

Unobserved Heterogeneity as an Alternative Explanation for “Reversal” Effects in Behavioral Research

J. WESLEY HUTCHINSON
WAGNER A. KAMAKURA
JOHN G. LYNCH, JR.*

Behavioral researchers use analysis of variance (ANOVA) tests of differences between treatment means or chi-square tests of differences between proportions to provide support for empirical hypotheses about consumer behavior. These tests are typically conducted on data from “between-subjects” experiments in which participants were randomly assigned to conditions. We show that, despite using internally valid experimental designs such as this, aggregation biases can arise in which the theoretically critical pattern holds in the aggregate even though it holds for no (or few) individuals. First, we show that crossover interactions—often taken as strong evidence of moderating variables—can arise from the aggregation of two or more segments that do not exhibit such interactions when considered separately. Second, we show that certain context effects that have been reported for choice problems can result from the aggregation of two (or more) segments that do not exhibit these effects when considered separately. Given these threats to the conclusions drawn from experimental results, we describe the conditions under which unobserved heterogeneity can be ruled out as an alternative explanation based on one or more of the following: a priori considerations, derived properties, diagnostic statistics, and the results of latent class modeling. When these tests cannot rule out explanations based on unobserved heterogeneity, this is a serious problem for theorists who assume implicitly that the same theoretical principle works equally for everyone, but for random error. The empirical data patterns revealed by our diagnostics can expose the weakness in the theory but not fix it. It remains for the researcher to do further work to understand the underlying constructs that drive heterogeneity effects and to revise theory accordingly.

Behavioral researchers use analysis of variance (ANOVA) tests of differences between treatment means or chi-square tests of differences between proportions to provide support for empirical hypotheses about consumer behavior. They typically test data from “between-subjects” experiments in which participants were randomly assigned to conditions. This practice carries an implicit assumption

that, aside from random error, each subject in the sample is a replication from a homogeneous population of consumers. Consumer researchers are well aware that aggregation biases are possible if the treatments interact with some “background variable” associated with individual differences, contexts, situations, or products (Lynch 1982). However, if the researcher lacks the insight to anticipate, measure, and analyze for such interactions, then the problem remains undetected and the resultant data patterns can be highly misleading. This danger is well known in consumer research. For example, most marketing texts warn managers that population means can be misleading and emphasize market segmentation to avoid this problem.

Unfortunately, consumer researchers who know (and perhaps teach) about the dangers of aggregation bias often do not make the connection between this truism and the analysis of their own experimental research data. They propose theories of individual behavior, but they report only group means from between-subjects designs, interpreting these to

*J. Wesley Hutchinson is professor of marketing at the Wharton School of University of Pennsylvania, Philadelphia, PA 19104-6371 (wes@marketing.wharton.upenn.edu). Wagner A. Kamakura is Wendell A. Smith Professor of Marketing at the University of Iowa, College of Business Administration, Iowa City, IA 52242-1000 (wagner-kamakura@uiowa.edu). John G. Lynch, Jr., is Hanes Corporation Foundation Professor of Business Administration at Duke University, Fuqua School of Business, Durham, NC 27708-0120 (john.lynch@duke.edu). The authors thank the associate editor and reviewers, and Phipps Arabie, Julie Edell, Tim Heath, Jordan Louviere, Itamar Simonson, and Bob Veryzer for comments on prior versions of this article. They also thank Jonathan Levav for assistance in data collection.

represent all or most individuals. These researchers correctly assume that internally valid experimental designs avoid certain forms of aggregation bias. Procedures like the random assignment of subjects to conditions minimize spurious correlations with background factors.¹ Although the sample mean is an unbiased estimate of the population mean (as defined by the sample frame of the experiment), researchers err by making untested assumptions about the generality of the observed pattern of treatment means. This is a problem of external validity that threatens construct (nomological) validity (Lynch 1982, 1983). Theory-consistent patterns of aggregate means may not correspond to the true patterns at the individual level, either within the sample or in the populations to which researchers would like to extrapolate. Individual level variation that is not represented well by some simple notion of error variance around a mean is called unobserved heterogeneity. We will propose methods to reduce the likelihood of making false inferences because of inattention to unobserved heterogeneity.

