$p=0.7506$		two-tailed		$p=1.0000$	
S#914		Like	Dislike	Row Totals		Like	Dislike	Row Totals
	Frequencies, "Not Enough"	20	7	27	Frequencies, "Not Enough"	15	6	21
	Percent of total	28.99%	10.15%	39.13%	Percent of total	71.43%	28.57%	
	Frequencies, "Too Much"	33	9	42	Frequencies, "Too Much"	0	0	0
	Percent of total	47.83%	13.04%	60.87%	Percent of total	0%	0%	
	Column totals	53	16	69	Column totals	15	6	21
	Percent of total	76.81%	23.19%		Percent of total	71.43%	28.57%	
	Chi-square (df=1)	0.19	$p=0.6657$		Chi-square (df=1)	na	na	
	Yates corrected Chi-square	0.02	$p=0.8888$		Yates corrected Chi-square	na	na	
	Phi-square	0.0027			Phi-square	na		
	Fisher exact p, one-tailed		$p=0.4401$		Fisher exact p, one-tailed		na	
	two-tailed		$p=0.7723$		two-tailed		na	

would lead to the conclusion that the "Too Much" group liked the product more than the "Not Enough" group and the recommendation would be to increase the level of that attribute.

The chi-square statistic is an approximation and is not recommended when any of the cells in the table are <5. In that case, Fisher's exact test should be used. The exact test confirms the results discussed above.

Results and Conclusions

Based on the results shown in Tables 3–6, there are no significant differences in liking between the "Not Enough" and "Too Much" for size, color, amount of flavor, thin/thick, and stickiness, and the Overall Liking using the following analyses: chi-square analysis, Yates' corrected chi-square, and Fisher's exact test. The first recommendation would be to run this data through Fisher's exact test because chi-square is not an appropriate tool when you have cells <5. After running the Fisher (see Table 6), there is no significant difference. The recommendation would be not to change any attributes based on the JAR ratings.

Pros and Cons

Pros

The chi-square test is easy test to execute and interpret. Unlike other statistical tests, the chi-2 square is a nonparametric test which makes no assumptions concerning the form of the original population distribution from which the test data are drawn.

Cons

The chi-square test suffers from limitations of small cell sizes (<5). If small cell sizes (<5) or near-zero cell frequencies, Fisher's test is required, but Fisher's test requires that all marginal totals be fixed, an assumption that is rarely met in practice. Additionally, in collapsing the scale from 5 points to 2 points, there is a loss of data.

Recommendation

The chi-square approach is recommended for determining whether liking differs among respondents rating the product "Not Enough" versus "Too Much."

Appendix

The chi-square tests give probability values for the relationship between two dichotomous variables. They calculate the difference between the data observed and the data expected, considering the given marginals and the assumptions of the model of independence. The chi-square tests give only an estimate of the true chi-square and associated probability value, an estimate which might not be very good in the case of the marginals being very uneven or with a small value ($\sim<5$) in one of the cells. In that case, Fisher's exact test is a good alternative for the chi-square. However, with a large number of cases the chi-square is preferred, as the Fisher test is difficult to calculate.

As long as all of the cells (frequencies per quadrant) have at least a count of 5, then the use of the Pearson chi-square for the significance level is appropriate. This requires having a big enough base so that you will get at least five per quadrant; the marginal totals have to be fixed. It is appropriate to use Fisher's exact test to compute when a table that does not result from missing rows or columns in a larger table has a cell with an expected frequency of less than 5. Yates, corrected chi-square is especially recommended when the same size is small. (Table 6 shows these various approaches.)

TABLE 4—Chi-Square, Yates' corrected chi-square, phi-square, and Fisher's exact tests summary results for Amount of Flavor and Thin/Thick, for samples 170, 458, 523, 896, and 914 of the ASTM data set for JAR scales, including frequencies.

		Overall Liking				Overall Liking		
Sample S#170	Amount Flavor	Like	Dislike	Row Totals	Thin/Thick	Like	Dislike	Row Totals
	Frequencies, "Not Enough"	7	9	16	Frequencies, "Not Enough"	7	6	13
	Percent of total	18.92%	24.32%	43.24%	Percent of total	41.18%	35.29%	76.47%
	Frequencies, "Too Much"	12	9	21	Frequencies, "Too Much"	2	2	4
	Percent of total	32.43%	24.32%	56.76%	Percent of total	11.77%	11.77%	23.53%
	Column totals	19	18	37	Column totals	9	8	17
	Percent of total	51.35%	48.65%		Percent of total	52.94%	47.06%	
	Chi-square (df=1)	0.65	p=0.4194		Chi-square (df=1)	0.02	p=0.8928	
	Yates corrected Chi-square	0.23	p=0.6344		Yates corrected Chi-square	0.19	p=0.6614	
	Phi-square	0.01762			Phi-square	0.00107		
	Fisher exact p, one-tailed		p=0.3175		Fisher exact p, one-tailed		p=0.6647	
	two-tailed		p=0.5148		two-tailed		p=1.0000	
S#458		Like	Dislike	Row Totals		Like	Dislike	Row Totals
	Frequencies, "Not Enough"	2	4	6	Frequencies, "Not Enough"	6	9	15
	Percent of total	5.71%	11.43%	17.14%	Percent of total	25.00%	37.50%	62.50%
	Frequencies, "Too Much"	14	15	29	Frequencies, "Too Much"	4	5	9
	Percent of total	40.00%	42.86%	82.86%	Percent of total	16.67%	20.83%	37.50%
	Column totals	16	19	35	Column totals	10	14	24
	Percent of total	45.71%	54.29%		Percent of total	41.67%	58.33%	
	Chi-square (df=1)	0.45	p=0.5036		Chi-square (df=1)	0.05	p=0.8307	
	Yates corrected Chi-square	0.05	p=0.8269		Yates corrected Chi-square	0.05	p=0.8307	
	Phi-square	0.01278			Phi-square	0.0019		
	Fisher exact p, one-tailed		p=0.4179		Fisher exact p, one-tailed		p=0.5818	
	two-tailed		p=0.6657		two-tailed		p=1.0000	
S#523		Like	Dislike	Row Totals		Like	Dislike	Row Totals
	Frequencies, "Not Enough"	3	1	4	Frequencies, "Not Enough"	7	7	14
	Percent of total	11.54%	3.85%	15.39%	Percent of total	41.18%	41.18%	82.35%
	Frequencies, "Too Much"	10	12	22	Frequencies, "Too Much"	1	2	3
	Percent of total	38.46%	46.15%	84.62%	Percent of total	5.88%	11.77%	17.65%
	Column totals	13	13	26	Column totals	8	9	17
	Percent of total	50.00%	50.00%		Percent of total	47.06%	52.94%	
	Chi-square (df=1)	1.18	p=0.2770		Chi-square (df=1)	0.28	p=0.5997	
	Yates corrected Chi-square	0.3	p=0.5867		Yates corrected Chi-square	0.01	p=0.9105	
	Phi-square	0.04545			Phi-square	0.0162		
	Fisher exact p, one-tailed		p=0.2965		Fisher exact p, one-tailed		p=0.5471	
	two-tailed		p=0.5930		two-tailed		p=1.0000	
S#896		Like	Dislike	Row Totals		Like	Dislike	Row Totals
	Frequencies, "Not Enough"	5	0	5	Frequencies, "Not Enough"	5	3	8
	Percent of total	29.41%	0.00%	29.41%	Percent of total	45.46%	27.27%	72.73%
	Frequencies, "Too Much"	8	4	12	Frequencies, "Too Much"	2	1	3
	Percent of total	47.06%	23.53%	70.59%	Percent of total	18.18%	9.09%	27.27%
	Column totals	13	4	17	Column totals	7	4	11
	Percent of total	76.47%	23.53%		Percent of total	63.64%	36.36%	
	Chi-square (df=1)	2.18	p=0.1399		Chi-square (df=1)	0.02	p=0.8982	
	Yates corrected Chi-square	0.72	p=0.3960		Yates corrected Chi-square	0.33	p=0.5648	

TABLE 4— (*Continued.*)

		Overall Liking				Overall Liking		
Sample S#170	**Amount Flavor**	**Like**	**Dislike**	**Row Totals**	**Thin/Thick**	**Like**	**Dislike**	**Row Totals**
	Phi-square	0.12821			Phi-square	0.00149		
	Fisher exact *p*, one-tailed		*p*=0.2080		Fisher exact *p*, one-tailed		*p*=0.7212	
	two-tailed		*p*=0.2605		two-tailed		*p*=1.0000	
S#914		Like	Dislike	Row Totals		Like	Dislike	Row Totals
	Frequencies, "Not Enough"	33	19	52	Frequencies, "Not Enough"	1	3	4
	Percent of total	60.00%	34.55%	94.55%	Percent of total	4.17%	12.50%	16.67%
	Frequencies, "Too Much"	1	2	3	Frequencies, "Too Much"	10	10	20
	Percent of total	1.82%	3.64%	5.46%	Percent of total	41.67%	41.67%	83.33%
	Column totals	34	21	55	Column totals	11	13	24
	Percent of total	61.82%	38.18%		Percent of total	45.83%	54.17%	
	Chi-square (df=1)	1.09	*p*=0.2963		Chi-square (df=1)	0.84	*p*=0.3596	
	Yates corrected Chi-square	0.19	*p*=0.6648		Yates corrected Chi-square	0.13	*p*=0.7141	
	Phi-square	0.01983			Phi-square	0.03497		
	Fisher exact *p*, one-tailed		*p*=0.3229		Fisher exact *p*, one-tailed		*p*=0.3634	
	two-tailed		*p*=0.5509		two-tailed		*p*=0.5963	

TABLE 5—Chi-Square, Yates' corrected chi-square, phi-square, and Fisher's exact tests summary results for Stickiness, for samples 170, 458, 523, 896, and 914 of the ASTM data set for JAR scales, including frequencies.

Sample	Stickiness	Overall Liking		
S#170		Like	Dislike	Row Totals
	Frequencies, "Not Enough"	1	2	3
	Percent of total	5.00%	10.00%	15.00%
	Frequencies, "Too Much"	8	9	17
	Percent of total	40.00%	45.00%	85.00%
	Column totals	9	11	20
	Percent of total	45.00%	55.00%	
	Chi-square (df=1)	0.19	p=0.6595	
	Yates corrected Chi-square	0.04	p=0.8502	
	Phi-square	0.0097		
	Fisher exact p, one-tailed		p=0.5789	
	two-tailed		p=1.0000	
S#458		Like	Dislike	Row Totals
	Frequencies, "Not Enough"	4	3	7
	Percent of total	15.39%	11.54%	26.92%
	Frequencies, "Too Much"	8	11	19
	Percent of total	30.77%	42.31%	73.08%
	Column totals	12	14	26
	Percent of total	46.15%	53.85%	
	Chi-square (df=1)	0.47	p=0.4951	
	Yates corrected Chi-square	0.06	p=0.8113	
	Phi-square	0.0179		
	Fisher exact p, one-tailed		p=0.4043	
	two-tailed		p=0.6652	
S#523		Like	Dislike	Row Totals
	Frequencies, "Not Enough"	1	1	2
	Percent of total	6.25%	6.25%	12.50%
	Frequencies, "Too Much"	5	9	14
	Percent of total	31.25%	56.25%	87.50%
	Column totals	6	10	16
	Percent of total	37.50%	62.50%	
	Chi-square (df=1)	0.15	p=0.6963	
	Yates corrected Chi-square	0.15	p=0.6963	
	Phi-square	0.00952		
	Fisher exact p, one-tailed		p=0.6250	
	two-tailed		p=1.0000	
S#896		Like	Dislike	Row Totals
	Frequencies, "Not Enough"	1	0	1
	Percent of total	7.69%	0.00%	7.69%
	Frequencies, "Too Much"	6	6	12
	Percent of total	46.15%	46.15%	92.31%
	Column totals	7	6	13
	Percent of total	53.85%	46.15%	
	Chi-square (df=1)	0.93	p=0.3352	
	Yates corrected Chi-square	0.01	p=0.9360	
	Phi-square	0.07143		
	Fisher exact p, one-tailed		p=0.5385	
	two-tailed		p=1.0000	
S#914		Like	Dislike	Row Totals
	Frequencies, "Not Enough"	3	7	10
	Percent of total	15.79%	36.84%	52.63%

TABLE 5— (*Continued.*)

Sample	Stickiness	Overall Liking		
	Frequencies, "Too Much"	6	3	9
	Percent of total	31.58%	15.79%	47.37%
	Column totals	9	10	19
	Percent of total	47.37%	52.63%	
	Chi-square (df=1)	2.55	p=0.1100	
	Yates corrected Chi-square	1.3	p=0.2551	
	Phi-square	0.13444		
	Fisher exact p, one-tailed		p=0.1276	
	two-tailed		p=0.1789	

TABLE 6—Chi-Square, Yates' corrected chi-square, phi-square, and Fisher's exact tests summary results for Size, Color, Amount of Flavor, Thin/Thick, and Stickiness, for samples 170, 458, 523, 896, and 914 of the ASTM data set for JAR scales.

		Size		Color		Amount Flavor		Thin/Thick		Stickiness	
Analysis	Sample #	Value	p	Value	p	Value	p	Value	p	Value	p
Chi-square (df=1)	#170	0.09	0.7649	0.03	0.8721	0.65	0.4194	0.02	0.8928	0.19	0.6595
Yates' corrected	#170	0	0.9754	0.16	0.6873	0.23	0.6344	0.19	0.6614	0.04	0.8502
Chi-square											
Phi-square	#170	0.00147		0.00185		0.01762		0.00107		0.0097	
Fisher exact p,	#170		0.4861		0.6573		0.3175		0.6647		0.5789
one-tailed											
two-tailed	#170		0.7940		1.0000		0.5148		1.0000		1.0000
Chi-square (df=1)	#458	0.08	0.7797	1.53	0.2165	0.45	0.5036	0.05	0.8307	0.47	0.4951
Yates' corrected	#458	0	1.0000	0.03	0.8599	0.05	0.8269	0.05	0.8307	0.06	0.8113
Chi-square											
Phi-square	#458	0.00135		0.12727		0.01278		0.0019		0.0179	
Fisher exact p,	#458		0.5000		0.4167		0.4179		0.5818		0.4043
one-tailed											
two-tailed	#458		1.0000		0.4167		0.6657		1.0000		0.6652
Chi-square (df=1)	#523	0.02	0.8891	1.25	0.2636	1.18	0.2770	0.28	0.5997	0.15	0.6963
Yates' corrected	#523	0.03	0.8595	0.31	0.5762	0.3	0.5867	0.01	0.9105	0.15	0.6963
Chi-square											
Phi-square	#523	0.00033		0.08333		0.04545		0.0162		0.00952	
Fisher exact p,	#523		0.5755		0.2867		0.2965		0.5471		0.6250
one-tailed											
two-tailed	#523		1.0000		0.3287		0.5930		1.0000		1.0000
Chi-square (df=1)	#896	0.39	0.5298	1.33	0.2482	2.18	0.1399	0.02	0.8982	0.93	0.3352
Yates' corrected	#896	0.1	0.7575	0	1.0000	0.72	0.3960	0.33	0.5648	0.01	0.9360
Chi-square											
Phi-square	#896	0.00669		0.33333		0.12821		0.00149		0.07143	
Fisher exact p,	#896		0.3851		0.5000		0.2080		0.7212		0.5385
one-tailed											
two-tailed	#896		0.7506		1.0000		0.2605		1.0000		1.0000
Chi-square (df=1)	#914	0.19	0.6657	na	na	1.09	0.2963	0.84	0.3596	2.55	0.1100
Yates' corrected	#914	0.02	0.8888	na	na	0.19	0.6648	0.13	0.7141	1.3	0.2551
Chi-square											
Phi-square	#914	0.0027		na		0.01983		0.03497		0.13444	
Fisher exact p,	#914		0.4401		na		0.3229		0.3634		0.1276
one-tailed											
two-tailed	#914		0.7723		na		0.5509		0.5963		0.1789

Note: None of the samples shows a significant association between rows and columns.

References

[1] Fisher, R. and Yates, F., *Statistical Tables for Biological, Agriculture and Medical Research*, Oliver and Boyd, Edinburgh, 1948, Table IV.

Appendix S: Biplots, Correspondence Analysis, and Principal Components Analysis

Elizabeth Horn[1] and Cindy Ford[2]

Introduction and Objectives

Multivariate graphical displays, also known as biplots, describe relationships among "Just About Right" attributes across samples. Biplots are a visual means to show multidimensional data relationships. There are many techniques that can be used to generate the necessary data for biplots, such as principal components analysis, multidimensional scaling, correspondence analysis, and discriminant analysis. Two of these methods, correspondence analysis (CA) and principal components analysis (PCA), were considered in this case study

Requirements

To develop a biplot using CA requires count data (frequencies) as input, whereas PCA requires interval data. The biplot of the CA results shows the samples and attributes plotted together. The first two components (or groups of highly correlated attributes) serve as the axes in the biplot. Researchers then can make interpretations on how attributes relate to one another and how sample formulations relate to one another.

"How to"

CA describes the relationships between two categorical variables in a correspondence table (i.e., a raw crosstabulation of the variables commonly containing frequency counts) in a low-dimensional space, while simultaneously describing the relationships between the categories for each variable. The analysis yields coordinates (x-values and y-values) for each attribute and sample. These coordinates are then plotted in a two-dimensional space.

PCA functions in much the same manner as CA. The objective for PCA is to extract two or more underlying components or "themes" from the data. PCA also yields coordinates for each attribute and sample that are then plotted in a two-dimensional space. PCA uses means of interval-scaled data to construct the biplot.

Most biplots allow the researcher to determine:

- the relationships among the samples in multidimensional space
- the relationships among the attributes in multidimensional space
- the degree to which the attributes differentiate the products

These interpretation hints can be used with most biplots:

- Many points clustering around the origin (intersection of the axes) suggest slight differentiation among the perceptions of the samples and their attributes. Conversely, the further away the attributes and samples are from the origin, the more one or more samples and attributes are differentiated from one another.
- The attribute vectors indicate the strength of the relationship between that attribute and the underlying factor/component.
  - The longer the vector, the stronger the attribute's relationship with the underlying factor/component.
  - Attributes close to one another may be seen as more substitutable by consumers.
  - Attributes that point in the same direction are seen as more similar.
- Products that are closer together are seen as more similar to one another on the attributes.
- A product located in the same direction as an attribute vector is characterized by that attribute. This relationship is stronger for those products positioned away from the center of the biplot space.

Analyses contained in this case study considered three samples—labeled 170, 896, and 914—and five JAR attributes—Size, Color, Flavor, Thin/Thick (Thickness), and Stickiness (Texture).

Biplots via Correspondence Analysis

Before the correspondence analysis can be performed, crosstabs resulting in the counts for the JAR attributes were obtained for the categories of "Too Much," "Not Enough," and "About Right." For example, the Color JAR data were the number of respondents that thought the particular sample had "Too Much Color," the number that thought there was "Not Enough Color," and the number that thought the color was "About Right."

[1] *Decision Analyst, Inc. 604 Avenue H East, Arlington, TX 76011.*
[2] *The Modellers, LLC 4505 Wasatch Blvd., Salt Lake City, UT 84124.*

The count data for each of the three sample formulations are shown below:

	Size			Thickness			Color			Flavor			Texture		
Sample (*n* for each sample = 102)	Too Large	About Right Size	To Small	Too Thick	About Right Thickness	Too Thin	Too Colorful	About Right Color	Too Drab	Too Much Flavor	About Right Flavor	Not Enough Flavor	Too Sticky	About Right Texture	Too Smooth
170	41	37	24	5	84	13	6	85	11	23	63	16	17	81	4
896	40	40	22	3	90	9	3	98	1	13	84	5	14	87	1
914	42	32	28	21	77	4	0	81	21	3	45	54	9	82	10

Note: There were missing data such that the "Too Much," "Not Enough," and "Just About Right" counts for a particular attribute may not add to the total sample size of 102. The original JAR scales were 5-point, fully anchored scales. Scales were recoded into three variables for each JAR attribute. "Not Enough" was created by collapsing the responses for scale point 1 ("Not Enough") and scale point 2 ("Somewhat Not Enough"). "Just About Right" was scale point 3 ("Just About Right"). "Too Much" was created by collapsing the responses for scale point 4 ("Somewhat Too Much") and scale point 5 ("Too Much"). Data sets with many cells that contain zero counts may cause unintended bias in the biplot results. Using this technique with small sample sizes (less than 100 respondents) is not recommended

Results

The following figure shows a biplot with the 15 attributes (five JAR attributes [size, color, amount of flavor, thickness, and texture] by three variables ["Too Much," "Not Enough," and "Just About Right"]). This biplot was generated using the singular value decomposition and biplot macros for Excel, available from http://www.stat.vt.edu/facstaff/epsmith.html [1]. Other software packages, such as SPSS® or SAS®, can also produce correspondence analyses and biplots.

The percent variance explained by the first component ("Too Drab," "Too Smooth," "Not Enough Flavor," and "Too Thick") was 88.8 %; the variance explained by the second component ("Too Colorful," and "Too Much Flavor") was 11.2 %. The second component added little to the interpretation of the map. The attributes and products may more appropriately occupy a one-dimensional space. Still, the map may yield insights that are unavailable using other relational methods.

Examining the map, the attribute vectors for "Too Colorful," "Too Much Flavor," "Too Thin," and "Too Sticky" project in the same direction and are thus considered to be related to one another. Sample 170 is associated with being too thin, having "Too Much Flavor," and "Too Much Color." Concerning the positions of the three samples in the map, Sample 914 is perceived to be different from the other two samples, while Sample 896 and Sample 170 are more similar.

Biplots via Principal Components Analysis

Similar to CA, PCA yields coordinates (*x*-values and *y*-values) for each attribute and each sample. These coordinates are then plotted in a two-dimensional space.

The means used in the PCA procedure are below:

Sample (*n* for each sample = 102)	Size	Color	Flavor	Thickness	Texture
170	3.22	2.95	3.07	2.92	3.14
896	3.24	3.02	3.09	2.93	3.14
914	3.10	2.79	2.44	3.17	2.97

Note: The original JAR scales were 5-point, fully anchored scales that ranged from 1 ("Not Enough") to 5 ("Too Much").

Results

The PCA was conducted using five attributes (Size, Color, Amount of Flavor, Thickness, and Texture). The analysis was performed using the singular value decomposition macro in Excel. The loadings, which are measures of the relationship of each attribute to each of the two principal components, are shown below:

Attribute	Component 1	Component 2
Flavor	**0.739**	−0.188
Thickness	**−0.669**	−0.159
Color	0.024	**0.858**
Texture	−0.016	**−0.447**
Size	−0.079	−0.063

The higher the magnitude of the loading for an attribute, the more that attribute describes the component. The sign (+/−) indicates the direction of the relationship of the attribute to the component. The first component is described primarily by Flavor and Thickness. The second component is described primarily by Color and Texture. The percent variance explained by the first principal component was 99.0 % (1.0 % for the second component). Thus, the samples are most differentiated on the first component, which is described most by the concepts of Flavor and Thickness working in opposition to one another.

Using the results from the PCA, a biplot was generated via the biplot macro for Excel.

The importance of Flavor and Thickness in discriminating among the three samples can be seen in the biplot. After examining the mean values for the Flavor attribute, we can determine that Samples 170 and 896 have "Too Much Flavor." Based on the high mean value for the thickness attribute, we can conclude that Sample 914 is "Too Thick" *compared to* Samples 170 and 896.

In this case study, CA and PCA generate biplots that are different from one another, owing mostly to the type of data used in the analyses. PCA relies on interval data (means). Although the mean describes the distribution of interval data completely, it may mask subtleties in the JAR data (which may not be truly interval) for reasons outlined in Appendix D. However, the PCA map is fairly uncluttered and broad differences among the samples may be easily observed.

In contrast, CA uses contingency table data that can completely describe the nuances of JAR data ("Not Enough," "Just About Right," "Too Much"). This creates three times the number of JAR attributes to be plotted on a map. Inter-

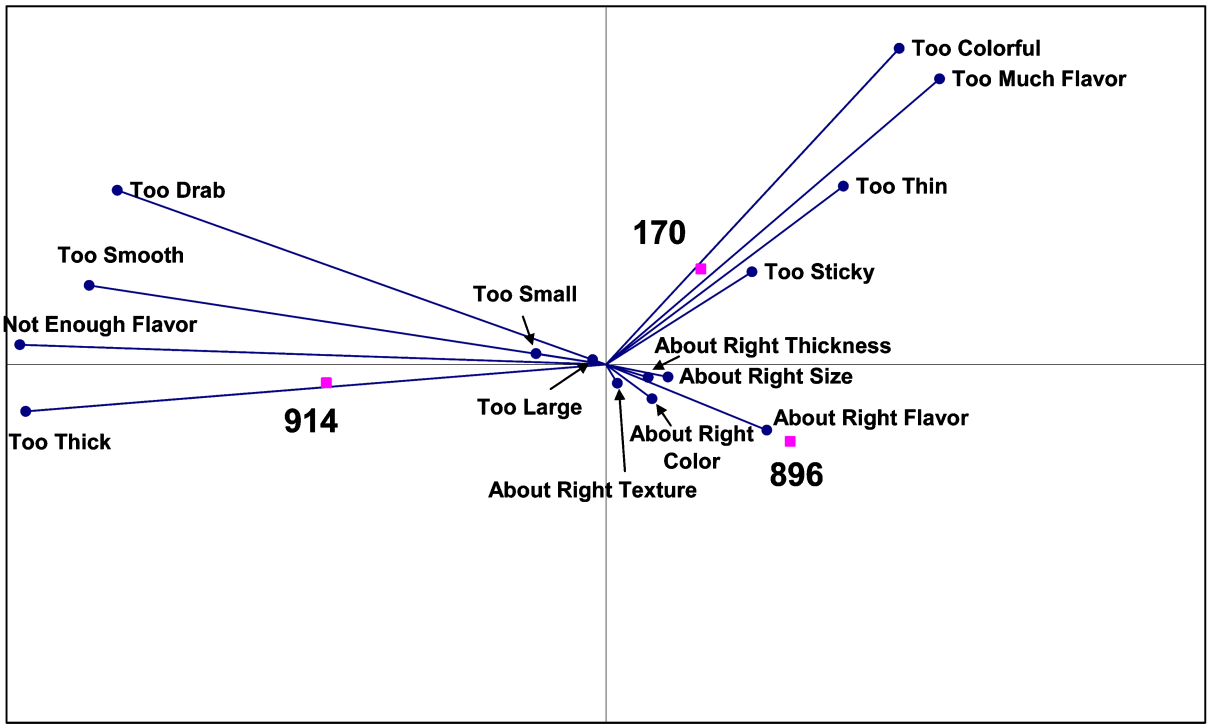

Fig. 1—Correspondence analysis biplot.

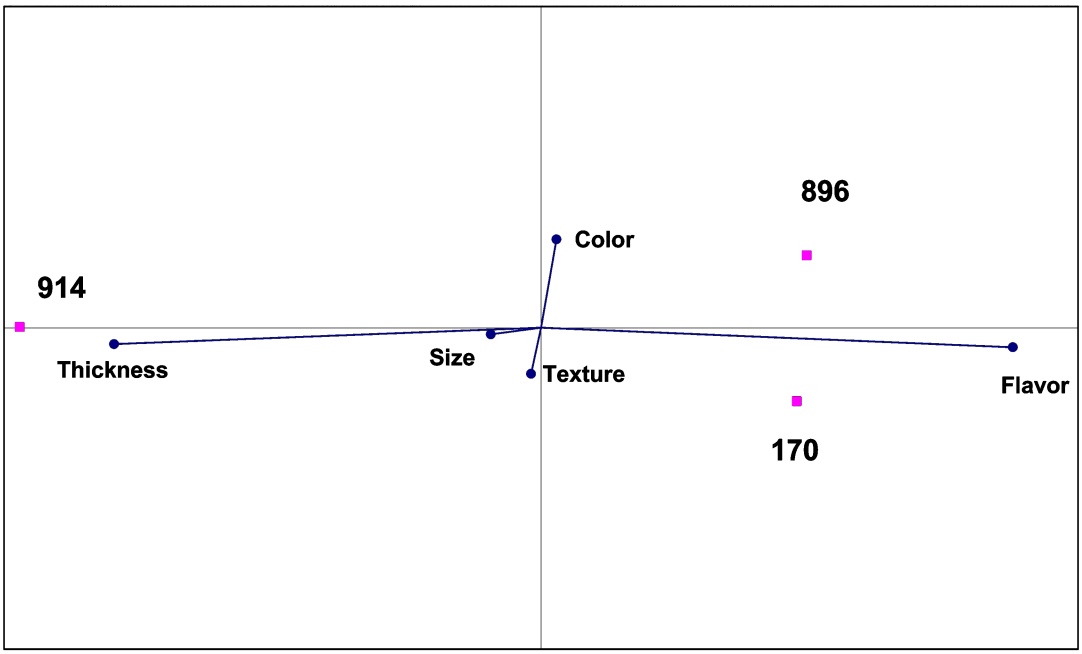

Fig. 2—Principal components analysis biplot.

pretation of relationships in the map becomes more challenging as the number of JAR attributes increases. Plotting only the "Too Much" variables, only the "Not Enough" variables or only the "Just About Right" variables is an alternative to plotting all three variables on one map.

Pros and Cons

The chief benefit to biplots is that each product can be evaluated within the context of other product. In contrast to the results produced by other single or bivariate techniques (e.g., mean, correlation), results yielded by biplot analyses allow researchers to identify products that are perceived similarly and attributes that are more associated with one another. Although most biplot methods yield similar insights, the choice of analytic method often depends on the scaling of the data (PCA for interval data and CA for frequency/contingency table data) and the advantages/disadvantages associated with the particular technique. For example, one of the disadvantages of CA is that the axes do not have a clear meaning. Interpreting axes in PCA biplots can be difficult as well. There also is some disagreement among experts as to whether the relationships among attributes and objects (or samples) are interpretable in CA biplots.

The interpretation of the biplots is limited to the samples and attributes that are included. In other words, the spatial relationships might change as different samples or attributes are involved or even if a different scale is used. Another limitation associated with biplots is legibility. Inclusion of numerous points creates a cluttered plot that can hinder interpretation. Also, the traditional biplot technique is purely descriptive in that it only forms a picture of perceptions and does not attempt to incorporate preferences or causality (i.e., consumers may think that Sample 170 is "Too Colorful," but is it worth reformulating if Sample 170 has a low preference or purchase intent among consumers?)

Recommendation

The use of biplots, especially those based on CA, is recommended as a descriptive tool to understand the relationships between attributes and samples.

References

[1] Lipkovich, I., and Smith, E. P., "Biplot and Singular Value Decomposition Macros for Excel," http://www.stat.vt.edu/facstaff/epsmith.html, Blacksburg, VA, 2001.

Appendix T: Correlation

Amy Takkunen[1]

Introduction and Objectives

Correlation analyses can be used to measure the linear association of JAR scale data to data measured using other types of scales, e.g., liking scales. The goal of these analyses is to assess the strength and direction of the relationship between the JAR scale and the other scale(s) of interest for each product and, possibly, across products.

Requirements

This analysis requires the raw data, arranged in a table with one line per assessor by product combination. The assessors must have evaluated the sample(s) using both the JAR scale(s) and the other scale(s) of interest.

"How to"

The formula for the correlation coefficient r is given below, where SXY = the corrected sum of cross products, SXX = the corrected sum of squares for the Xs (here, the JAR scale scores), and SYY = the corrected sum of squares for the Ys (here, the other scale scores):

$$r = \frac{SXY}{\sqrt{(SXX)(SYY)}}$$

Most statistics programs, as well as Microsoft Excel, include procedures to calculate this statistic. This statistic has $n - 2$ degrees of freedom, where n is the number of assessors. If the calculated r-value is larger than the correlation table r-value for the chosen alpha level, the correlation between scales is considered to be statistically significant. Correlation coefficients range –1 and +1. Positive values of r indicate that as the JAR score goes up, so does the score on the other scale. Negative r-values indicate that as the JAR score goes up, the score on the other scale goes down. An r close to 0 means that there is no relationship between the scales. The closer an r is to –1 or +1, the stronger the linear relationship between the scales.

Case Study Example

This example uses the JAR scale and the liking scales for Products 170, 896, and 914. Table 1 displays the correlation and p-value of each JAR scale with its corresponding liking scale within each product. Table 2 repeats that analysis, pooling across products.

Results and Conclusions

Data in Table 1 indicate that, for Product 170, there are no significant linear relationships between JAR scales and liking scales. For Product 896, there is a weak but significant negative linear relationship between JAR color and overall liking, and JAR flavor and Overall Liking. For Product 914, there is a weak but significant positive relationship between JAR color and color liking, and between JAR flavor and flavor liking. There is also a somewhat stronger significant positive relationship between JAR flavor and Overall Liking.

Data in Table 2 indicate that, across products, there is a significant positive relationship between JAR size and size liking, and between JAR color and color liking.

Pros and Cons

Calculating correlation coefficients allows the researcher to evaluate the strength of a linear relationship between a JAR scale and other scales used by the same assessors on the same products.

A limitation of this analysis is its use of a linear relationship when the relationship between the JAR scale and the Liking scale is expected to be highest in the middle of the scale and lowest at the ends. It is not hard to demonstrate that there can be a perfect association between the JAR and liking that has a correlation of zero. This problem can be circumvented by alternate encodings of the JAR scale (e.g., change {1,2,3,4,5} to {–2, –1, 0, –1, –2}) or by breaking the JAR scale up into two scales, each of which is unidirectional (i.e., "Too Weak" to "Just About Right" and "Just About

TABLE 1—Correlation of JAR scales with other scales by product.

	Product					
	170		896		914	
Scales Compared	*r*	*p*	*r*	*p*	*r*	*p*
JAR Size, Liking Size	–0.15	NS	–0.06	NS	–0.04	NS
JAR Color, Liking Color	0.04	NS	0.09	NS	0.23	0.02
JAR Flavor, Liking Flavor	0.07	NS	–0.12	NS	0.23	0.02
JAR Size, Liking Overall	–0.04	NS	–0.03	NS	–0.03	NS
JAR Color, Liking Overall	0.02	NS	–0.22	0.02	0.02	NS
JAR Flavor, Liking Overall	0.05	NS	–0.19	0.05	0.33	0.001

NS = Not significant

[1] *General Mills, 9000 Plymouth Avenue North, Minneapolis, MN 55427.*

TABLE 2—Correlation of JAR scales with other scales.

Scales Compared	*r*	*p*
JAR Size, Liking Size	0.99	<0.01
JAR Color, Liking Color	0.99	<0.01
JAR Flavor, Liking Flavor	–0.78	NS
JAR Size, Liking Overall	–0.56	NS
JAR Color, Liking Overall	–0.41	NS
JAR Flavor, Liking Overall	–0.64	NS

NS=Not significant

Right" to "Too Strong," but these should be considered carefully.

Recommendation

Correlation analysis is not recommended unless it has been demonstrated (via graphical examination of liking versus JAR data or other analysis) that the relationship between the JAR scale and liking is linear). Where such data are not available, it is recommended to re-code the scale, as discussed above.

Appendix U: Regression

Joseph E. Herskovic[1]

Background and Introduction

Many of the other techniques in this guideline treat the effect of JAR scales on the overall rating one scale at a time. Regression analysis allows the researcher to evaluate the joint effects of the scales levels on overall response. The stronger the relationship between a JAR scale and the overall response, the more important that "Just-Right" attribute is in explaining the liking attribute, even after controlling for the other attributes.

The regression can be either non-parametric (ordinal) or parametric (ordinary regression) and the JAR scales can have either a linear or non-linear effect on the response. The examples below use linear regression because of ease of use and widespread availability in many statistical packages. The more general approaches require more statistical sophistication.

Regression analysis can be done for the entire data set (all samples combined) or for each individual sample. Conducting the analysis on all samples combined gives a general overview of how the "Just Right" attributes work together to explain liking. This analysis can be conducted using either the individual respondent data or product mean scores.

Requirements

These approaches generally require the individual level data for the overall rating scale and the attributes of interest.

Example Analysis

All data from example data set were used in these analyses. Both an overall model and single sample analyses were fitted. For simplicity, the examples did not include terms to adjust for the repeated measures on each panelist. This means that the significance tests are rather conservative, and may miss some significant effects. In each case a stepwise linear regression approach was used to select the terms to be included in the models.

Results and Conclusions

All Samples Combined

For the current data set, the regression of Overall Liking on the five attributes of Size, Color, Amount of Flavor, Thick/Thin, and Stickiness was fit. Note that Overall Liking is a 9-point scale and all scales on the right side of the equation are 5-point "Just About Right" scales. The input contains main effects only and is based on the raw data (not means). The output is summarized in the following printout:

*** Stepwise Regression ***

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	8.8216	0.6614	13.3372	0.0000
JAR.Flavor	−0.3615	0.1235	−2.9278	0.0036
JAR.Stickiness	−0.5656	0.1814	−3.1173	0.0019

Residual standard error: 1.946 on 505 degrees of freedom
Multiple R-Squared: 0.03731
F-statistic: 9.785 on 2 and 505 degrees of freedom, the p-value is 0.00006771

Of the five attributes entered into the equation, only two had a significant effect on overall liking: Flavor (negative) and Stickiness (negative). Thus, it is recommended that the researcher further investigate these attributes and pay less attention to the other three (Size, Color, and Thickness). Note that the sign (positive or negative) yields clues as to how to reformulate. If the sign is positive, then in these product samples more is better (up to a certain point, of course). Likewise, If the sign is negative, then less is better.

Individual Product Models

The above analysis was done without segmenting by sample. Often the researcher wants to know the importance of sensory attributes on specific test samples. This involves the same stepwise regression analysis separately for each sample of interest.