When randomly assigning subjects to conditions, it is common practice to collect ancillary covariate data about participants to reduce within-cell error and to use in post hoc tests to explore the possibility of background factor by treatment interactions. The central problem addressed in this article is that biases due to unobserved heterogeneity can exist even when these standard precautions are taken and the internal validity of the experiment is high. We will show that the problem is serious and can easily lead to conclusions that are distorted or simply wrong.

Interestingly, researchers analyzing naturally occurring data with econometric tools have made significant progress in understanding how to model unobserved heterogeneity (Ailawadi, Gedenk, and Neslin 1998; Allenby, Arora, and Ginter 1998; Kamakura and Russell 1989; Rossi and Allenby 1993). Similarly, marketing research methods that use conjoint analysis to predict consumer response have long emphasized the need to capture heterogeneity of preferences at both the segment and individual levels. Experimental researchers working within an ANOVA tradition have lagged.

Three Basic Problems Resulting from Unobserved Heterogeneity

This article is aimed at experimental researchers who are concerned about external validity—the extent to which internally valid treatment mean differences generalize across subpopulations (Cook and Campbell 1979). The fundamental problem is that statistically significant differences in means may appropriately represent none, some, or all of the subjects in sample data, and none, some, or all of the subpopulations in the larger population of interest. Thus, we refer to this as the none-some-all problem.

Researchers who do not investigate this problem directly

¹One of the earliest and most famous forms of aggregation biases of this type, that is, stemming from a confounding of unobserved individual difference characteristics with treatments, is called Simpson's paradox (Simpson 1951).

exhibit what we call the ignorance-is-bliss problem. They implicitly assume that their treatment mean results are representative of the individuals in their sample and in the population, but they do not test this assumption. Cook and Campbell (1979) and Lynch (1982, 1983) therefore recommended several methodological procedures for testing directly for background factor by treatment interactions. In particular, these authors discuss the advantages and disadvantages of explicitly manipulating certain background factors while holding others constant as compared to post hoc analyses of data for which background factors vary naturally because of representative sampling. Calder, Phillips, and Tybout (1982), however, referred to this advice as a “counsel of despair” because the search for the right background variables to use in such analyses is unending. Moreover, the search for the right set of variables is made difficult by the need to examine all possible interactions between treatments and background variables. Thus, in most situations there are simply too many possibilities to search explicitly for moderators, and it is very likely that key sources of heterogeneity in the population will remain unobserved. We refer to this as the needle-in-a-haystack problem.

In summary, heterogeneity in individual level responses to experimental treatments creates a none-some-all problem in interpreting aggregate analyses, especially simple averages. Researchers who are unaware of the potential effects of heterogeneity suffer from the ignorance-is-bliss problem. Researchers who are aware of heterogeneity effects and attempt to incorporate them into their analyses using manipulated or measured background variables face the needle-in-a-haystack problem. Our goal is to solve all three problems.

Approaches to Solving the Problems of Unobserved Heterogeneity

In this article, we draw on classic advice about data analysis (e.g., Tukey 1977) and recent advances in econometric modeling of unobserved heterogeneity (e.g., De Sarbo and Cron 1988; Wedel and Kamakura 1997) to attack the none-some-all problem in experimental data. These methods solve the needle-in-a-haystack problem because they do not require the researcher to anticipate, measure, and model the specific sources of heterogeneity contributing to none-some-all problems. Thus, they also solve the “counsel of despair” critique of Calder et al. (1982). Calder et al. (1982) and Lynch (1982, 1983) stressed the need for within-cell homogeneity in background factors to increase statistical power and avoid the ignorance-is-bliss problem. The methods we describe allow the researcher to avoid the ignorance-is-bliss problem without loss of statistical power even when unmeasured background factors vary within treatments and interact with those treatments.

The proposed diagnostics require within-subjects designs. When these diagnostics provide evidence of subgroup differences in response to experimental manipulations, we can reject the assumption of homogeneity (except for random

error) required by standard statistical analyses. Consequently, finding evidence of unobserved heterogeneity should then trigger an effort to explain the person by treatment interactions, presumably by revising one's theory and conducting further research. Thus, variation due to unmeasured factors can become an asset insofar as it reveals heterogeneity and advances subsequent research and theory development.

AN ILLUSTRATION OF THE PROBLEM IN A 2 × 2 FACTORIAL DESIGN

We illustrate the general problem via a hypothetical experiment described in Table 1. In this example, a 2 × 2 between-subjects design manipulating price (low vs. high) and the type of advertising appeal (image-oriented vs. quality-oriented) was used to assess the effects of these variables on brand attitude (see part A). To simplify our discussions, we introduce the following notation to represent the general

linear model for 2 × 2 experimental designs (whether within or between subjects):

$$R_{ijk} = \beta_{0k} + \beta_{rk}X_r(i) + \beta_{ck}X_c(j) + \beta_{rck}X_r(i)X_c(j) + \epsilon_{ijk}, \tag{1}$$

where *i* and *j* index the row and column experimental conditions, respectively (i.e., appeal and price in Table 1), *k* indexes individuals, R_{ijk} is the observed dependent measure for condition (*i, j*) for individual *k*, X_r is a contrast-coded dummy variable for Row, X_c is a contrast-coded variable for Column, β_{0k} , β_{rk} , β_{ck} , and β_{rck} are model parameters, and ϵ_{ijk} is normal error (which is here assumed to be independently, but not necessarily identically, distributed across *i, j,* and *k*). Thus, with the contrast coding shown in part A of Table 1, the model equation can be simplified as follows:

TABLE 1

HYPOTHETICAL EXPERIMENT (A) CONTRAST CODING OF CELLS, MEANS, AND REGRESSION PARAMETERS SHOWN FOR (B) WHEN UNOBSERVED HETEROGENEITY DUE TO GENDER IS IGNORED OR (C) WHEN COEFFICIENTS ESTIMATED WITHIN GENDER SEGMENTS

				Cell mean	Intercept	Row	Column	Interaction
A. Experimental design:								
	Price high	Price low		R_{11}	1	1	1	1
Image appeal	R_{11k}	R_{12k}		R_{12}	1	1	-1	-1
Quality appeal	R_{21k}	R_{22k}		R_{21}	1	-1	1	-1
				R_{22}	1	-1	-1	1
				Coefficients ignoring gender				
	Price high	Price low	Row average		Intercept, β_0	Row, β_r	Column, β_c	Interaction, β_{rc}
B. Aggregate results:								
Total sample (<i>n</i> = 48 per cell)								
	Image appeal	3.0	3.0	3.0				
	Quality appeal	3.0	3.0	3.0				
	Column average	3.0	3.0	3.0	3.0	.0	.0	.0
C. Unobserved segments:								
Males (<i>n</i> = 24 per cell)					Male coefficients			
	Image appeal	3.0	.0	1.5				
	Quality appeal	6.0	5.0	5.5				
	Column average	4.5	2.5	3.5	3.5	-2.0	1.0	.5
Females (<i>n</i> = 24 per cell):					Female coefficients			
	Image appeal	3.0	6.0	4.5				
	Quality appeal	.0	1.0	.5				
	Column average	1.5	3.5	2.5	2.5	2.0	-1.0	-.5

$$R_{11k} = \beta_{0k} + \beta_{rk} + \beta_{ck} + \beta_{rck} + \epsilon_{11k}, \quad (2)$$

$$R_{12k} = \beta_{0k} + \beta_{rk} - \beta_{ck} - \beta_{rck} + \epsilon_{12k}, \quad (3)$$

$$R_{21k} = \beta_{0k} - \beta_{rk} + \beta_{ck} - \beta_{rck} + \epsilon_{21k}, \quad (4)$$

$$R_{22k} = \beta_{0k} - \beta_{rk} - \beta_{ck} + \beta_{rck} + \epsilon_{22k}. \quad (5)$$

Part B shows aggregate results from the hypothetical experiment (i.e., when gender is ignored). Part C of Table 1 shows means and coefficients for two unobserved segments, men and women. Let us assume that the obtained sample sizes resulted from probability sampling from some population with equal numbers of men and women combined with random assignment of individuals to conditions.