Sample 170

Coefficients:

	Value	Std.Error	t value	Pr(>|t|)
(Intercept)	8.1163	1.4669	5.5328	0.0000
JAR.Stickiness	−0.8277	0.4625	−1.7895	0.0766

Residual standard error: 2.173 on 100 degrees of freedom
Multiple R-Squared: 0.03103
F-statistic: 3.202 on 1 and 100 degrees of freedom, the p-value is 0.07656

Comments: Only Stickness has a significant (negative) effect on overall liking Note that the coefficient (−0.8) is both larger and less reliable than that Stickiness coefficient in the overall regression.

[1] *Sensory ConAgra Foods, Inc. Six ConAgra Drive Omaha, NE 68102-5094.*

Sample 458

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	9.6205	1.2640	7.6114	0.0000
JAR.Flavor	−0.7467	0.2229	−3.3497	0.0011
JAR.Stickiness	−0.5270	0.2958	−1.7815	0.0779

Residual standard error: 1.645 on 98 degrees of freedom

Multiple R-Squared: 0.1189

F-statistic: 6.614 on 2 and 98 degrees of freedom, the p-value is 0.00202

Comments: As with the joint model, only Flavor and Stickiness have significant effects on Overall Liking. As with the previous example, the coefficients are both larger and less reliable than the joint regression.

The researcher can continue conducting this analysis for each of the samples to determine the most important sensory attributes to investigate.

From the regression data, we can determine that product liking is most sensitive to the amount of flavor and the stickiness of the products. The other attributes, Size, Color, and Thickness, did not have a strong effect after controlling for Flavor and Stickiness (within the range tested here) and should be of secondary importance in product development efforts.

Discussion

As was demonstrated here, it can be the case that only a few of the attributes have a direct effect on the output, while the rest of the attributes have little impact once the direct attributes have been accounted for. This implies that the effects of these other attributes may be indirect and may be realized through their impact on the other attributes. Untangling "what causes what" requires some substantive understanding of the attributes and product use in the field.

Two ways of analyzing the data were demonstrated here: the entire data set and each individual sample, both using raw data. A more thorough analysis would include a test for product × attribute interactions to determine if there is any benefit in evaluating individual product models. This analysis also treats the center point of a JAR scale (the "Just-About-Right" point) as another point on the intensity continuum, affording it no special significance.

There are other ways to analyze these data. For example, one can use "indicator" variables to evaluate the joint effect of deviating from the JAR values. Additionally, as mentioned earlier, a more careful analysis could include dummy variables to capture some of the individual variability in using the scales. Further, the response can be treated as an ordinal response, dropping the requirement that the panelist have an interval-level response.

Pros and Cons

Regression measures provide the benefit of simultaneous analysis of all product attributes, resulting in an understanding of those that have the most impact on overall liking. Regression analysis provides predicted overall liking ratings based on JAR scale ratings. Curvilinear relationships can be modeled using these techniques, a distinct over correlational analysis. Regression techniques are widely available in most software packages.

Regression techniques assume that the data are unimodal and provide interval-level information; in practice, these assumptions may not be true. While widely available, these analyses require some statistical sophistication on the part of the analyst.

Recommendation

It is recommended that regression analysis be considered when there are multiple attributes that can affect the overall response and the research is interested in untangling which attributes have direct effects on the response and which do not.

Appendix V: Preference Mapping from JAR Data Using Tree-Based Regressions

Jean-Francois Meullenet[1] and Rui Xiong[1]

Background and Objectives

Penalty analysis offers a method to consider the individual effects of JAR ratings on Overall Liking (OAL), but does not provide a way to assess the impacts of simultaneous changes in JAR ratings on Overall Liking. Standard multiple regression is of limited use in this situation because of its strong assumptions of linearity.

A form of non-parametric regression, which we will refer to as "tree-based" regression, removes that assumption and allows you to determine the combinations of the JAR ratings that have the strongest impact on Overall Liking.

There are wide variety of "tree-based" regressions packages available, such as CART, MARS, KnowledgeSeeger, and SPSS AnswerTree, as well as free implementations such as part in R. This example will use MARS (multivariate adaptive regression splines) as its example [1]. This is commercial software, sold by Salford Systems (http://www.salford-systems.com/) [2].

Requirements for the Analysis

This analysis requires raw data arranged in a table, with one row for each rater by product combination. Each row should have both the liking and JAR ratings for a particular product evaluation.

"How to"

This is a computer intensive procedure and we will only give an overview of the method. Tree-based functions typically proceed by examing each of the predictors in the whole set in turn. For each variable it will use each of its levels to find a split that will make the resulting subgroups most different. It keeps the best split overall of the variables and recursively repeats the process on each of the subgroups until the resulting subgroups are too small. Many of the tree programs fit each subgroup with a simple mean, but MARS goes further and fits a linear regression on the splitting variable within the subgroups.

A general MARS model for a single response and a vector of predictors may take the following form:

$$Y = \beta_0 + \beta_1 \mathrm{BF}_1(X) + \beta_2 \mathrm{BF}_2(X) + \cdots + \beta_M \mathrm{BF}_M(X) + \varepsilon$$

where Y is the response variable (e.g., Overall Liking) X is the vector of predictors (such as Size, Color, Flavor, Salt, Thin/Thick, Stickiness, etc), BF_k denotes the kth basis function, a function of all of the splits that lead to one of the final subgroups, and M is the number of basis functions included in the final model. The regression coefficients β_k are estimated by minimizing the sum of squared residuals ε. This can be used on either individual products or on multiple products.

Example From Case Study Data

In this example, MARS was applied to each of the Products 170, 458, 596, 823, and 914 individually. This provides an optimum regression for each product's Overall Liking (OAL) individually

Results and Conclusions

For Sample 170, the "best fit" MARS regression equation was as follows:

$$\mathrm{OAL} = 5.515 \quad (R^2 = 0, n = 101)$$

The coefficient of determination $R^2 = 0$ means that there was no predictive relationship between the JAR variables and OAL, indicating that the JAR attributes (Color, Size, Flavor, Thin/Thick, Stickiness) in this case did not significantly affect the OAL. The predicted OAL mean was 5.5150, which was actually identical to the observed mean (OAL = 5.5149). This sample was the second least preferred product, implying that the JAR attributes tested were unable to explain the low OAL scores obtained for this product.

For Sample 458, MARS gave the following "best fit" regression equation:

$$\mathrm{OAL} = 6.048 - 0.946\mathrm{BF}_1 - 1.197\mathrm{BF}_2 \quad (R^2 = 0.2, n = 101)$$

where the two basis functions were $\mathrm{BF}_1 = \max(0, \text{Flavor-3.0})$ and $\mathrm{BF}_2 = \max(0, 3.0\text{-Thin/Thick})$ ($\max(x_1, x_2)$ is interpreted as the maximum value of the two elements x_1 and x_2). $\mathrm{BF}_1 = \max(0, \text{Flavor-3.0})$ split the flavor JAR scale at 3 (the JAR score) into two scale regions: region of 1 to 3 and region of 3 to 5. BF_1 is constant over the region of 1 to 3, but linearly increased over the region of 3 to 5. Since the regression coefficient (–0.946) is negative OAL score decreased by 0.946 per unit change in BF_1. This relationship between the flavor and OAL is shown in Fig. 1(*a*). This indicates that a "Too Strong" flavor was more detrimental to the OAL than a "Too Weak" flavor. Similarly, $\mathrm{BF}_2 = \max(0, 3.0\text{-Thin/Thick})$ split the JAR scale into two regions (Fig. 1(*b*)). Over the region of 3 to 5, the effect of Thin/Thick on OAL is roughly constant, but the OAL scores decreased at a rate of 1.197 (–1.197 was the regression coefficient for BF_2) as the Thin/Thick scores increased from 1 to 3. This meant that being "Too Thin" was more detrimental to OAL than being "Too Thick." Other JAR attributes (Color, Size, and Stickiness) had no predictive effects on OAL. The observed and predicted means of Overall Liking scores were 5.4653 and 5.4650, respectively. Since the regression intercept (6.048) could be interpreted as the potential maximum OAL mean score if all the attributes were JAR, the difference (0.583) between the regression intercept (6.048) and the predicted mean (5.4650) of the OAL scores can be explained as the average potential improvement in OAL scores if the thin/thick and flavor were adjusted to be "Just About Right."

[1] *Department of Food Science, University of Arkansas, Fayetteville, AR 72704.*

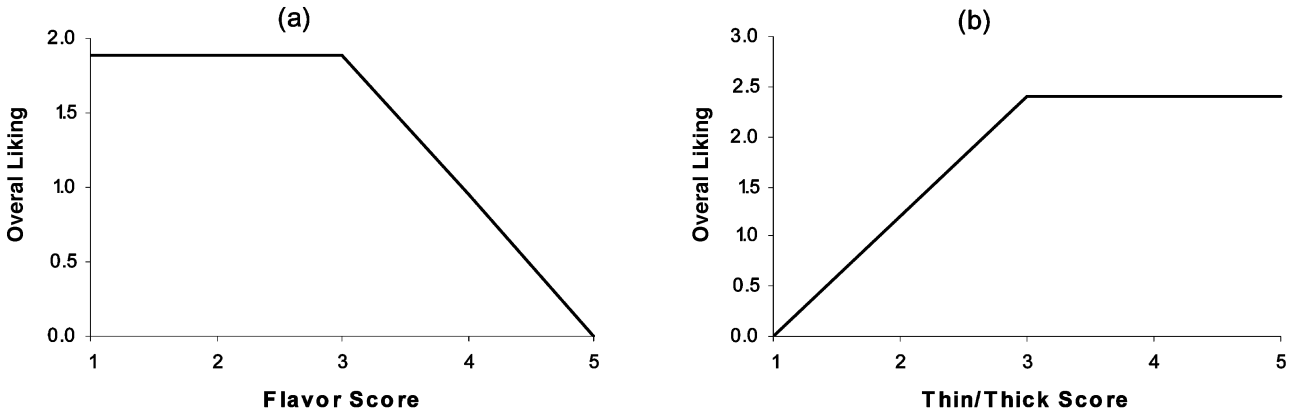

Fig. 1—Contribution to overall liking from flavor (a) and thin/thick (b).

For sample 523, MARS estimated the best regression equation to be:

$$\text{OAL} = 6.445 - 1.024\text{BF}_1 - 1.1587\text{BF}_2 \quad (R^2 = 0.22, n = 102)$$

where the two standard basis functions were $\text{BF}_1 = \max(0, \text{Flavor-3.0})$ and $\text{BF}_2 = \max(0, \text{Stickiness-3.0})$. The predicted relationships between OAL and Flavor/Stickiness are displayed in Fig. 2. It is evident from the figure that both flavor and stickiness significantly decreased OAL over the region of 3 to 5 in the rates of 1.024 and 1.159, respectively, but did not have a significant influence on OAL over the region of 1 to 3. This suggested that reducing the intensity of flavor and stickiness from being "Too High" to being JAR would increase average consumer OAL score up to 6.445. The observed and predicted means of Overall Liking scores were 6.0 and 5.991, respectively.

For Sample 896, the "best fit" MARS regression equation was as follows:

$$\text{OAL} = 7.949 - 1.446\text{BF}_1 \quad (R^2 = 0.01, n = 102)$$

where one standard basis functions was $\text{BF}_1 = \max(0, \text{Stickiness-2.0})$. The relationship between OAL and Stickiness is presented in Figure 3(*a*). The figure shows that BF_1 split the Stickiness scale at the point of 2. This means that the OAL mean score was lower at the JAR score of 3 than at the score of 2, which was not expected. This deviation from the JAR score of 3 could be due to the noise in data, so two MARS models were fitted separately over the regions of 1 to 3 and the region of 3 to 5 for stickiness. It was found that stickiness had no predictive value for OAL over the region of 1 to 3, but had a significantly negative impact on OAL over the region of 3 to 5. By combining these two models, the following regression equation was obtained:

$$\text{OAL} = 6.523 - 1.52\text{BF}_1 \quad (R^2 = 0.01, n = 102)$$

where the only standard basis functions was $\text{BF}_1 = \max(0, \text{Stickiness-3.0})$. The relationship between OAL and the JAR variable is illustrated in Fig. 3(*b*). The observed and predicted means of overall liking scores were 6.3039 and 6.3144, respectively. Results like those in Fig. 3(*a*) could occur if there is significant noise in the data; consumer scoring of the sample is not consistent and/or the JAR variables are highly correlated.

For sample 914, the "best fit" MARS regression equation was as follows:

$$\text{OAL} = 7.730 - 1.638\text{BF}_1 - 2.928\text{BF}_2 - 2.305\text{BF}_3 - 1.202\text{BF}_4 \quad (R^2 = 0.35, n = 100)$$

where the four standard basis functions were $\text{BF}_1 = \max(0, \text{Thin/Thick-3.0})$, $\text{BF}_2 = \max(0, \text{3.0-Thin/Thick})$, $\text{BF}_3 = \max(0, \text{Flavor-3.0})$ and $\text{BF}_4 = \max(0, \text{3.0-Flavor})$. The observed and predicted means of Overall Liking scores were 6.5600 and 6.5603, respectively. The relationships between

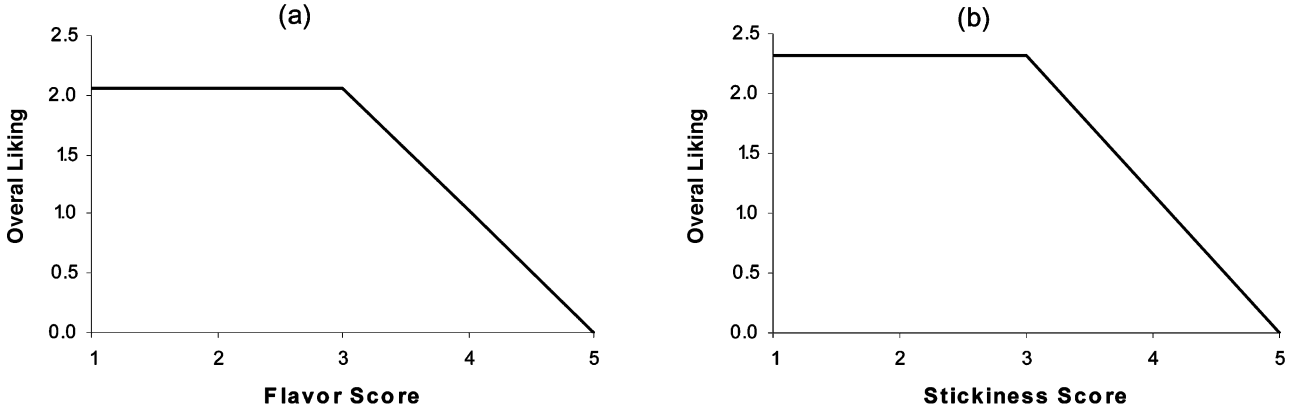

Fig. 2—Contribution to overall liking from flavor (a) and stickiness (b).

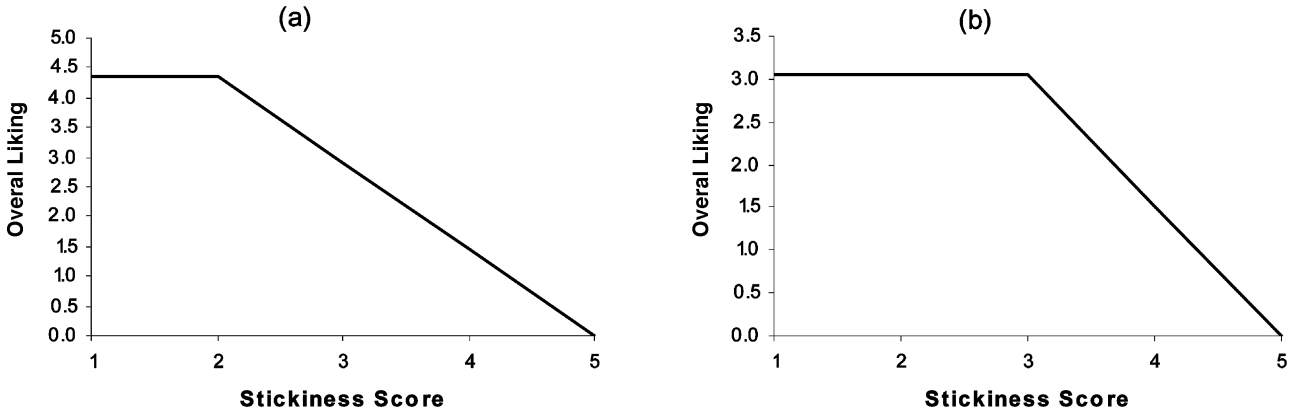

Fig. 3—Contribution to overall liking from stickiness: (a) with one model fit and (b) with two-models fit.

OAL and thin/thick and flavor are given in Fig. 4. The OAL mean was the highest at the JAR score of 3 and decreased over either side of the JAR score. This figure suggests that there was a disagreement between two segments of consumers. Some consumers found the product to be too thin while others found it to too thick. However, being too thick seemed slightly more detrimental to OAL than being too thin. Similarly for Flavor, one group of consumers found the flavor to be too strong while another group found the flavor to be too weak. Overall, a weak flavor had a more negative impact on OAL than a too strong flavor.

A comparison of all the samples was done by tabulating the results from the above data analysis for each individual product (Table 1). The table clearly shows that size and color did not significantly contribute to OAL scores. To compare consumer preference to the samples, the plot of observed means versus predicted means of OAL scores for all the samples are presented in Fig. 5. The predicted means were nearly identical to the observed mean with $R^2 = 0.99$. It was apparent that the overall acceptance order of the samples was $914 \geqslant 896 \geqslant 523 \geqslant 170 \geqslant 458$. Of the five samples, Sample 914 was the most liked sample, whereas Sample 458 was the least liked. However, no useful information was contained in the JAR data to explain the weaknesses of Sample 170. The potential improvements varied from sample to sample. Sample 914 had the largest potential improvement of 1.170, which means that the OAL mean score could be potential im-

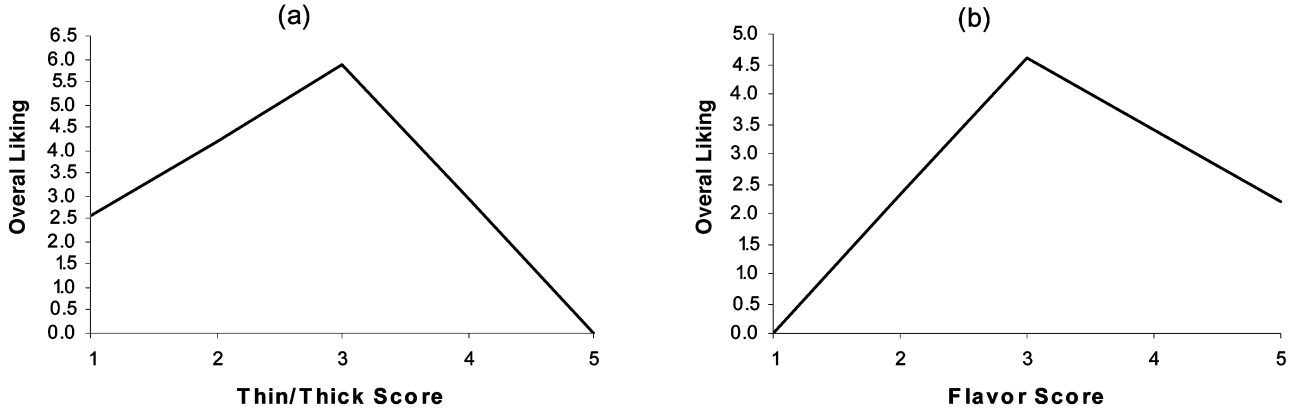

Fig. 4—Contribution to overall liking from think/thick (a) and flavor (b).

TABLE 1—Comparison of the samples.

Sample	Size	Color	Flavor	Thin/Thick	Stickiness	Observed OAL Mean	Predicted OAL Mean	Potential Improvement
170	○	○	○	○	○	5.5149	5.5150	0
458	○	○	×	×	○	5.4653	5.4650	0.583
523	○	○	×	○	×	6.0000	5.9909	0.454
896	○	○	○	○	×	6.3039	6.3144	0.209
914	○	○	×	×	○	6.5600	6.5603	1.170

○ stands for no predictive effect on OAL
× stands for detrimental effect on OAL due to not being JAR
Potential improvement is the potential improvement in OAL mean score by adjusting contributing JAR attributes to JAR levels

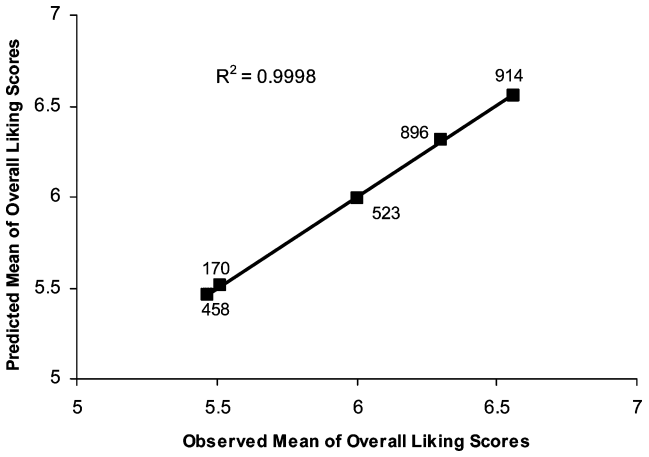

Fig. 5—Comparison of observed and predicted means of overall liking scores.

proved by 1.170 by improving flavor and stickiness to JAR levels.

Conclusions from the Analysis

- According to MARS, the JAR attributes (Size, Color, Flavor, Thin/Thick, Stickiness) evaluated for sample 170 did not seem to explain the OAL scores. This sample had the second lowest OAL mean score of all the five samples tested.
- Sample 458 was found to have only two attributes (Flavor and Thin/Thick) not being JAR that contributed to determining OAL. Too strong flavor and being too thin were detrimental to OAL. It was the least preferred sample.
- Sample 523 had only two attributes (Flavor and Stickiness) not being JAR that contributed to determining OAL. A "Too Strong Flavor" and being "Too Sticky" had negative impacts on OAL.
- Sample 896 had only one attribute (Stickiness) not being JAR that contributed to lowering OAL and it was the second most preferred sample. Being too sticky decreased OAL.
- Sample 914 had two out of five attributes (Thin/Thick and Flavor) not being JAR and it was the most liked sample. Consumers tended to disagree about the product weaknesses.

Pros and Cons

The tree-based regression approaches, such as MARS analysis, allows the researcher to examine the joint effect of the JAR ratings on Overall Liking. The MARS regression intercept estimates the maximum OAL mean that a sample could achieve if the attributes were adjusted to JAR levels.

The limitations of these approaches are three-fold. These include conceptual complexity of the results, the requirement for specialist software, often expensive, and the need for larger sample sizes. Additionally, as presented, the program treats the JAR scales as being continuous, rather than ordinal.

Recommendation

This method is recommended for JAR scale regression analysis when the researcher wishes to understand the joint action of multiple attributes on "Overall Liking" under conditions that the independent variables are not correlated with each other and the number of independent variables is larger than the number of observations. Additionally, it is recommended to analyze the products individually.

References

[1] Friedman, J., "Multivariate Adaptive Regression Splines," *Annu. Stat.*, Vol. 19, 1991, pp. 1–141.

[2] MARS, *MARS User's Guide*, Salford Systems, 2001.

Appendix W: Application of JAR Data to Preference Mapping Using Dummy Variables

Rui Xiong[1] and Jean-Francois Meullenet[1]

Objectives

The dummy variable approach is a regression method that can be applied to map relationships between JAR and liking scores. The approach models the relationship between JAR and liking variables, estimates the mean drop of Overall Liking (OL) as a function of JAR scores and determines the drivers of liking. The dummy variable approach is a very flexible method that can be performed either for individual products or for all products to determine the effects of JAR attribute scores on Overall Liking.

Requirements for the Analysis

Individual respondent data for each attribute/product combination are required and the number of observation must be at least twice as large as the number of predictive variables. A statistical package that implements regression and partial least-squares (PLS) models is required.

"How to"

Penalty analysis is a graphical technique that provides a list of critical product characteristics that most negatively impact product liking. However, penalty analysis has many pitfalls. It is not a regression-based method, ignores correlations among product characteristics, and cannot be used to predict consumer overall acceptance from JAR data. In addition, the mean drop estimated by penalty analysis for a specific attribute is not an estimate of the "true" mean drop on overall liking. A dummy variable approach with two models (analysis of covariance or partial least-squares regression) is proposed as an extension to penalty analysis.

Regression Analysis with Dummy Variables

Regression analysis with dummy variables is a method that subsumes both ordinary regression and analysis of variance. It does this by transforming the each JAR rating, with, say, k categories into up to k indicator or dummy variables, preferably independent. As an example, consider the following encoding for a $k = 3$ JAR scale (see Table 1).

This example converts a three-level JAR scale into three nominally independent (uncorrelated) variables (Z_1–Z_3). The dummy variable Z_1 captures the usual mean. The dummy variable Z_2 compares the JAR value with the average of the non-JARs. If this variable is used to predict a liking variable, It would measure the difference between liking when the attribute is at the JAR value and liking is not at the JAR value (the average penalty). Finally Z_3 compares the liking when the attribute is "Too High" with the liking when the attribute is "Too Low." When used as a response in a regression model comparing several products or attributes, the first column is constant and non-informative, the second column compares the drops, and the third column compares the asymmetry.

To use this technique, the analyst needs to recode the data to explode each of the JAR variables into a larger set of dummy variables. This typically requires some programming to perform. These dummy variables are subsequently supplied to a regression program. As mentioned earlier, we are using regression in the general sense, which can include ordinary least-squares, partial least-squares, multivariate least-squares, and ordinal regression (proportional odds, probit, and proportional hazards).

It is used to assess the statistical significance of mean differences among treatment groups with an adjustment made for initial differences on one or more covariates. When analysis of covariance applies to relating Overall Liking to JAR variables, the covariance analysis model is expressed as

$$Y_{ij} = \mu_i + \sum_{k=1}^{p} \beta_{ik} X_{ijk} + \eta_{ij} + \varepsilon_{ij} \quad (i = 1,2 \ldots ,t; j = 1,2 \ldots ,m) \tag{1}$$

where the response variable Y_{ij} is the Overall Liking score given by the jth consumer for the ith product; μ_i is the mean of the response variable for the ith product; independent variable X_{ijk} is the JAR score given by the jth consumer for the kth JAR variable of the ith product; t, m, and p are the numbers of products, consumers, and JAR variables used in the test, respectively; β_{ik} is the regression coefficient for the kth JAR variable and the ith product; η_{ij} and ε_{ij} are the random effect term and residual for the jth consumer and the ith product, respectively. This model (Eq (1)) is unable to correctly describe the non-linear relationship between the response variable (Y_{ij}) and covariates ($X_{ij1}, X_{ij2}, \ldots, X_{ijp}$) because the JAR scale has its "best/ideal" score in the middle (Fig. 1(*a*)) of the scale. Take a 5-point JAR scale (1 to 5) as an example. The best/ideal score, also called the JAR score, is 3. As the scores of a JAR variable (X_{ijk}) are away from the JAR score over both sides/regions ("Too Little" region and "Too Much" region), consumer acceptance scores (Y_{ij}) would be expected to drop or stay constant (Fig. 1). It is possible that the drop rates over both regions of the JAR score may be different. To describe this phenomenon, two dummy variables (Z_{ijk1} and Z_{ijk2}) are introduced to represent each original JAR variable (X_{ijk}). The presentation scheme is provided in Table 1. Over the region (1 to 3) of X_{ijk}, Z_{ijk1} changes from –2 to 0, and Z_{ijk1} is 0 over the region (3 to 5). In contrast, Z_{ijk2} is 0 over the region (1 to 3) and changes from 0 to 2 over the region (3 to 5). If the drop rates over the two regions are the same, Z_{ijk1} and Z_{ijk2} are combined into a single dummy variable $Z_{ijk} = Z_{ijk1} - Z_{ijk2}$ to represent X_{ijk}.

Using dummy variables instead of the original JAR variables, the new model is given by

[1] *Department of Food Science, University of Arkansas, Fayetteville, AR 72704.*

TABLE 1—Example dummy variables for a 3-point JAR scale.

Response	Z_1	Z_2	Z_3
"Too Low"	1/3	−1/2	−1
"Just About Right"	1/3	1	0
"Too High"	1/3	−1/2	1

$$Y_{ij} = \mu_i + \sum_{k=1}^{p} (\alpha_{ik1} Z_{ijk1} + \alpha_{ik2} Z_{ijk2}) + \eta_{ij} + \varepsilon_{ij}$$
$$(i = 1,2 \ldots, t; j = 1,2 \ldots, m) \quad (2)$$

where the pair (Z_{ijk1}, Z_{ijk2}) are dummy variables for the original JAR variable X_{ijk}; α_{ik1} and α_{ik2} are regression coefficients for the kth pair of dummy variables, respectively. For each pair of regression coefficients $(\alpha_{ik1}, \alpha_{ik2})$, the sign (+) of α_{ik1} must be opposite to the sign (−) of α_{ik2}, as shown in Fig. 1(*b*). The same signs of α_{ik1} and α_{ik2} indicate that the original JAR variables $(X_{ij1}, X_{ij2}, \ldots, X_{ijp})$ are highly correlated with one another or there is noise in the data. As such, this covariance analysis model (2) is appropriate for mapping relationships between JAR and liking variables only if the original JAR variables $(X_{ij1}, X_{ij2}, \ldots, X_{ijp})$ are independent of each other (Fig. 1(*b*)). Correlation coefficients can be used to check the independence between JAR variables. A stepwise method (such as backward elimination) can apply to this model for selection of important variables. If Eq (2) is appropriate, the term $(\alpha_{ij1} Z_{ijk1} + \alpha_{ij2} Z_{ijk2})$ is always either zero or negative. A zero mean of the terms $(\alpha_{ij1} Z_{ijk1} + \alpha_{ij2} Z_{ijk2})$ $(j = 1, 2, \ldots, m)$ across all consumers indicates that there is no significant relationship between the kth JAR variable and the response (OL), whereas a negative mean can be explained as the estimate of the mean drop on Overall Liking due to the kth JAR variable not being JAR for the ith product. Since Eq (2) includes all dummy variables involved, it is usually called the full model. Based on the full model, many hypotheses can be formed and tested using an F-test. For example, if it is assumed that the effect of each dummy variable on the response is the same (common slope) for all products, then the model (Eq (2)) becomes

$$Y_{ij} = \mu_i + \sum_{k=1}^{p} (\alpha_{k1} Z_{ijk1} + \alpha_{k2} Z_{ijk2}) + \eta_{ij} + \varepsilon_{ij}$$
$$(i = 1,2 \ldots, t; j = 1,2 \ldots, m) \quad (3)$$

This model (Eq (3)) is called the reduced model because it contains only a subset of the variables used in the full model. The F-test is used to determine which model (full model or reduced model) fits the data. The testing hypotheses are given below:

H_0: the reduced model (Eq (3) in this case) fits

H_a: the full model (Eq (2) in this case) fits

Reject H_0 if the calculated $F > F_{\alpha, t-1, N-t}$, where the F-value is calculated by

$$F = \frac{(SSE_R - SSE_F)/(DF_R - DF_F)}{SSE_F/DF_F} \quad (4)$$

where SSE_R and SSE_F are the sum of squares of errors for the reduced and full models, respectively; DF_R and DF_F are the degrees of freedom for the reduced and full models, respectively. Similarly, an F-test can be used to test if a pair or some pairs of dummy variables have the same effect on the response for all products.

For a JAR scale, it is usually important to determine if the drop rates over two JAR regions ("Too Little" region and "Too Much" region) are the same, which forms the following testing hypotheses:

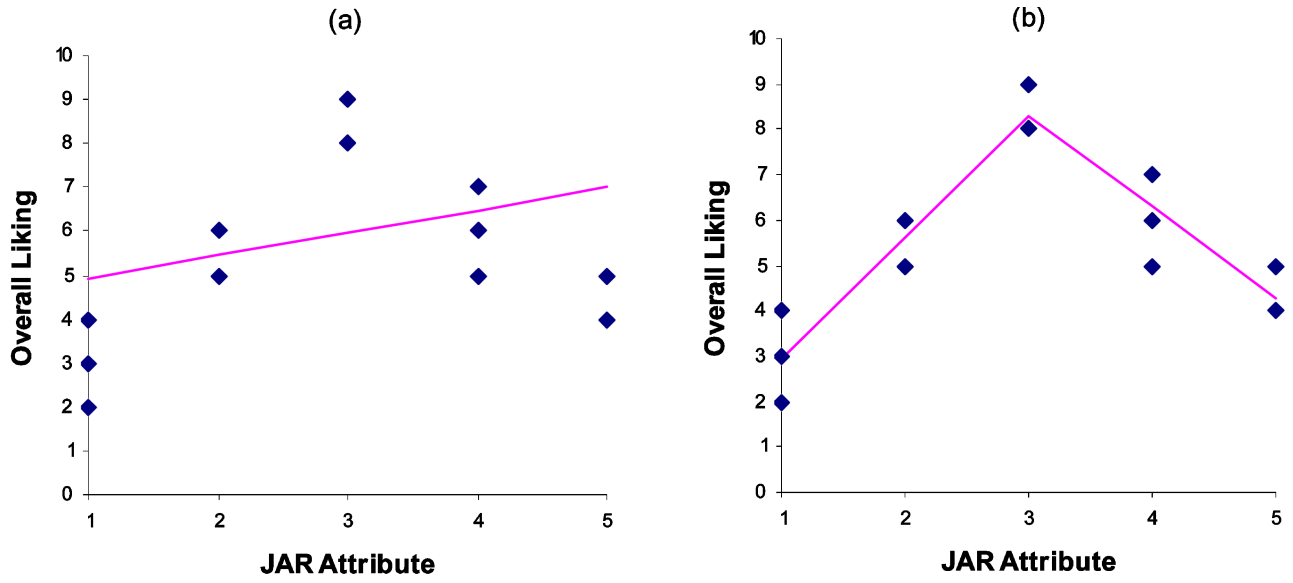

Fig. 1—Examples of linear regression models using the original variable (a) and dummy variables (b) to map relationships between JAR and hedonic scores.

$$H_0: \alpha_{ik1} = -\alpha_{ik2}, \text{ or } |\alpha_{ik1}| = |\alpha_{ik2}| \text{ for all pairs of dummy variables}$$

$$H_a: \alpha_{ik1} \neq -\alpha_{ik2}, \text{ or } |\alpha_{ik1}| \neq |\alpha_{ik2}| \text{ for at least some pairs of dummy variables}$$

Under the null hypothesis (H_0), the model (Eq (2)) can be simplified as

$$Y_{ij} = \mu_i + \sum_{k=1}^{p} \alpha_{ik1}(Z_{ijk1} - Z_{ijk2}) + \eta_{ij} + \varepsilon_{ij} = \mu_i + \sum_{k=1}^{p} \alpha_{ik1} Z_{ijk} + \eta_{ij} + \varepsilon_{ij} \quad (i = 1,2 \ldots, t; j = 1,2 \ldots, m) \tag{5}$$

where $Z_{ijk} = Z_{ijk1} - Z_{ijk2}$ (see Table 1). By using the above F-test (Eq (4)), the full (Eq (2)) and reduced (Eq (5)) models can be compared to determine if the two drop rates over the two JAR regions are the same for all pairs of dummy variables simultaneously. Similarly, an F-test can be used to test if the drop rates are the same only for a pair or some pairs of dummy variables or if all products have the same means. If other factors (such as gender, age, etc.) are of interest, they can be added to the models. Once the final model is determined, the mean score of Overall Liking for each individual product can be predicted by

$$\bar{Y}_i = \mu_i + \sum_{k=1}^{p} (\alpha_{ik1}\bar{Z}_{ik1} + \alpha_{ik2}\bar{Z}_{ik2}) \quad (i = 1,2 \ldots, t) \tag{6}$$

where $\bar{Y}_i$ is the predicted mean score of Overall Liking for the ith product; $\bar{Z}_{ik1}$ and $\bar{Z}_{ik2}$ are the means of dummy variables Z_{ijk1} and Z_{ijk2} across consumers for the ith product and the kth JAR variable, respectively; $\alpha_{ik1}\bar{Z}_{i.k1}$ and $\alpha_{ik2}\bar{Z}_{ik2}$ are defined as the estimates of the mean drop on Overall Liking due to the kth JAR variable being "Too Little" and being "Too Much," respectively. As was pointed out previously, $\sum_{k=1}^{p}(\alpha_{ik1}\bar{Z}_{ik1} + \alpha_{ik2}\bar{Z}_{ik2}) \leq 0$ is always true if the model fits the data appropriately. For the ith product, $\sum_{k=1}^{p}(\alpha_{ik1}\bar{Z}_{ik1} + \alpha_{ik2}\bar{Z}_{ik2}) = 0$ holds only if $\bar{Z}_{ik1} = \bar{Z}_{ik2} = 0$ for all k ($k = 1,2,\ldots,q$), which means that all attributes of the ith product are JAR.