In this extreme example, the manipulations have equal and opposite effects for each of the unobserved segments. We refer to this type of aggregation bias as effect cancellation. Because the sample in each cell is a probability sample, the expected value of the sample means matches the population means, as do the associated regression coefficients. However, they do not describe any individual (or segment) in the population. A standard ANOVA of these data would find no significant effects nor would there be any indicator of a heterogeneity problem. It would be natural for the researcher to reach a conclusion of null results based on such analyses, exemplifying the ignorance-is-bliss problem. In this particular example, ignorance leads to type 2 errors (i.e., the researcher falsely concludes that appeal and price do not matter). We subsequently show how similar aggregation biases can create type 1 errors as well.

The problem of effect cancellation is very general. However, one specific form is of particular interest to behavioral researchers because it calls into question the conclusions that are typically drawn from crossover interactions. In the next section, we discuss this problem in detail and use it to introduce our proposed solutions.

CROSSOVER INTERACTIONS

Crossover interactions of theoretical variables are often taken as definitive evidence of moderating effects (e.g., Tybout 1995). In terms of the general linear model introduced earlier, if $\max\{|\beta_r|, |\beta_c|\} > |\beta_{rc}| > \min\{|\beta_r|, |\beta_c|\}$, then exactly one of the main effects reverses its sign when the level of the other main effect changes (here called a single crossover interaction).

To understand this mathematical point, consider the example of classic tests of the elaboration likelihood model. Petty and Cacioppo (1979) showed that involvement increased persuasion for pro-attitudinal messages, with mean attitude scores of 1.6 and 0.6 for high and low involvement, respectively. Involvement decreased persuasion for counter-attitudinal messages, with means of -1.8 and -0.4 for high and low involvement, respectively (see the bottom panel of Fig. 1). Pro-attitudinal messages induce more positive attitudes than counter-attitudinal ones at all levels of involvement, so the sign of the simple effect of message does not

“cross over.” However, the sign of the effect of involvement changes with message type. Consequently, the absolute value of the parameter for the main effect of message ($|\beta_r| = |1.1|$) exceeds that for the interaction ($|\beta_{rc}| = |0.6|$), which, in turn, exceeds that for the main effect of involvement ($|\beta_c| = |-0.1|$).

A different inequality characterizes double crossover interactions, wherein the simple effects of the row variable change signs at different levels of the column variable, and the simple effect of the column variable changes signs at different levels of the row variable. It can be readily shown that for double crossover interactions, $|\beta_{rc}| > \max\{|\beta_r|, |\beta_c|\}$. Keppel (1991, p. 235) provides a related discussion of “ordinal” and “disordinal” interactions. These inequalities undergird the diagnostic tests for unobserved heterogeneity that we will present in the remainder of the article.

Unobserved Heterogeneity and Nonlinear Output Functions Combine to Create Spurious Aggregate Crossover Interactions