Partial Least-Squares Regression with Dummy Variables

In sensory evaluation, it is often found that some variables are highly correlated with each other. This correlation or dependence between variables violates the independence assumption for covariance analysis models, so analysis of covariance is no longer valid. Partial least-squares (PLS) regression or principal component regression (PCR) are often used to handle this kind of collinearity problems. For a single response, PLS regression models for each individual product can be expressed as

$$Y_j = \beta_0 + \beta_1 PC_{j1} + \beta_2 PC_{j2} + \ldots + \beta_s PC_{js} + \varepsilon_j \quad (j = 1,2, \ldots, m; s \leq p) \tag{7}$$

where the response variable Y_j is the OL score given by the jth consumer for this product; PC_{jk} is the score of the kth principal component for the jth consumer; s is the number of principal components and is less than or equal to p number of the original variables; β_0 is the regression intercept; β_k is the regression coefficient for the kth PC; ε_j is the residual for the jth consumer. If the original JAR variables are used for preference mapping of Overall Liking for each product (not all products), principal components (PCs) are calculated as follows

$$PC_{jk} = a_{k1}X_{j1} + a_{k2}X_{j2} + \ldots + a_{kp}X_{jp} \quad (j = 1,2, \ldots, m; k = 1,2, \ldots, s) \tag{8}$$

where a_{kl} ($k = 1,2,\ldots,s$; $l = 1,2,\ldots,p$) is the loading for the kth PC and the lth original JAR variable (X_l); X_{jl} is the score given by the jth consumer for the lth JAR variable. As discussed previously, non-linear relationships between the original JAR variables and the response (OL) cannot be appropriately described by a linear regression model (such as PLS or PCR) using the original variables. Dummy variables can be used in PLS or PCR models to estimate these non-linear relationships, and principal components (PCs) using dummy variables are calculated as

$$PC_{jk} = (b_{k11}Z_{j11} + b_{k12}Z_{j12}) + (b_{k21}Z_{j21} + b_{k22}Z_{j22}) + \ldots + (b_{kp1}Z_{jp1} + b_{kp2}Z_{jp2}) \quad (j = 1,2, \ldots, m; k = 1,2, \ldots, s) \tag{9}$$

where the pair (b_{kl1}, b_{kl2}) ($k = 1,2,\ldots,s$; $l = 1,2,\ldots,p$) represents the loadings for the kth PC and the lth pair of dummy variables (Z_{l1}, Z_{l2}) which represents the lth JAR variable (X_l). By substituting PC_{jk} in Eq (9) into the PLS model (Eq (7)), we obtain the following PLS model using p pairs of dummy variables:

$$Y_j = \gamma_0 + (\gamma_{11}Z_{j11} + \gamma_{12}Z_{j12}) + (\gamma_{21}Z_{j21} + \gamma_{22}Z_{j22}) + \ldots + (\gamma_{p1}Z_{jp1} + \gamma_{p2}Z_{jp2}) + \varepsilon_j \quad (j = 1,2, \ldots, m) \tag{10}$$

where γ_0 is the regression intercept which can be interpreted as the estimated mean of Overall Liking for the product if all JAR attributes are JAR; (γ_{l1}, γ_{l2}) are regression coefficients for the lth pair of dummy variables (Z_{l1}, Z_{l2}), respectively. If the scheme in Table 1 is used for each pair of dummy variables, regression coefficient γ_{l1} should be positive, while γ_{l2} should be negative, which means that the response (OL) has its maximum at the JAR score. Since not all variables in Eq (9) have equal influences on the response (OL), those unimportant or insignificant variables to the response need to be removed from the PLS model. The jackknife optimization method is one of popular methods to remove unimportant variables from PLS models or select important variables for PLS models. The PLS models using q ($q \leq p$) important pairs of dummy variables are given by

$$Y_j = \gamma'_0 + (\gamma'_{11}Z'_{j11} + \gamma'_{12}Z'_{j12}) + (\gamma'_{21}Z'_{j21} + \gamma'_{22}Z'_{j22}) + \ldots + (\gamma'_{q1}Z'_{jq1} + \gamma'_{q2}Z'_{jq2}) + \varepsilon_j \quad (j = 1,2, \ldots, m) \tag{11}$$

where γ'_0 is the regression intercept, the pair (γ'_{l1}, γ'_{l2}) are regression coefficients for the lth important pair of dummy variables (Z'_{l1}, Z'_{l1}). A pair of dummy variables is defined to be important or significant if at least one of a pair of dummy variables is selected into the PLS model by the jackknife or other optimization method. Important pairs of dummy variables imply that the intensities of the corresponding JAR attributes are not "Just About Right," while unimportant pairs

TABLE 2—Example dummy variables for a 5-point JAR scale.

JAR Rating	Z_1	Z_2	Z_3	Z_4	Z_5
1 ("Too Low")	1/5	−1/4	−1/2	1	0
2 ("Slightly Low")	1/5	−1/4	−1/2	−1	0
3 ("Just About Right")	1/5	1	0	0	0
4 ("Slightly High")	1/5	−1/4	1/2	0	−1
5 ("Too High")	1/5	−1/4	1/2	0	1

Dummy Variable	Interpretation
Z_1	Average rating
Z_2	Average difference between JAR and non-JAR values
Z_3	Average difference between "High" and "Low" values
Z_4	Difference between "Too Low" and "Slightly Low" values
Z_5	Difference between "Too High" and "Slightly High" values

of dummy variables can be interpreted as having no significant effect on Overall Liking or as the attribute to be at a JAR level. When the paired regression coefficients (γ'_{l1}, γ'_{l1}) are not equal for an important pair of dummy variables, the JAR attribute is more detrimental to OL over one JAR region than another. The PLS model (Eq (11)) is called the F-model (or pseudo full model) because it includes all important pairs of dummy variables. Based on this F-model, various null hypotheses can be formed and tested just as for the covariance analysis models. For example, if the null hypothesis is that the two drop rates (in absolute values) over the two JAR regions are the same for all important pairs of dummy variables in Eq (11) simultaneously (i.e., H_0: $\gamma'_{l1} = -\gamma'_{l1} = \varphi'_{l1}$ for all important pairs, $l = 1, 2, \ldots, q$; see Eq (5)), then a pair of two dummy variables can be combined into one single dummy variable ($Z'_{l1} = Z'_{l1} - Z'_{l1}$) and Eq (11) becomes

$$Y_j = \varphi'_0 + \varphi'_1 Z'_{j1} + \varphi'_2 Z'_{j2} + \ \ldots\ + \varphi'_q Z'_{jq} + \varepsilon_j \quad (j = 1, 2, \ldots, m) \tag{12}$$

where $\varphi'_0, \varphi'_1, \varphi'_2, \ldots, \varphi'_q$ are regression coefficients for Eq (12). This PLS model (Eq (12)) is called the R-model (pseudo reduced model) because it contains only a subset of the all dummy variables used in the above F-model (Eq (11)). Similarly, if the null hypothesis of interest is that the two drop rates are the same only for the first important pair of dummy variables (Z'_{j1}, Z'_{j2}), the PLS model is given by

$$Y_j = \phi'_0 + \phi'_1 Z'_{j1} + (\phi'_{21} Z'_{j21} + \phi'_{22} Z'_{j22}) + \ \ldots\ + (\phi'_{q1} Z'_{jq1} + \phi'_{q2} Z'_{jq2}) + \varepsilon_j \quad (j = 1, 2, \ldots, m) \tag{13}$$

where $\phi'_0, \phi'_1, \phi'_2, \ldots, \phi'_q$ are regression coefficients for Eq (13). The PLS model ((13)) is another R-model. Unlike the above analysis of covariance, however, there is no *F*-test available for testing the F-model and R-model. The root mean square error (RMSE) statistic can be used to assess which model is more appropriate. If RMSE values are "substantially" different between F- and R-models, the F-model is more appropriate. Otherwise, the R-model is more appropriate. Because F- and R-models fit the same data, the residuals from the two models are correlated with each other. In addition, RMSE_R (RMSE for R-model) is at least equal to or greater than RMSE_F (RMSE for F-model) because the R-model uses fewer dummy variables than the F-model. A paired *t*-test described by Snedecor and Cochran [1] for comparing two correlated variances can be used to test if RMSE_R and RMSE_F are significantly different at a significance level α. The *t* value for the paired *t*-test is computed as

$$t = r_{FR}\sqrt{\frac{N-2}{1-r_{FR}^2}}$$

$$\text{where } r_{FR} = \frac{F-1}{\sqrt{(F+1)^2 - 4r^2F}}$$

$$F = \frac{s_R^2}{s_F^2} = \frac{\text{DF}_R}{\text{DF}_F}\left(\frac{\text{RMSE}_R}{\text{RMSE}_F}\right)^2 \tag{14}$$

N is the number of observations; r is the correlation coefficient between the residuals for F- and R-models; S_R^2 and S_F^2 are standard deviations of the residuals for R- and F-models, respectively; DF_R and DF_F are the degrees of freedom for the R- and F-models, respectively, and they are calculated as the difference between the number of observations and the number of principal components in the PLS model. If computed *t*-value is equal to or greater than the table *t*-value (one-tailed test, degree of freedom $\text{DF} = N - 2$) at α, RMSE_R is significantly larger than RMSE_F, suggesting that F-model fits the data best. If the computed *t*-value is less than the table *t*-value at α, RMSE_R is not significantly larger than RMSE_F, suggesting that the R-model fits the data equally as well as the F-model. Once the final PLS model is determined, the mean score of overall liking for the product can be predicted by

$$\bar{Y} = \gamma'_0 + (\gamma'_{11}\bar{Z}'_{11} + \gamma'_{12}\bar{Z}'_{12}) + (\gamma'_{21}\bar{Z}'_{21} + \gamma'_{22}\bar{Z}'_{22}) + \ \ldots\ + (\gamma'_{q1}\bar{Z}'_{q1} + \gamma'_{q2}\bar{Z}'_{q2}) \tag{15}$$

where $\bar{Y}'$ is the predicted mean score of overall liking; $\bar{Z}'_{ik1}$ and $\bar{Z}'_{ik2}$ are the means of dummy variables Z_{jk1} and Z_{jk2} for the kth JAR variable, respectively; $\alpha_{ik1}\bar{Z}_{ik1}$ and $\alpha_{ik2}\bar{Z}_{ik2}$ are defined as the estimates of the mean drop on Overall Liking due to the kth JAR variable being "Too Little" and being "Too Much," respectively. When all attributes of the product are JAR, the intercept γ'_0 is equal to the predicted mean score of overall liking. When at least one or some attributes are not

TABLE 3—Example dummy variables for a 7-point JAR scale.

Jar Rating	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7
1 ("Too Low")	1/7	−1/6	−1/3	1	0	−1/2	0
2 ("Low")	1/7	−1/6	−1/3	0	0	1	0
3 ("Slightly Low")	1/7	−1/6	−1/3	−1	0	−1/2	0
4 ("Just About Right")	1/7	1	0	0	0	0	0
5 ("Slightly High")	1/7	−1/6	1/3	0	−1	0	−1/2
6 ("Low")	1/7	−1/6	1/3	0	0	0	1
7 ("Too Low")	1/7	−1/6	1/3	0	1	0	−1/2

Dummy Variable	Interpretation
Z_1	Average rating
Z_2	Average difference between JAR and non-JAR values
Z_3	Average difference between "High" and "Low" values
Z_4	Linear Trend over "Too Low" and "Slightly Low" values
Z_5	Linear Trend over "Too High" and "Slightly High" values
Z_6	Curvature in the "Low" values (Deviation from a linear trend)
Z_7	Curvature in the "High" values (Deviation from a linear trend)

JAR, the intercept is larger than the predicted OL mean score. The difference between the intercept and the predicted OL mean score can be interpreted as the total/overall mean drop due to not being JAR or as the maximum potential improvement margin on OL if the attributes that are not JAR are modified to be JAR.

Similarly to penalty analysis, important dummy variables in a PLS model can initially be selected according to the pre-specified percent of consumers (e.g., 20 %) who rate the attribute not to be JAR and then the jackknife optimization method applied to select the final important dummy variables. To compare with penalty analysis, it is recommended that dummy variables with 20 % or more consumers scored be used in the PLS model.

The results of the Analysis

Correlation Analysis

Correlation analysis is important for determining which model (covariance analysis or PLS model) is appropriate for preference mapping between JAR and hedonic scores. This is because covariance analysis requires the independence assumption on JAR variables. Correlation coefficients among the five original JAR variables (Size, Color, Flavor, Thin/Thick, and Stickiness) are presented in Table 2. The maximum correlation coefficient was −0.4043 between Thin/Thick and Stickiness, which was highly significantly different from zero with $p < 0.0001$, and the remaining correlation coefficients were less than ±0.12. Since some of the five JAR variables were correlated with each other, the independence assumption for analysis of covariance was somewhat violated. In this case, it is more appropriate to use a PLS model than a covariance analysis model. The following section will focus only on use of PLS models with dummy variables to map relationships between JAR and liking variables for each individual sample.

Partial Least-Squares Regression With Dummy Variables

Five pairs of dummy variables for the five original JAR variables were created using the scheme for a 5-point JAR scale in Table 1. PLS regression models using the five pairs of dummy variables were separately fitted by the Unscrambler software program (Unscrambler, version 7.5, CAMO, Norway) to the data for each individual product. The jackknife optimization method was applied to the PLS models to identify important pairs of dummy variables to Overall Liking. The fitted PLS models with significant dummy variables are presented in Table 3. For Sample 170, the jackknife method found that only one important pair of dummy variables (Z_{51}, Z_{52}) representing Stickiness significantly affected OL. For

TABLE 4—Scheme for using two dummy variables (Z_1 and Z_2) or one dummy variable (Z) to represent one JAR variable (X) on a 5-point or 7-point JAR scale.

5-Point JAR Scale				7-Point JAR Scale			
X	Z_1	Z_2	Z	X	Z_1	Z_2	Z
1	−2	0	−2	1	−3	0	−3
2	−1	0	−1	2	−2	0	−2
3	0	0	0	3	−1	0	−1
4	0	1	−1	4	0	0	0
5	0	2	−2	5	0	1	−1
				6	0	2	−2
				7	0	3	−3

TABLE 5—Correlation coefficients between JAR variables.

	Size	Color	Flavor	Thin/Thick	Stickiness
Size	1	0.0004	−0.0040	−0.0039	0.0947[a]
Color		1	0.0255	−0.1115[a]	0.0163
Flavor			1	−0.0280	0.0648
Thin/Thick				1	−0.4043[b]
Stickiness					1

[a]Significant at $p < 0.05$;
[b]Significant at $p < 0.0001$ ($n = 102$).

the pair (Z_{51}, Z_{52}), regression coefficient of 1.2692 (it stands for the absolute value hereafter) for Z_{52} means that the OL mean score decreased in the rate of 1.2692 per unit increase in stickiness over the region of 3 to 5, whereas regression coefficient for Z_{51} was zero, indicating that the OL mean score was not affected by Z_{51}. The combined effects of the pair of dummy variables (Z_{51}, Z_{52}) for stickiness on OL are graphically presented in Fig. 2. The figure shows that the OL mean score was constant at 5.7436 (regression intercept) over the JAR region of 1 to 3 for stickiness, and dropped linearly over the JAR region of 3 to 5, implying that "Too Sticky" was more detrimental to OL than "Not Sticky Enough." The estimated mean drop on OL for this sample was 0.224 (=1.2692*0.1765, where 0.1765 was the mean of Z_{52}; see Eq (15)) due to the product being "Too Sticky." The regression intercept was 5.7436, indicating that the potential maximum mean score was 5.7436 if all JAR attributes were overall "Just About Right." Although the PLS model did not give very good prediction to the OL scores for individual consumers (low R value in Table 3), it predicted well the observed OL mean score (Table 3) for sample 170. As far as consumers were concerned, 17 % of consumers rated this product "Too Sticky" and 1 % of consumers "Not Sticky Enough." Like penalty analysis, if only attributes for which 20 % or more consumers found the attribute intensity not to be JAR were initially selected into the PLS model, no attributes would be found by the jackknife method to have significant effects on OL.

For Sample 458, three important pairs of dummy variables (Z_{31}, Z_{32}) for Flavor, (Z_{41}, Z_{42}) for Thin/Thick, and (Z_{51}, Z_{52}) for Stickiness significantly affected OL (Table 3) because at least one of two paired regression coefficients was not zero. For the Flavor attribute, regression coefficients for the pair (Z_{31}, Z_{32}) were 0 and 0.6716, respectively, suggesting that "Too Strong" flavor had a negative impact on OL, while "Too Weak" had no impact. "Too Strong" flavor was rated by 32 % of consumers, whereas "Too Weak" flavor by only 6 %. The estimated mean drop on OL was 0.283 (=0.6716*0.4216, where 0.4216 was the mean of Z_{32}) due to the "Too Strong" flavor. For the Stickiness attribute, 4 % and 14 % of consumers rated the sample "Not Sticky Enough" and "Too Sticky," respectively. The PLS model shows that only dummy variable Z_{52} for Stickiness significantly decreased the OL mean score, suggesting that "Too Sticky" texture was detrimental to OL. The estimated mean drop on OL was 0.161 (=0.7837*0.2059, where 0.2059 was the mean of Z_{52}). For the Thin/Thick attribute, the two regression coefficients (0.8586 and 0.8737) for Z_{41} and Z_{42} were significantly different from zero. Since the regression coefficient of 0.8737 for Z_{42} was slightly larger than the regression coefficient of 0.8586 for Z_{41}, the paired t-test (Eq (14)) was conducted to test whether the two regression coefficients were the same or not. It was found that there was no significant difference in RMSE value between F-model (containing Z_{32}, Z_{41}, Z_{42}, Z_{52}) and R-model

TABLE 6—Results from PLS models using dummy variables for the five samples.

		170		458		523		896		914	
JAR Attribute		% of Panelists	Estimate	% of Panelists	Estimate	% of Panelists	Estimate	% of Panelists	Estimate	% of Panelists	Estimate
Intercept			5.7436		6.1246		6.4828		6.5282		7.4183
Size	Z_{11}	24	0	32	0	25	0	22	0	27	0
	Z_{12}	40	0	32	0	39	0	39	0	41	0
Color	Z_{21}	11	0	13	0	11	0	1	0	21	0
	Z_{22}	6	0	1	0	5	0	3	0	0	0
Flavor	Z_{31}	16	0	6	0	5	0	5	0	53	0.8151
	Z_{32}	23	0	32	−0.6716	22	−0.8067	13	0	3	0
Thin/Thick	Z_{41}	13	0	16	0.8586	15	0.7670	9	0	4	0
	Z_{42}	5	0	9	−0.8737	3	0	3	0	21	−1.2391
Stickiness	Z_{51}	4	0	7	0	2	0	1	0	10	1.1624
	Z_{52}	17	−1.2692	20	−0.7837	16	−0.9864	14	−1.5249	9	0
R			0.24		0.47		0.46		0.30		0.53
RMSE			2.13		1.52		1.54		1.82		1.77
N			102		101		102		102		101
Observed OL mean			5.52		5.48		6.00		6.30		6.51
Predicted OL mean			5.52		5.47		6.00		6.30		6.55
Rank of preference			4		5		3		2		1
Potential Improvement			0.224		0.655		0.483		0.228		0.868

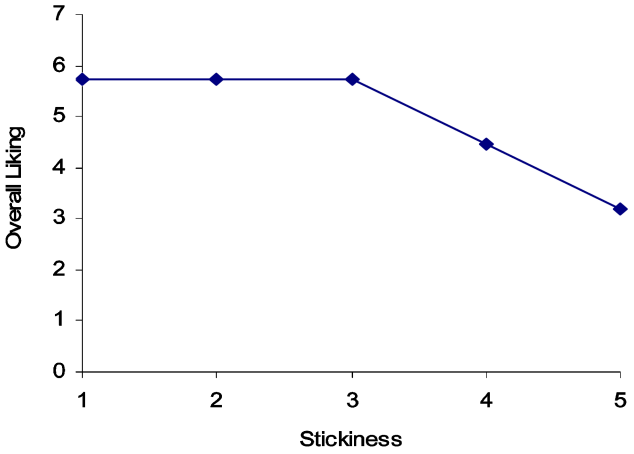

Fig. 2—Effect of stickiness on overall liking for sample 170.

(containing Z_{32}, $Z_4 = Z_{41} - Z_{42}$, and Z_{52}) at $\alpha = 0.05$, suggesting no significant difference between the two regression coefficients. It was concluded that "Too Thin" and "Too Thick" textures statistically decreased OL at the same rate and that the R-model was more appropriate for sample 458 than the F-model. The results from the R-model are presented in Fig. 3(*a*), which was made to look like the graphical presentation of the results from penalty analysis. The figure is the plot of the mean drop on OL versus percent of consumers who rated the product not to be JAR. For a dummy variable, overall mean drop on Overall Liking is the product of the regression coefficient by the mean of the dummy variable across consumers. Figure 3(*a*) shows that "Too Strong" flavor caused the most mean drop on OL, while "Too Thick" texture the least mean drop for this sample. Although the drop rates for "Too Thin" and "Too Thick" were the same, "Too Thick" texture dropped more OL mean score than "Too Thin" texture because more consumers scored the product "Too Thick"

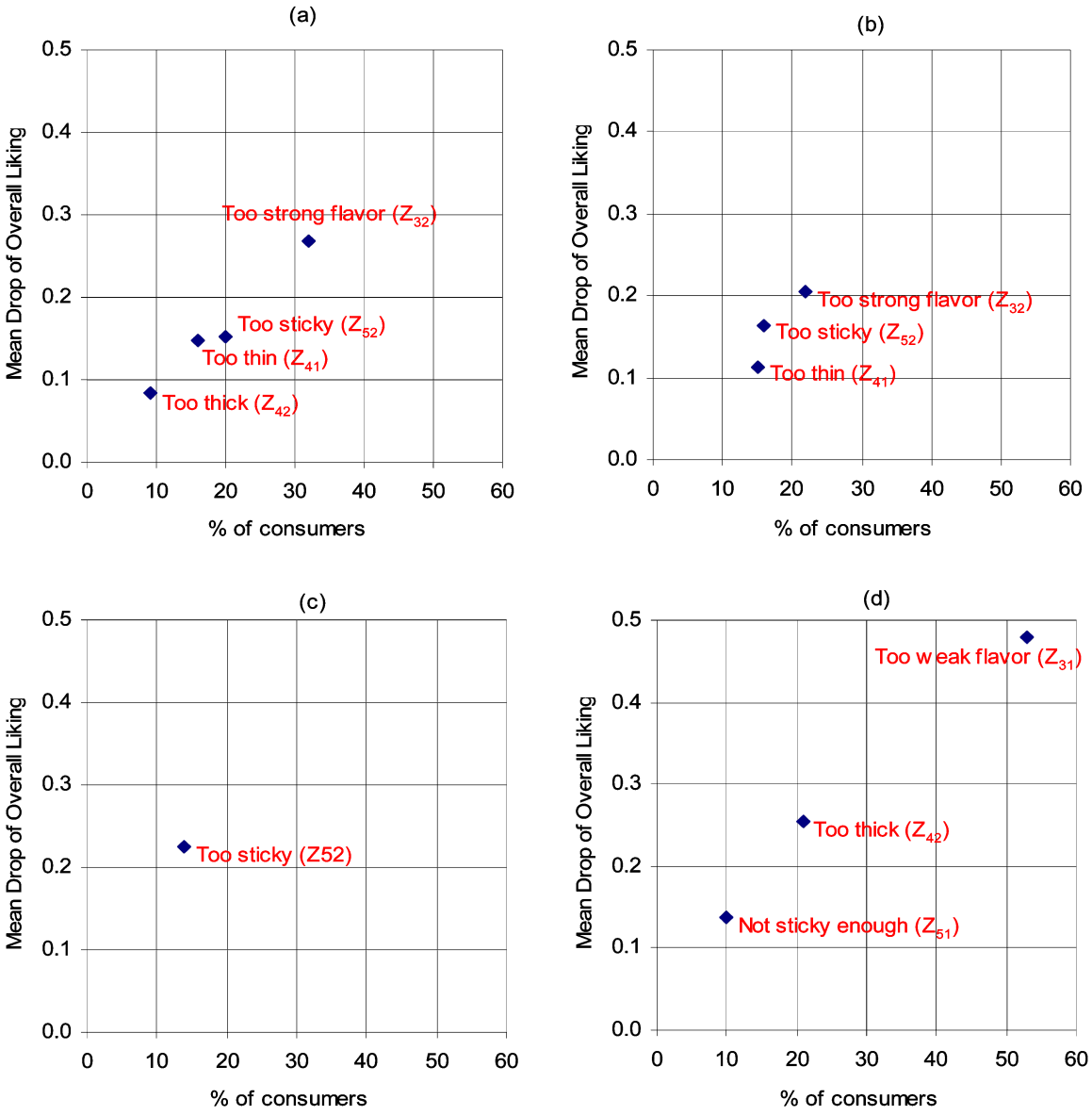

Fig. 3—The results from the final PLS models: (a) for sample 458, (b) for sample 523, (c) for sample 896 and (d) for sample 914.

than "Too Thin." This sample was the least liked product in terms of the predicted OL mean scores.

For sample 523, the jackknife method selected three important pairs of dummy variables (Z_{31}, Z_{32}) for Flavor, (Z_{41}, Z_{42}) for Thin/Thick, and (Z_{51}, Z_{52}) for Stickiness (Table 3). The regression coefficients suggested that "Too Strong" flavor had more negative influence on OL than "Too Weak" flavor, "Too Thick" texture was more detrimental to OL than "Too Thin" texture, and "Too Sticky" texture was also more detrimental to OL than "Not Sticky Enough" texture. The mean drop plot (Fig. 3(*b*)) shows that more consumers rated "Too Strong" flavor for this sample and "Too Strong" flavor resulted in a greater drop of OL than "Too Thick" or "Too Sticky" texture.

For Sample 896, only one important pairs of dummy variables (Z_{51}, Z_{52}) was found by the jackknife method for Stickiness (Table 3). Dummy variable Z_{51} had no significant effect on OL, but dummy variable Z_{52} significantly decreased OL (Table 3). This implies that "Too Thick" texture was detrimental to OL. The mean drop plot (Fig. 3(*c*)) shows that 14 % of consumers rated the product "Too Sticky," which dropped the OL mean score by 0.224.

For Sample 914, the jackknife method determined three important pairs of dummy variables (Z_{31}, Z_{32}) for Flavor, (Z_{41}, Z_{42}) for Thin/Thick, and (Z_{51}, Z_{52}) for Stickiness (Table 3). By comparing the regression coefficients for each pair of dummy variables, it was found that "Too Weak" flavor had more negative influence on OL than "Too Strong" flavor, "Too Thick" texture was more detrimental to OL than "Too Thin" texture, and "Not Sticky Enough" texture was also more detrimental to OL than "Too Sticky" texture. The mean drop plot (Fig. 3(*d*)) shows that 53 % of consumers rated the flavor of this sample "Too Strong" and "Too Strong" flavor dropped the OL mean score by 0.479, 21 % of consumers scored the texture of the product "Too Thick" with an estimated mean drop of 0.255, and 10 % of consumers scored the texture to "Not Sticky Enough" with an estimated mean drop of 0.138. Figure 3 also shows the trends that the mean drops on OL increased as the percent of consumers increased because the means of dummy variables are dependent on the number of consumers who rated the variables not to be JAR. It is evident from Table 3 that sample 914 had the highest regression intercept (7.4183) and predicted OL mean score (6.55). This sample was the most liked product, with the actual OL mean score of 6.51.

Conclusions from the Analysis

- Sample 170 was found to have only one attribute (Stickiness) not being JAR. Being "Too Sticky" significantly dropped the OL mean score, while being "Not Sticky Enough" had no significant effect on OL. This sample was the least liked product of the five samples tested.
- Sample 458 had three attributes (Flavor, Thin/Thick and Stickiness) significantly affecting OL. "Too Strong" flavor and "Too Sticky" texture were more detrimental to OL than "Too Weak" flavor and "Not Sticky Enough" texture, respectively. Although the drop rates for "Too Thin" and "Too Thick" were the same, "Too Thick" texture resulted in a greater drop in OL than "Too Thin" texture.
- For Sample 523, Flavor, Thin/Thick, and Stickiness significantly decreased the OL mean score. "Too Strong" flavor, "Too Thin," and "Sticky" textures had more negative impact on OL.
- "Too Sticky" texture had significantly negative contribution to overall liking of Sample 896.
- "Too Strong" flavor of Sample 914 was more detrimental to OL than "Too Weak" flavor. Being "Too Thin" and "Not Sticky Enough" significantly decreased the OL mean score. This sample was the most liked product.

Benefits from the Analysis

- The dummy variable method is flexible in that it can be used with many regression models. If the JAR variables are independent of each other, dummy variables are used with covariance analysis models. If JAR variables are correlated with each other, dummy variables are used with PCR or PLS regression models for mapping relationships between JAR and liking variables.
- The method also provides a tool to perform various hypothesis tests.
- Like penalty analysis, it uses a similar graphical presentation of relationships between JAR and liking variables, but unlike penalty analysis it is a regression method, which can estimate the "true" mean drop of OL.
- The difference between the regression intercept and actual OL mean score indicates the average potential improvement margin if the product is modified to be JAR.

Caveats from the Analysis

The drawback of this method is that there is no single software program to implement it.

Recommendations

Correlation analysis of JAR variables is recommended before selecting a model to determine whether the independency assumption is met. If the assumption is met, it is recommended to use covariance analysis model with dummy variables because the effects of treatments/products, consumer panelists and other factors on OL can be tested simultaneously. Otherwise, principal components-based regression models (such as PLS or PCR models) with dummy variables are recommended.

References

[1] Snedecor, G. W., and Cochran, W. G., *Statistical methods*, The Iowa State University Press, Ames, Iowa, 1976.

Appendix X: Collecting Intensity and Hedonic Information Separately

Gloria A. Gaskin[1] and Joni L. Keith[1]

This example utilizes attribute intensity data rather than "Just About Right" data in conjunction with Overall Liking in order to determine attribute intensities that maximize Overall Liking. For this example, previously published data were utilized [1].

By relating Overall Liking to attribute intensities on an attribute-by-attribute basis and on a product-by-product basis, information is gained regarding the drivers of Overall Liking. Simple linear regression can be used to determine an equation relating Overall Liking to attribute intensity (Overall Liking $= b_0 + b_1$Attribute Intensity).

The slope b_1 measures importance. High values of b_1 indicate an important attribute, while low values of b_1 indicate an attribute that is less important in predicting Overall Liking [2].

From Table 1 below for Product C the attribute Roasted Garlic Intensity (Int) significantly affects Overall Liking ($p \leq 0.05$). The regression model shows that a one-point increase in Roasted Garlic Intensity corresponds to an increase of 0.293 points in Overall Liking. In contrast, Flavor Strength Intensity (Int) for Product C is insignificant ($p > 0.05$). A one-point increase in Flavor Strength Intensity corresponds to an increase of only 0.048 points in Overall Liking.

A visual representation of the relationship between Overall Liking (1 = "Dislike Extremely", 9 = "Like Extremely") and Roasted Garlic Intensity (Int) (1 = "Not at All", 9 = "Extremely") for Product C (Code 998) is shown in the frequency scatterplot in Figure 1. As Roasted Garlic Intensity (Int) increases, Overall Liking increases.

To further utilize this method, a multiple regression model may be built using the significant intensity attributes generated by the simple regression analysis (Table 1). Significant parameters from multiple regression may be plotted on a 3D contour plot to determine the locations where Overall Liking is maximized.

Both Roasted Garlic Intensity (Int) (1 = "Not at All", 9 = "Extremely") and Chunkiness Intensity (Int) (1 = not at all, 9 = extremely) significantly impacted Overall Liking as shown in Table 1 and were selected for inclusion in a multiple regression model. A plot of the model is shown in Figure 2. Overall Liking for Product C is maximized by keeping Chunkiness Intensity (Int) low while increasing Roasted Garlic Intensity (Int).

TABLE 1—Product Example (Product C, Code 998).

Attribute	Slope: b_1	Significant at $p \leq 0.05$
Q#21. ROASTED GARLIC INT	0.29265	Yes
Q#19. TOMATO FLAVOR INT	0.28156	Yes
Q#9. COLOR INT	0.24536	Yes
Q#34. THICKNESS (IN THE MOUTH) INT	0.21578	Yes
Q#6. THICK APPEARANCE INT	0.12699	No
Q#23. HERB FLAVOR INT	0.11060	No
Q#25. SWEETNESS INT	0.10147	No
Q#38. FIRMNESS OF VEGETABLE PIECES INT	0.06854	No
Q#31. HEAT (SPICE) INT	0.06557	No
Q#14. AROMA STRENGTH INT	0.05260	No
Q#17. FLAVOR STRENGTH INT	0.04784	No
Q#27. SALTINESS INT	–0.00365	No
Q#29. SOURNESS/TANGINESS INT	–0.01357	No
Q#36. CHUNKINESS INT	–0.43286	Yes

[1] *Bush Brothers and Company, 1016 E. Weisgarber Road, Knoxville, TN 37909.*

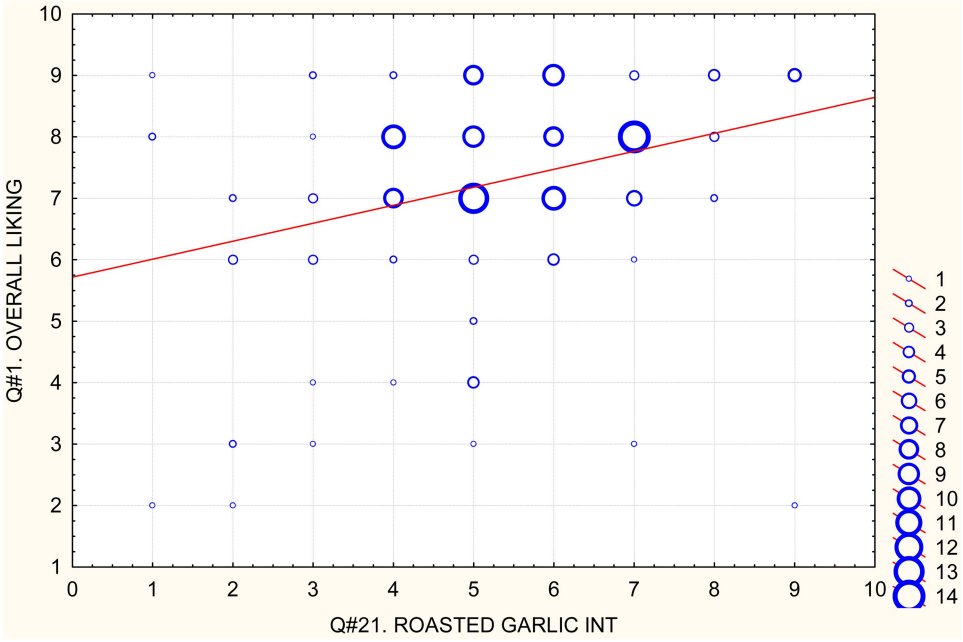

Fig. 1—Frequency scatterplot (Pangborn JAR workshop data 46v*1727c). Include condition: v5=998. Q#1. Overall liking=5.7167+0.2927*x.

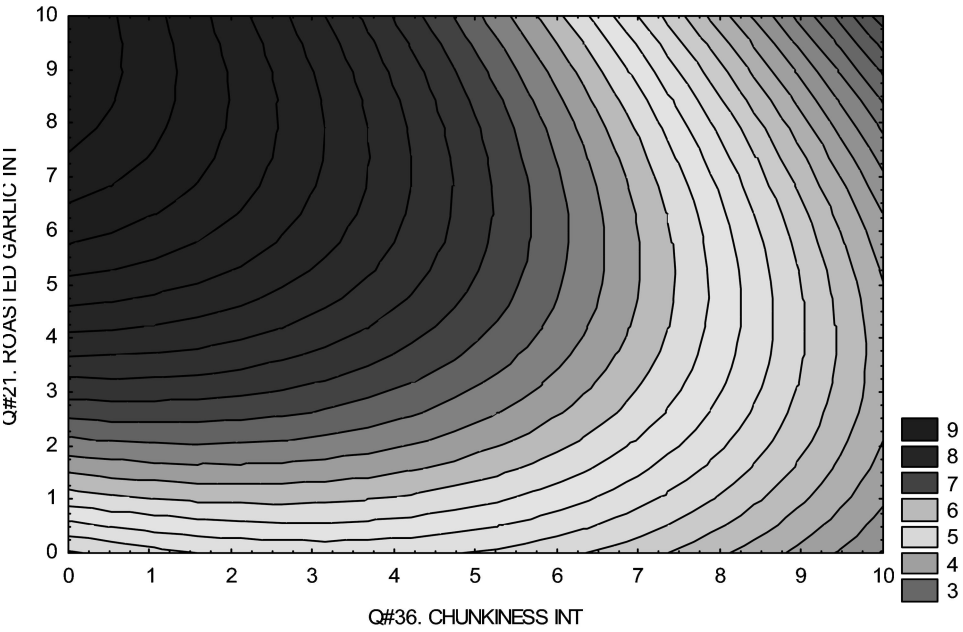

Fig. 2—3D contour plot (Pangborn JAR Workshop Data 46v*1727c). Include condition: v5=998. Q#1. Overall liking=4.7118+0.2311*x +0.9367*y-0.0354*x*x-0.0589*x*y-0.0485*y*y.

Recommendation

This technique is appropriate when multiple products are being evaluated outside of a design of experiments; i.e., where variables have not been systematically varied.

References

[1] Popper, R., "Workshop Summary: Data Analysis Workshop: Getting the Most Out of Just-About-Right Data," *Food Quality Preference*, Vol. 15, 2004, pp. 891–899.

[2] Moskowitz, H. R., "Learning from the Competition through Category Appraisal: One Practitioner's Keys to Faster and More Efficient Product Development," *Food Service Technol.*, Vol. 1, No. 2, 2001, pp. 103–118.

Appendix Y: Designed Experiments

Merry Jo Parker[1] and B. Thomas Carr[2]

Objective

The objective of a designed experiment is to determine the optimal level of a variable or combination of variables within the experimental range tested. Designed experiments obviate the need for JAR scales because there is a direct link between the experimental variables and consumer response. Experimental variables are generally product ingredients or components.

Requirements

One or more experimental variables are chosen and products are produced in accordance with the appropriate statistical design. For analysis, only the mean liking scores for each product are required along with the associated variable levels.

"How to"

- Test products are defined by systematic variations in ingredients and/or process settings.
- The results are valid within the experimental ranges chosen in advance of a study.
- Experimental design is typically used when the number of potential variables is small or when one knows which factors affect product characteristics that are important to consumers.
- This method can assist R&D in optimizing levels of ingredients and process settings.

Example

A researcher wants to determine the optimum levels of sugar and citric acid in a product being developed. An experimental design is being utilized to determine the optimum levels of each attribute. High, medium, and low levels of each attribute have been provided.