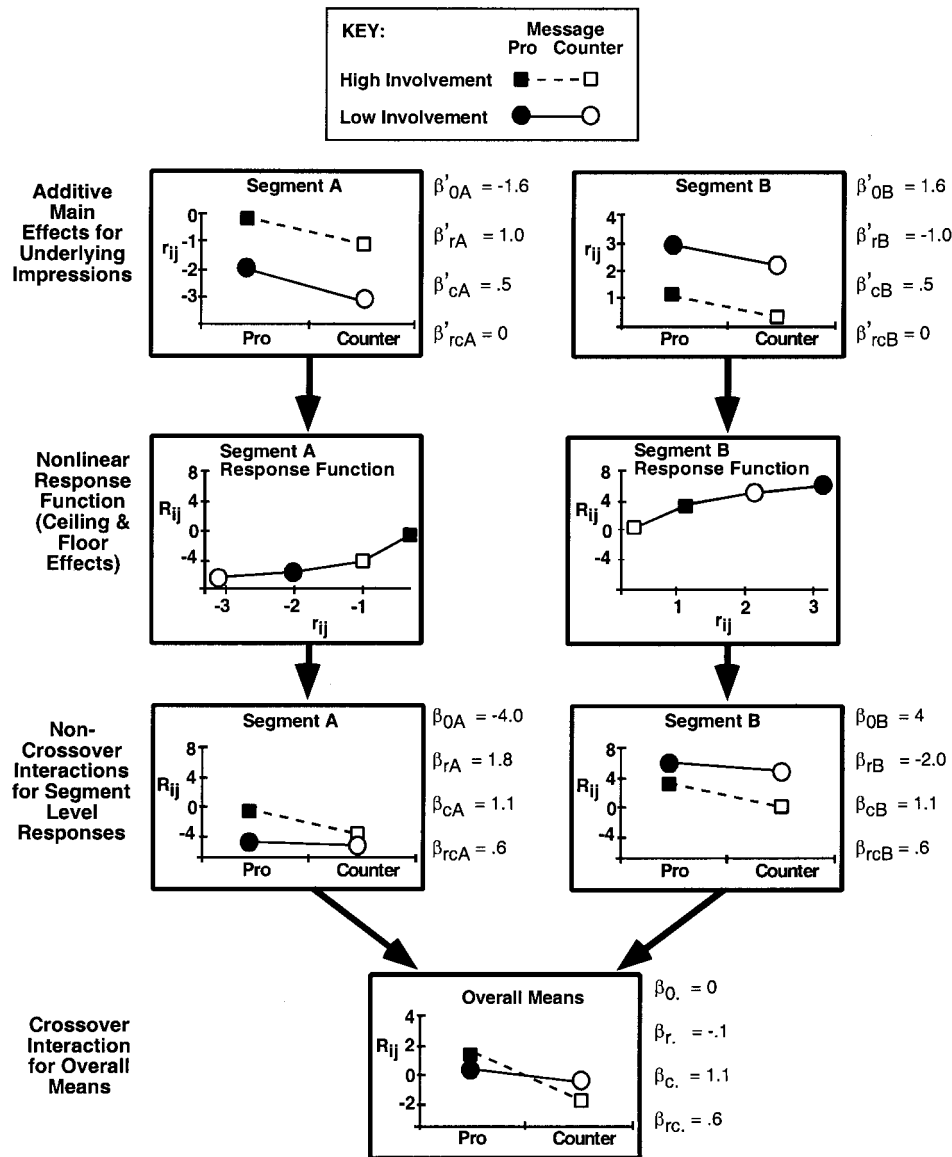
There is a good reason why experimental results are persuasive when predicted crossover interactions are observed. First, as Sternthal, Tybout, and Calder (1994) and Tybout (1995) have noted, it is often more difficult to pose alternative explanations for findings that have crossover interactions than for those relying on main effects. Second, if the four values observed in a 2×2 design exhibit a (single or double) crossover interaction, no monotonic transformation of these values can remove the interaction or change its sign (Krantz and Tversky 1971; Lynch 1985). We label this property “crossover invariance.”

Assume that measured responses are related to underlying and unobservable impressions through some type of “output function,” $R_{ij} = O(r_{ij})$, where r_{ij} is an impression resulting from row condition i and column condition j , and O is the monotonic function mapping underlying impressions into overt responses, R_{ij} . At the level of a specific individual, a crossover interaction guarantees that an interaction in the R_{ij} is also present in the true underlying impressions, r_{ij} . Alternatively, an interaction that does not exhibit a crossover may result from the output function alone despite no interaction in the underlying impressions (Lynch 1985). Because there is much research to suggest that the functions mapping psychological impressions into observed responses are nonlinear (e.g., Gescheider 1988), robust indicators, such as crossover interactions, are important.

Previous discussions of this property of crossover interactions have not assessed the problem of unobserved heterogeneity. Unfortunately, crossover invariance does not hold for means when individual-level transformations are allowed, even if the transformation $O(r)$ has the same functional form for each individual. Heterogeneity in individual responses can produce mean values exhibiting a crossover interaction, even though no individual exhibits such an interaction. Thus, crossover interactions of aggregate means

FIGURE 1

AN ALTERNATIVE EXPLANATION OF PETTY AND CACIOPPO'S (1979) SINGLE CROSSOVER INTERACTION



NOTE.—Based on unobserved heterogeneity coupled with a nonlinear output function, mapping from additive unobservable impressions, r_{ij} to overt ratings, $R_{ij} = O(r_{ij})$.

may arise from nonlinear output functions applied to psychological impressions that exhibit main effects but no interaction.

Figure 1 illustrates how such misleading means can arise using the Petty and Cacioppo (1979) results discussed earlier. In this example, the underlying impressions exhibit main effects of message and involvement, but no interaction. The effect coefficients for each segment (labeled β' in Fig. 1) are the same for the main effect of message and the

interaction (i.e., $\beta' = \beta' = 0.5$ and $\beta' = \beta' = 0$), but are equal and opposite for the main effect of involvement and the intercept (i.e., $\beta' = 1$, $\beta' = -1$, $\beta' = -1.6$, and $\beta' = 1.6$). If the response function was linear, the opposite effects would cancel and only the main effect of message would be nonzero in the aggregate means. However, as can be seen in Figure 1, if the output function introduces so-called floor and ceiling effects, the floor effect creates a positive interaction in the observed responses for segment

A and the ceiling effect creates a positive interaction in the observed responses for segment B. The interaction is not a crossover for either segment (i.e., $\min\{|\beta_{rk}|, |\beta_{ck}|\} > |\beta_{rc}|$) because monotonic output functions cannot create crossover interactions. When aggregate means are computed, however, the main effects of involvement cancel, but the interactions do not. This results in a spurious single crossover in the aggregate means.

One interpretation of our example is that the effect of high involvement is to reduce each segment's reliance on prior attitude, rather than to increase "central processing" of message arguments as argued by Petty and Cacioppo (1979) and many later researchers. Thus, segments A and B have opposite prior attitudes and involvement results in a "regression" toward neutrality (in addition to the content-based effect of message). We do not assert that our interpretation is preferable to the standard account. It is, however, plausible and demonstrates that consideration of unobserved heterogeneity can lead to very different inferences about theoretical processes.

It is important to note that we have used nonlinear output functions in this example because they are a common concern among behavioral researchers.² However, heterogeneity can arise from a variety of sources and lead to similarly spurious aggregate results. In general, effect cancellation can create spurious effects of any type. For example, the null results in Table 1 change to a spurious double crossover interaction when $\beta_{rcMale} = \beta_{rcFemale} > 0$, because opposite main effects in the two segments would cancel, but the interaction now would not, so that $|\beta_{rc}| > \max\{|\beta_r|, |\beta_c|\}$.

Sign Homogeneity

How can researchers rule out unobserved heterogeneity and nonlinear output functions as an explanation for a crossover interaction observed in aggregate means? In a between-subjects experiment with random assignment of subjects to groups, the tools available are limited. In some cases, unobserved heterogeneity leads to unequal variance across the cells of the design (i.e., heteroscedasticity). Thus, the standard tests for homogeneity of variance might alert the researcher to potential problems. Unfortunately, unequal within-cell variance is neither necessary nor sufficient as an indicator of aggregation bias. Many patterns of heterogeneity exert only minimal effects on the pattern of within-cell variances (particularly when there are more than two underlying segments), and whatever differences are present may be masked by error variance. Conversely, differences in within-cell variance can arise from a homogeneous population with a heteroscedastic error distribution.

Unobserved heterogeneity can be ruled out a priori as an alternative explanation of a crossover interaction in aggregate means if it is known or reasonably assumed that the

²A misleading double crossover interaction can arise when both of the underlying main effects are reversed across segments and the output function is concave for both segments. See Hutchinson, Kamakura, and Lynch (1999) for an illustration.

main effects are homogeneous with respect to sign (but not necessarily magnitude). More specifically, if sign homogeneity holds then we can assume without loss of generality that all main effects are positive, $0 < \min\{\beta_{rk}, \beta_{ck}\}$. If for each segment, k , $|\beta_{rc}| < \min\{\beta_{rk}, \beta_{ck}\}$ (i.e., there are no segment level crossover interactions), then it is easy to show that $|\sum_k \pi_k \beta_{rc}| < \min\{\sum_k \pi_k \beta_{rk}, \sum_k \pi_k \beta_{ck}\}$ (where π_k is the proportion of individuals in segment k), and, therefore, the means do not exhibit a crossover interaction.³ Thus, if a crossover interaction is observed in the means and sign homogeneity holds, then at least some individuals must exhibit a crossover interaction of the same type.