Attribute Levels	Sugar	Acid
High	50 g	4.5 g
Medium	22 g	2 g
Low	6 g	0.5 g

The following is an example of a two-variable, three-level factorial design that could be used for the consumer research.

SUGAR LEVEL → ACID LEVEL ↓	50 g	28 g	6 g
4.5 g	Sample 1	Sample 2	Sample 3
2.5 g	Sample 4	Sample 5	Sample 6
0.5 g	Sample 7	Sample 8	Sample 9

All other product attributes levels are held constant. Only the above attributes are varied per the design. The analysis of the data from this design would identify the optimum levels of each attribute tested.

Results

Average Overall Liking ratings of the test products are presented in the table below.

SUGAR LEVEL → ACID LEVEL ↓	50 G	28 G	6 G
4.5 G	6.2	6.4	5.2
2.5 G	6.5	6.7	5.1
0.5 G	5.8	5.3	4.4

A second-order polynomial response-surface model was fit to the data. Overall Liking is the dependent, response variable. Sugar Level and Acid Level are the independent, predictor variables. The form of the regression equation is:

$$\text{Liking} = B_0 + B_1(\text{Sugar}) + B_2(\text{Acid}) + B_{11}(\text{Sugar})2 + B_{22}(\text{Acid})2 + B_{12}(\text{Sugar})(\text{Acid}),$$

where "Liking" = Overall Liking, "Sugar" = Sugar Level, "Acid" = Acid Level and the B_i's are the regression coefficients whose values are estimated using regression analysis.

The resulting regression model is:

$$\text{Liking} = 3.38 + 0.10(\text{Sugar}) + 0.88(\text{Acid}) - 0.00124(\text{Sugar})^2 - 0.138(\text{Acid})^2$$

(the (Sugar) (Acid) cross-product term was not statistically significant). The model explains 95 % of the variability in Overall Liking. A graphical representation of the results, called a contour plot, is presented below. The plot illustrates the location of the optimal levels of sugar (42 gm) and acid (3 gm). The plot also illustrates how sensitive consumers are to deviations from the optimal levels. Note that because the optimal variable levels are predicted based on the products' hedonic ratings, the use of JAR scales is obviated. (See Fig. 1.)

For more information on design and analysis of experimental design see Gacula et al. [1], Meilgaard et al. [2], and Myers and Montgomery [3].

[1] *Food Perspectives, 2880 Vicksburg lane, Plymouth, MN 55443.*
[2] *Carr Consulting, 1215 Washington Ave., Suite 203, Wilmette, IL 60091.*

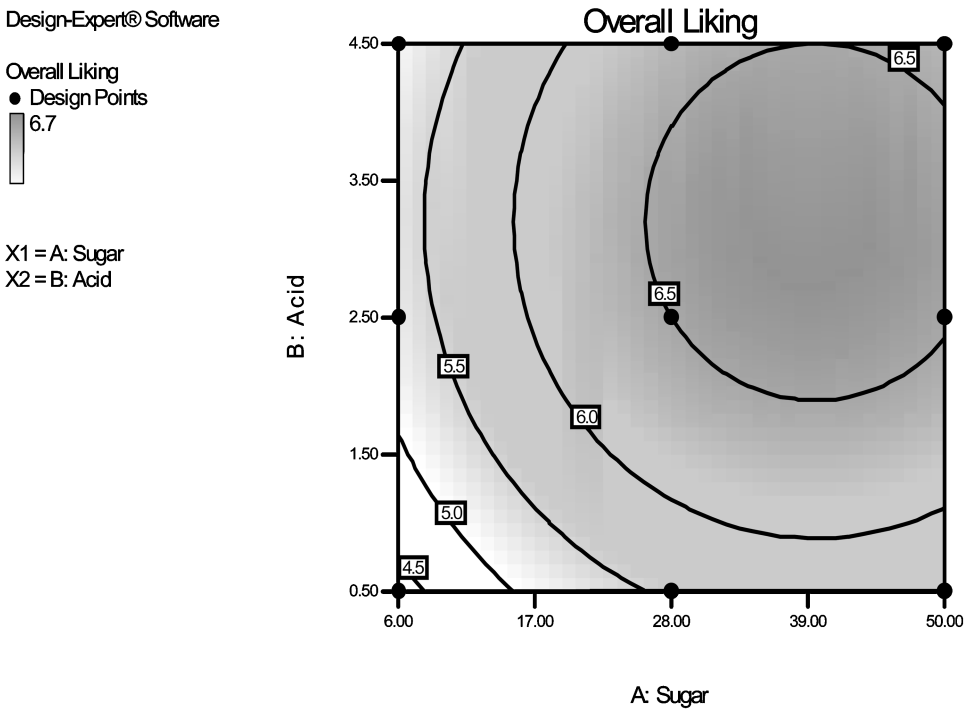

Fig. 1—Contour plot of overall liking by sugar and acid levels.

References

[1] Gacula, Jr., M. C. and Singh, J., Bi, J., Altan, A., *Statistical Methods in Food and Consumer Research*, 2nd edition, Academic, San Diego, CA, 2009.

[2] Meilgaard, M., Civille, G. V., and Carr, B. T., *Sensory Evaluation Techniques*, 4th edition, CRC Press, Boca Raton, FL, 2006.

[3] Myers, R. H. and Montgomery, D. C., *Response Surface Methodology: Process and Product Optimization using Designed Experiments*, John Wiley & Sons, New York, 1995.

Appendix Z: Ideal Scaling

Anne Goldman[1] and Jagoda Mazur[1]

Objective

In place of attempting to gauge attribute intensity and acceptability in one scale, ideal point modeling involves separating out the hedonic component of the response from the intensity evaluation. The "Ideal" product ratings are compared to the actual ratings. Attribute liking can be used to supplement findings from this technique.

"How to"

In using this method, mean attribute data and "Ideal" attribute data are required for each product/attribute combination of interest. Attribute liking data may also be collected. It is postulated that the greater the distance between the perceived and ideal intensities, the greater the change that must be made to "fix" the attribute. Response to the prior liking question, if asked, may suggest the significance of the discrepancy between the perceived and ideal intensities with respect to product acceptance.

Example

The following is an example of using the "Ideal" rating technique for evaluating the product diagnostics of four brands of chocolate brownies A, B, C, and D.

Background

The objective of this research was to measure the competitive performance of Product A against three other chocolate brownie brands (B, C, and D). The four products were rated on structural intensity scales for appearance, texture, and flavor attributes for the icing and the cake components of this product. In addition, they were rated for hedonic questions using a 9-point hedonic scale ("Dislike Extremely"—"Like Extremely"). At the very end of the testing session (this can also be done before product ratings), consumers completed an "Ideal" questionnaire for an "Ideal" brownie.

Below is an example of the "Ideal" question related to overall flavor:

1. The *overall flavor* of this chocolate brownie is…

1	2	3	4	5	6	7	8	9
Very weak								Very strong

TABLE 1—Diagnostic ratings for the four products.

	Product A	Product B	Product C	Product D	Ideal
		Appearance			
Size	6.0a	6.5a	5.4b	5.1b	6.0a
Amount of surface nuts	5.5a	4.7b	4.4b	4.3b	5.7a
Amount of chocolate icing	5.9a	5.7a	4.2c	5.0b	5.8a
Color of chocolate icing	6.8ab	7.1a	4.9c	6.5b	7.2a
Smoothness of chocolate icing	7.5a	7.0ab	4.4d	6.4c	7.4a
Shininess of chocolate icing	5.0b	5.0b	2.6d	4.0c	6.1a
Color of the cake	7.5a	6.6b	4.5c	4.8c	6.9b
		Texture of the Chocolate Icing			
Firmness	4.8c	5.8b	6.7a	5.6b	4.3c
Smoothness	6.9a	6.5ab	4.6c	6.2b	7.6a
Meltability	5.3a	4.7b	3.7c	4.5b	5.8a
		Texture of Cake			
Moistness	7.0b	6.0c	4.0d	6.1c	7.6a
Crumbliness	3.3a	2.8a	2.9a	3.0a	3.3a
Chewiness	5.5c	6.3b	6.9a	6.1b	5.3c
Stickiness while chewing	3.9c	5.0b	5.9a	5.3b	3.7c
		Flavor of Cake			
Overall sweetness	5.4b	5.4b	5.6b	6.1a	4.9c
Overall chocolate flavor	6.2a	5.8ab	4.5c	5.6b	7.2a
Nut flavor	4.7b	3.6c	3.2c	3.4c	5.3a
Overall flavor (naturalness)	5.8a	5.1b	3.3c	5.0b	8.3a
Overall flavor balance	6.2a	5.8b	4.1c	5.4b	7.9a
Freshness	7.4a	6.6b	4.5c	6.4b	8.7a

Note: Average scores in a row per flavor set followed by different letters are significantly different ($p < 0.05$). (Suggested color scheme: Green=parity to Ideal, within row, Blue=score significantly lower than Ideal, Red=score significantly higher than Ideal.)

[1] *Applied Consumer & Clinical Evaluations, 2575 B. Dunwin Dr., Mississauga, ON L5L3N9.*

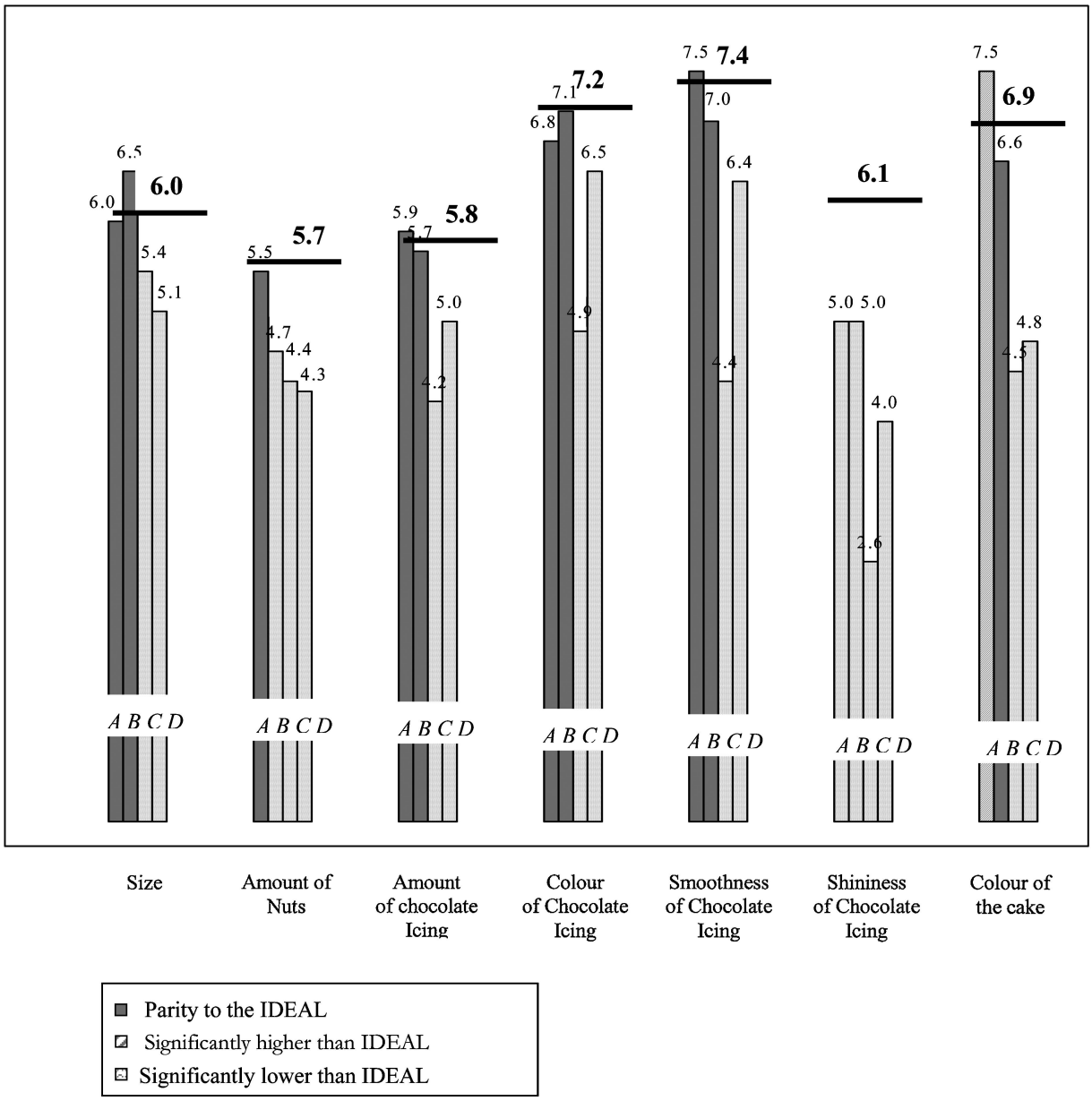

Fig. 1—Meeting the ideal for appearance dimensions.

2. The *overall flavor* of the "Ideal" chocolate brownie should be…

1	2	3	4	5	6	7	8	9
Very weak								Very strong

Results

Results are presented in Table 1 and in Figs. 1–3. They show that Product A performed better than the other three products based on hedonic scores and as shown by a number of product ratings that satisfied the consumer's "Ideal" ratings for this product. Product A met the "Ideal" for the appearance of size, amounts of surface nuts and chocolate icing, smoothness, firmness, and meltability of the chocolate icing, as well as crumbliness, chewiness, and stickiness of the cake. However, product A was rated too low for shininess of chocolate icing, moistness of the cake, and the majority of flavor dimensions. Product A was also perceived as being too sweet and too dark for color relative to the "Ideal." The superior performance of Product A for product diagnostics, relative to the other three brands, was also reflected by the highest scores for overall opinion and liking of appearance.

The only other brand which was rated above 6.0 for "Overall Opinion" was Product B. This product met the

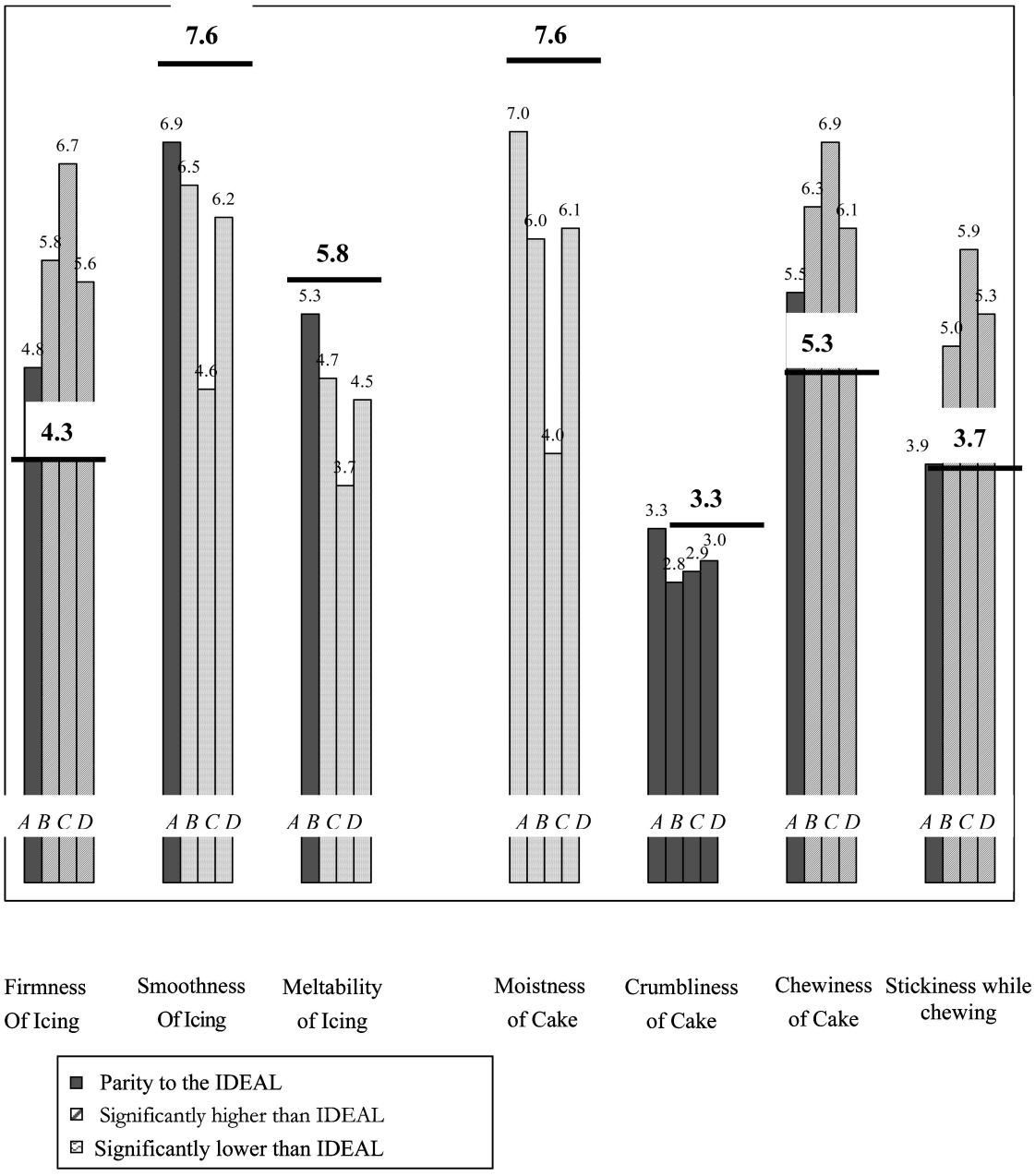

Fig. 2—Meeting the ideal for texture flavor.

"Ideal" for a series of appearance attributes including size, amount of chocolate icing, color of chocolate icing, smoothness of chocolate icing, color of the cake, and crumbliness. The other two products both performed very poorly based on hedonic ratings and failed to meet the "Ideal" for all attributes except for the crumbliness.

Benefits

This method allows for easy visual assessment of the "Ideal" and the product attribute ratings. It may require less mental processing for the respondent than JAR scales because the ideal intensity and the actual intensity are scaled separately. Obtaining attribute liking may assist in postulating the relationship between the difference from "Ideal" and overall liking.