Even when sign homogeneity holds, the appearance of an aggregate crossover interaction does not guarantee that all, or even most, subjects exhibit a crossover interaction. Therefore, sign homogeneity rules out unobserved heterogeneity as a "complete" explanation of an observed crossover interaction, but the aggregate means may nevertheless poorly represent individual-level behavior. Put differently, it reduces the none-some-all problem to a some-all problem with respect to whether crossover interactions exist. In addition, in most situations, the main effects are also being submitted to empirical test. In fact, many experiments are designed to discriminate between competing models on the basis of the observed sign of a main effect. Because individuals may differ in the model that represents them best, sign homogeneity cannot be assumed.

In general, the researcher must use prior results about experimental treatments to determine whether or not sign homogeneity can be assumed. It is not sufficient, however, that most or even all prior research has found main effects of the same sign; prior research must also have ruled out sign heterogeneity for those effects. For example, if price differences are an experimental treatment, pretesting might demonstrate that for the stimuli being used all subjects prefer lower prices and that no subjects engage in price quality inferences.

Current Practice in Behavioral Research Is at Risk

In recent years, the *Journal of Consumer Research* (March 1996–September 1998) and the *Journal of Marketing Research* (February 1996–February 1999) have published 45 articles in which two-way crossover interactions were reported as key hypothesis tests.⁴ Of the 45 articles we identified, 32 used a completely between-subjects design, nine used a mixed design in which only one key factor was repeated within subjects, and five manipulated both factors

³Our claim follows from $|\sum_k \pi_k \beta_{rc}| \leq \sum_k \pi_k |\beta_{rc}| < \sum_k \pi_k \min\{\beta_{rk}, \beta_{ck}\} \leq \min\{\sum_k \pi_k \beta_{rk}, \sum_k \pi_k \beta_{ck}\}$. The first inequality holds because for all k , $\beta_{rc} \leq |\beta_{rc}|$. The second inequality follows from the assumption that for all k , $|\beta_{rc}| < \min\{\beta_{rk}, \beta_{ck}\}$. The third inequality holds because for all k , $\min\{\beta_{rk}, \beta_{ck}\} \leq \beta_{rk}$ and $\min\{\beta_{rk}, \beta_{ck}\} \leq \beta_{ck}$ (which implies $\sum_k \pi_k \min\{\beta_{rk}, \beta_{ck}\} \leq \sum_k \pi_k \beta_{rk}$ and $\sum_k \pi_k \min\{\beta_{rk}, \beta_{ck}\} \leq \sum_k \pi_k \beta_{ck}$).

⁴The designs were mainly 2^k. Virtually all articles predicted and observed crossover interactions; however, at least one affirmed the null hypothesis as a test of method validity.

within subjects in at least one experiment. Subsequently, we describe various statistical tests that can be used to rule out unobserved heterogeneity as an alternative explanation. These methods require within-subjects observations for all effects of interest, however. Thus, such alternative explanations cannot be rejected on the basis of observed data for 91 percent of these articles. Although we will describe several nonstatistical approaches that do not require within-subjects observations, and between-subjects designs have many strengths, one conclusion of the present work is that the current trend in behavioral research away from within-subjects designs deserves serious scrutiny. We should note that within-subjects designs, when used alone, are just as prone to aggregation biases as are between-subjects designs. It is only when within-subjects designs are coupled with the diagnostic tests we will describe below that they permit a solution to the problem of unobserved heterogeneity.

In the following section, we use experimental data to illustrate (a) how aggregation can bias observed means, (b) how nonparametric statistical tests can assess the extent of individual-level interactions, and (c) how latent class modeling can be used to test the homogeneity assumption of ANOVA and, if heterogeneity problems are indicated, to provide exploratory information about segment specific effects.

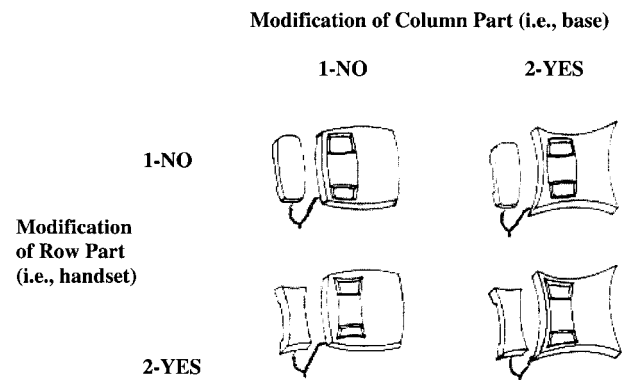
An Empirical Example: The Effects of Unity on Aesthetic Response