Disadvantages

This technique assumes that consumers know their ideal level of an attribute. It further assumes that reformulation will improve the Overall Liking of the product. However, without a link between the attribute and Overall Liking, this

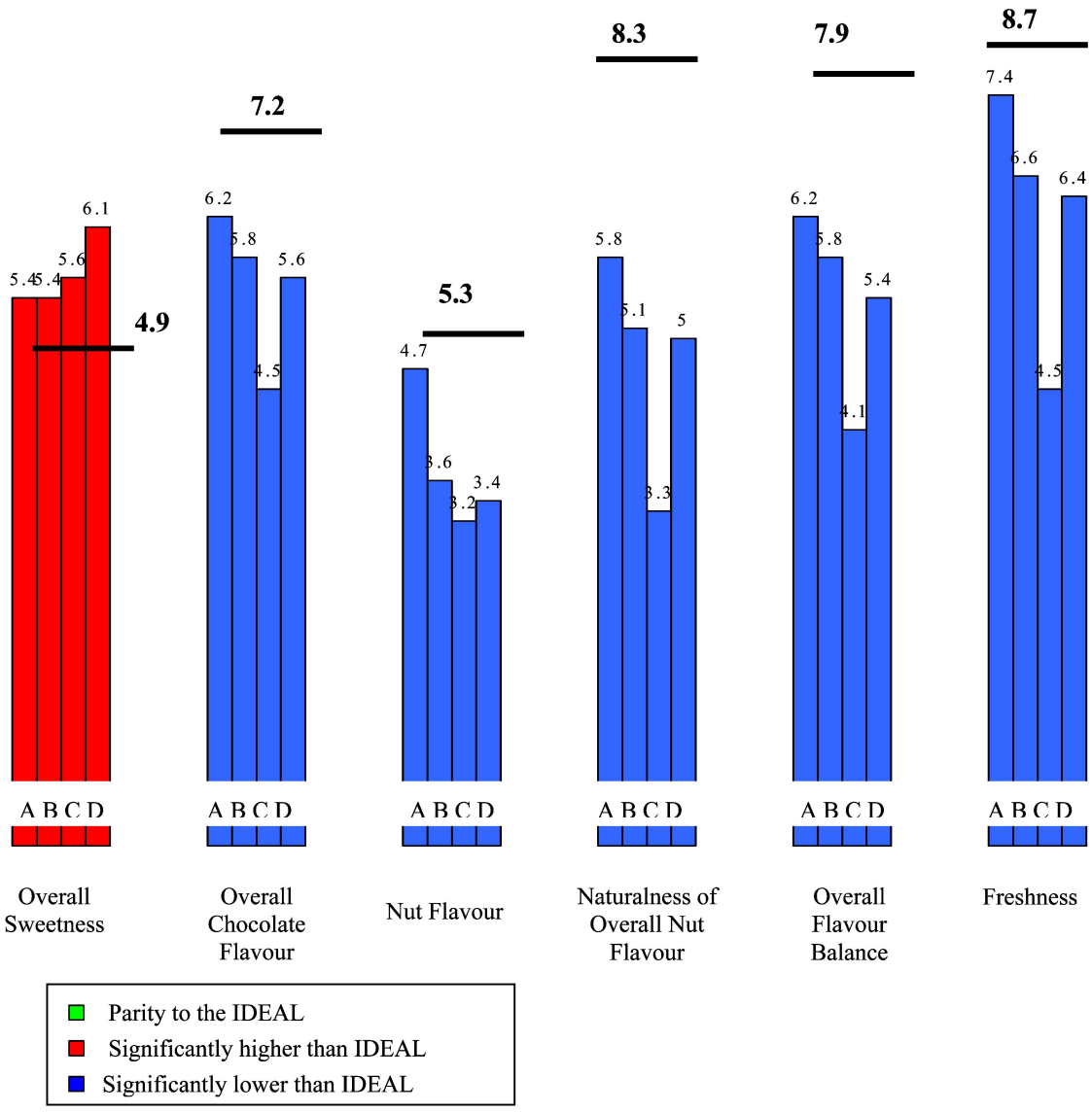

Fig. 3—Meeting the ideal for flavor dimensions.

may be a false assumption. The method does not provide the amount of the proposed attribute change, other than using the difference between ideal and actual intensity as a gauge. Similar to data obtained from JAR scales, the data may show evidence of bimodality as responses may suggest both lower and higher than "Ideal" direction.

Recommendation

This is a good alternative method to JAR scales in that the mental work of simultaneously judging the ideal and actual attribute intensities are separated. However, it suffers from many of the same pitfalls as other methods of JAR scale analysis,

Bibliography

ASTM E253 Terminology Relating to Sensory Evaluation of Materials and Products, ASTM International, W. Conshohocken, PA.

ASTM E456 Terminology Relating to Quality and Statistics, ASTM International, W. Conshohocken, PA.

Albaum, G., "The Likert Scale Revisited: An Alternate Version (Product Preference Testing)," *Journal of the Market Research Society*, Vol. 39, No. 2, 1997, p. 331.

Armstrong, R. L., "The Midpoint on a Five-Point Likert-type Scale," *Perceptual Motor Skills*, Vol. 64, 1987, pp. 359-362.

Baxter, I. A., Jack, Schroder, M. J. A., "The Use of Repertory Grid Method to Elicit Perceptual Data from Primary School Children," *Food Quality Preference*, Vol. 9, 1998, pp. 73-80.

Bendig, A. W. and Hughes, J. B., II, "Effect of Amount of Verbal Anchoring and Number of Rating-Scale Categories upon Transmitted Information," *J. Exp. Psychol.*, Vol. 40, No. 2, 1953, pp. 87-90.

Booth, D. A., Thompson, A., and Shahedian, B., "A Robust, Brief Measure of an Individual's Most Preferred Level of Salt in an Ordinary Foodstuff," *Appetite: J. Intake Res.*, Vol. 4, 1983, pp. 301-312.

Connor, M. T. and Booth, D. A., "Preferred Sweetness of a Lime Drink and Preference for Sweet Over Non-Sweet Foods," *Appetite*, Vol. 10, 1988, pp. 25-35.

Coombs, C. H., "Psychological Scaling Without a Unit of Measurement," *Psychol. Rev.*, Vol. 57, 1950, pp. 145-158.

Coombs, C. H., *A Theory of Data*, John Wiley & Sons, Inc., New York, 1964.

Cooper, L. G. and Nakanishi, M., "Two Logit Models for External Analysis and Preferences" *Psychometrika*, Vol. 48, 1983, pp. 607-620.

Devlin, S. J. and Dong, H. K., "Selecting a Scale for Measuring Quality," *Marketing Res.*, Vol. 5, No. 3, Summer 1993, p. 12.

Dixon, P. N., Bobo, M., and Stevick, R. A., "Response Differences and Preferences for All-Category-Defined and End-Defined Likert-Formats," *Educ. Psychol. Measurement*, Vol. 44, 1984, pp. 61-67.

Earthy, P. J., MacFie, J. H., and Duncan, H., "Effect of Question Order on Sensory Perception and Preference in Central Location Trials," *J. Sens. Stud.*, Vol. 12, 1997, pp. 215-237.

Ennis, Daniel M., *Analytic Approaches to Accounting for Individual Ideal Points*, IFPress, Vol. 8, No. 2, 2005, pp. 2-3.

Epler, S., Chambers, E., IV, and Kemp, K. E., "Hedonic Scales are a Better Predictor than Just-About-Right Scales of Optimal Sweetness in Lemonade," *J. Sens. Stud.*, Vol. 13, 1998, pp. 191-197.

Gacula, M. C., Jr., "Analysis of Incomplete Block Designs with Reference Sample in Every Block," *J. Food Sci.*, Vol. 43, 1978, pp. 1461-1466.

Gacula, M. C., Jr., Rutenbeck, S., Pollack, L., Resurreccion, A., and Moskowitz, H. R., "The just-about-right intensity scale: Functional analyses and relation to hedonics," *J. Sens. Stud.*, Vol. 22, 2007, pp. 194–211.

Gacula, M. C., Jr., Singh, J., Bi, J., and Atlan, A., *Statistical Methods in Food and Consumer Research*, 2nd edition, Academic, San Diego, CA, 2009.

Grapentine, T., "Problematic Scales: When Measuring Quality, Expectations Scales Exhibit Several Drawbacks," *Marketing Res.*, Fall 1994.

Gridgeman, N. T., "A Comparison of Some Taste-Test Methods," *J. Food Sci.*, Vol. 16, 1961, pp. 171-177.

Holbrook, M. B., "Situation-Specific Ideal Points and Usage of Multiple Dissimilar Brands," *Res. Marketing*, Vol. 7, 1984, pp. 93-131.

Jones, L.V., Peryam, D. R., and Thurstone, L. L., "Development of a Scale for Measuring Soldiers' Food Preferences," *Food Res.*, Vol. 20, 1955, pp. 512-520.

Johnson, J. and Vickers, Z., "Avoiding the Centering Bias or Range Effect When Determining an Optimum Level of Sweetness in Lemonade," *J. Sens. Stud.*, Vol. 2, 1987, pp. 283-292.

Kamakura, W. K., "Estimating Flexible Distributions of Ideal-Points with External Analysis of Preferences," *Psychometrika*, Vol. 56, pp. 419-431.

Kim, W., Ennis, D., and O'Mahony, M., "A New Approach to Category Scales of Intensity II: Use of *d* Values," *J. Sens. Stud.*, Vol. 13, 1998, pp. 251-267.

Lawless, H. T. and Heyman, H., *Sensory Evaluation of Food Principles and Practices*, Chapman and Hall, New York, 1998.

Lawless, H. T. and Malone, G. J., "The Discriminative Efficiency of Common Scaling Methods," *J. Sens. Stud.*, Vol. 1, 1986, pp. 85-98.

Likert, R., "A Technique for the Measurement of Attitudes," *Arch. Psychol.*, 1932, p. 140.

MacKay, D. B., "Probabilistic Unfolding Models for Sensory Data," *Food Quality Preference*, Vol. 12, 2001, pp. 427-436.

MacKay, D. B., Easley, R. F., and Zinnes, J. L., "A Single Ideal Point Model for Market Structure Analysis," *J. Marketing Res.*, November 1995, pp. 32, 433-443.

McBride, R.L., "Range Bias in Sensory Evaluation," *J. Food Technol.*, Vol. 17, 1982, pp. 405-410.

McBride, R. L., "Stimulus Range Influences Intensity and Hedonic Ratings of Flavour," *Appetite*, Vol. 6, 1985, pp. 125-131.

McBride, R. L. and Booth, D. A., "Using Classical Psychophysics to Determine Ideal Flavour Intensity," *J. Food Technol.*, Vol. 21, 1986, pp. 775-780.

McKelvie, S. J., "Graphic Rating Scales-How Many Categories?," *Br. J. Psychol.*, Vol. 69, 1978, pp. 185-202.

Meilgaard, M., Civille, G.V., and Carr, B.T., *Sensory Evaluation Techniques*, 4th Edition, CRC Press, Boca Raton, FL, 2006.

Moskowitz, H. R., "Subjective Ideals and Sensory Optimization in Evaluating Perceptual Dimensions in Food," *J. Appl. Psychol.*, Vol. 56, 1972, pp. 60.

Moskowitz, H. R., *Food Concepts and Products: Just In Time Development*, Food & Nutritions Press, Trumbull, CT, 1994.

Moskowitz, H. R., "Learning from the Competition through Category Appraisal: One Practitioner's Keys to Faster and More Efficient Product Development," *Food Service Technol.*, Vol. 1, No. 2, 2001a, pp. 103-118.

Moskowitz, H. R., "Sensory Directionals for Pizza: A Deeper Analysis," *J. Sens. Stud.*, Vol. 16, No. 6, 2001b, pp. 583-600.

Moskowitz, H. R., "On the Analysis of Product Test Results: The Relation among Liking, Sensory and Directional Attributes," http://www.mji-designlab.com/articles/lang6.htm.

Moskowitz, H. R., Munoz, M. S., and Gacula, M. C., *Viewpoints and Controversies in Sensory Science and Consumer Product Testing*, Food & Nutrition Press, Inc., Trumbull, CT, 2003.

Moskowitz H. R., "Just About Right (JAR) Directionality and the Wandering Sensory Unit. In Data Analysis Workshop: Getting the Most Out of Just-About-Right-Data," *Food Quality Preference*, Vol. 15, 2004, pp. 891-899.

Myers, R. H. and Montgomery, D. C., *Response Surface Methodology: Process and Product Optimization using Designed Experiments*, John Wiley & Sons, Inc., New York, 1995.

Pangborn, R. M. and Giovanni, M. E., "Dietary Intake of Sweet Foods and of Dairy Fats and Gustatory Responses to Sugar in Lemonade and to Fat in Milk," *Appetite*, Vol. 5, 1984, pp. 317-327.

Pangborn, R. M., Guinard, J. X., and Meiselman, H. L., "Evaluation of Bitterness of Caffeine in Hot Chocolate Drink by Category, Graphic, and Ratio Scaling," *J. Sens. Stud.*, Vol. 4, 1989, pp. 31-53.

Parducci, A. and Perrett, L. F., "Category Rating Scales: Effects of Relative Spacing and Frequency of Stimulus Values," *J. Exp. Psychol. Monogr.*, Vol. 89, 1971, pp. 427-452.

Pearce, J. H., Korth, B., and Warren, C. B., "Evaluation of Three Scaling Methods for Hedonics," *J. Sens. Stud.*, Vol. 1, 1986, pp. 27-46.

Piggott, J. R., *Sensory Analysis of Foods*, 2nd Edition, Elsevier Applied Science, New York, 1989.

Pokorny, J. and Davidek, J., "Application of Hedonic Sensory Profiles for the Characterization of food Quality," *Die Nahrung*, Vol. 8, 1986, pp. 757-763.

Popper, R., Rosenstock, W., Schraidt, M., and Kroll, B. J., "The Effect of Attribute Questions on Overall Liking Ratings," *Food Quality Preference* Vol. 15, 2004, pp. 853-858.

Popper, R., Schraidt, M., and Kroll, B. J., "When do Attribute Ratings Affect Overall Liking Ratings," presented at the Sixth Pangborn Sensory Sciences Symposium, Harrogate International Center, York, August 7-11, 2005.

Poulton, E. C., "Models for Biases in Judging Sensory Magnitude," *Psychol. Bull.*, Vol. 86, 1979, pp. 777-803.

Riskey, D. R., "Use and Abuses of Category Scales in Sensory Measurement," *J. Sens. Stud.*, Vol. 1, 1986, pp. 217-236.

Rothman, L. R., "The Use of Just-About-Right (JAR) Scales in Food Product Development and Reformulation," in *Consumer Led Food Product Development*, Woodhead Publishing Limited, Cambridge, England, 2007.

Schutz, H. G., "Consumer DataSense and Nonsense," 1999, http://147.46.94.112/journal/sej/full/fqp9907v10i4-503.pdf.

Shepherd, R., Farleigh, C. A., Land, D. G., and Franklin, J. G., "Validity of Relative-to-Ideal Rating Procedure Compared with Hedonic Rating," *Progress in Flavor Research*, 4th Weurman Flavour Research Symposium, 1984.

Shepherd, R., Farleigh, C. A., and Wharft, S. G., "Effect of Quality Consumed on Measures of Liking for Salt Concentrations in Soup," *J. Sens. Stud.*, Vol. 6, 1991, pp. 227-238.

Shepherd, R., Griffiths, N. M., and Smith, K., "The Relationship Between Consumer Preference and Trained Panel Responses," *J. Sens. Stud.*, Vol. 3, 1988, pp. 19-35.

Stewart, D. W., Shamdasani, P. N., and Rook, D. W., *Focus Groups: Theory and Practice*, 2nd Edition, Thousand Oaks California: Sage Publications, 2007.

Stone, H. and Sidel, J. L., *Sensory Evaluation Practices*, 3rd Edition, Academic, San Diego, CA, 2004.

Stone, L. J. and Pangborn, R. M., "Preferences and Intake Measures of Salt and Sugar, and Their Relation to Personality Traits," *Appetite*, Vol. 15, 1990, pp. 63-79.

Takane, Y., "Ideal Point Discriminant Analysis and Ordered Response Categories," *Behaviormetrika*, Vol. 26, 1989, pp. 31-46.

Tang, Chen, Heymann, and Hildegarde, "Multidimensional Sorting, Similarity Scaling and Free-Choice Profiling of Grape Jellies," *J. Sens. Stud.*, Vol. 17, No. 6, 2002, 493-509.

Teas, R. K., "Expectations, Performance Evaluation, and Consumers' Perceptions of Quality," *J. Marketing*, Vol. 57, 1993, pp. 18-34.

Teas, R. K., "Expectations as a Comparison Standard in Measuring Service Quality: An Assessment of a Reassessment," *J. Marketing*, Vol. 58, 1994, pp. 132-139.

Thurstone, L. L., "Attitudes Can Be Measured," *Am. J. Sociol.*, Vol. 33, No. 4, 1928, pp. 529-552.

van Trip, H., Punter, P., Mickartz, F., and Kruithof, L., "The Quest for the Ideal Product," *J. Food Quality Preference*, 2007.

Vickers, Z., "Sensory Specific Satiety in Lemonade using a Just Right Scale for Sweetness," *J. Sens. Stud.*, Vol. 3, 1988, pp. 1-8.

Vickers, Z., Holton, E., and Wang, J., "Effect of Ideal-Relative Sweetness on Yogurt Consumption," *Food Quality Preference*, Vol. 12, No. 8, 2001, pp. 521-526.

Lori Rothman is a Section Manager for Kraft Foods in the Department of Perceptual and

Applied Quantitative Sciences and Innovative Applications, a part of Research and Development. For the past 14 years, Lori has worked in the area of consumer research, conducting both quantitative and qualitative studies for many of the Kraft brands. Lori has a B.S. degree from Cornell University in Nutritional Sciences and an M.S. degree from the University of California, Davis in Food Science. At Davis, Lori researched the language of basic tastes and its transfer to novel tastants with trained panelists. Prior to joining Kraft Foods, Lori worked for Philip Morris (now Altria), researching the sensory impact of flavor degradation in carbonated beverages, for Kellogg's, where she managed the Sensory Evaluation and Shelf Life departments and for Brach's Candies at the inception of its Product Performance group, where she built a state of the art sensory laboratory including shelf life testing chambers and computerized panel evaluation booths. Lori has authored a number of publications in refereed and industry journals and is a frequent speaker at universities and conferences. Lori is a longstanding professional member of the Institute of Food Technologists and its Sensory Evaluation Division as well as ASTM International where she cochairs the Accceptance Preference Task Group. She is an active reviewer for the Journal of Quality and Preference and has authored a chapter on "Just About Right Scales" for the book Consumer Led Food Product Development published in 2007.

Merry Jo Parker has over 25 years experience in applying sensory principles and practices to

consumer research. In 2008 she retired as the founder, owner and CEO of Food Perspectives Inc., a guidance research and consumer insights consulting and testing firm. Founded in 1990, Food Perspectives has clients across the United States from Fortune 500 companies to small emerging companies. Food Perspectives works with product guidance and marketing consumer insights professionals offering a variety of research techniques, ranging from fieldwork to complete research services that include test design, interpretation and consulting.

Prior to founding Food Perspectives Inc. Ms. Parker was, an independent consultant, a Senior Scientist at General Mills Inc. and Research Scientist at Sandoz Nutrition focusing on product development and consumer research on nutritional, food service and retail food products.

Ms. Parker received Bachelors and Masters Degrees in Food Science from the University of Minnesota with an emphasis in sensory science. She has been a member of ASTM since 1999 and serves as the Chairman for the ASTM Subcommittee on Fundamentals of Sensory. In 2007 Ms. Parker co-chaired the 7th Pangborn Sensory Science Symposium. This symposium is considered the most important international scientific gathering for sensory and consumer scientists with over 900 attendees from 54 different countries. She has also been a long standing professional member of the Institute of Food Technologies and IFT's Sensory Evaluation Division.

www.astm.org

ISBN: 978-0-8031-7010-0
Stock #: MNL63