Veryzer and Hutchinson (1998) report the results of several experiments on judgments of the visual attractiveness of product designs. In one experiment, they used a within-subjects design and a large number of different sets of stimuli. This design provides a rich source of data for illustrative purposes because we can “create” unobserved heterogeneity by ignoring differences between stimuli. Also, we can search across the different stimulus sets to find particularly instructive examples. Overall, the analyses we report reinforce their conclusions and raise interesting problems for future research on heterogeneity in aesthetic response. They do not imply that the published research was flawed because of unobserved heterogeneity.

In Veryzer and Hutchinson’s second experiment, 27 sets of line drawings were used. One set, telephones with shape modifications, is illustrated in Figure 2. Each set consisted of four line drawings and was generated according to the following factorial design of stimuli (S_{ij}). First, two distinct parts (referred to here as row and column parts) and a common visual attribute of each were identified in a line drawing of an existing product. Then, the drawing was either left in its original form (called the prototype, S_{11}), or modified in one of three ways—by changing only the column part with respect to the visual attribute (S_{12}), by changing only the row part (S_{21}), or by changing both parts in the same way (S_{22}). Thus, the four stimuli in each set (i.e., the original prototype and the three new products) resulted from a 2×2 factorial design in which the two factors were mod-

FIGURE 2

AN EXAMPLE OF THE STIMULI USED BY VERYZER AND HUTCHINSON (1998)



ification of the row part and modification of the column part.

Most prior research suggested that familiar product designs would be preferred (e.g., Martindale, Moore, and West [1988]; Nedungadi and Hutchinson [1985]; see Veryzer and Hutchinson [1998] for a full discussion). Thus, in the present analysis, original parts are coded +1 and modified parts are coded -1. A familiarity effect is indicated when $\beta_r > 0$ and $\beta_c > 0$ (i.e., when respondents prefer alternatives that have the same level as the prototype stimulus, S_{11} , on the row and column attributes, respectively). However, there was also prior support for novelty effects (i.e., $\beta_r < 0$ and $\beta_c < 0$; e.g., Meyers-Levy and Tybout 1989); therefore, sign homogeneity could not be assumed. Regardless of the main effects, the aesthetic principle of unity suggested that there should be a positive effect when parts matched with respect to the visual attribute used in the design (i.e., $\beta_{rc} > 0$; this interaction was called the unity effect).

In the original Veryzer and Hutchinson (1998) experiment, the use of standard rating scales led to many ties (i.e., two or more of the four stimuli given the same rating). The presence of ties significantly increases the complexity of the analyses we propose. Therefore, we replicated the experiment with a more refined rating scale and instructions that subjects use a unique response for each of the four stimuli in a given set. Preliminary analyses of the Veryzer and Hutchinson (1998) data identified three stimulus sets that were suitable for our purposes because they varied in their degree of heterogeneity. These sets were dressers with two drawers as parts that were modified by adding trim, telephones with base and handset as parts that were modified by changing shape (as in Fig. 2), and TV remote controls with the front end and the back end as parts that were modified by changing the shape from rectangular to rounded.⁵

⁵For these experiments, new line drawings were created that closely resembled those of Veryzer and Hutchinson (1998).

TABLE 2

A COMPARISON OF INTERVAL-SCALE AND ORDINAL ANALYSES OF HETEROGENEITY IN TYPES OF INTERACTION

Interval-scale analysis				Ordinal analysis			Item means			
Type of interaction	β_r	β_c	β_{rc}	Type of interaction	%	β_{rc}	R_{11}	R_{12}	R_{21}	R_{22}
Dressers:										
Average	-.22	-.34	1.93	Average		1.93	1.80	-1.38	-1.62	2.93
				Double	77	2.41	2.29	-2.10	-2.36	2.92
Positive ^a	28	23	87	Single	6	1.04	1.33	1.00	1.00	4.83
Zero ^b	15	18	6	Neutral	13	-.08	-.50	1.42	.67	2.25
Negative ^c	57	59	6	R-single	1	-.75	-2.00	-1.00	3.00	1.00
				R-double	1	-1.00	2.00	4.00	5.00	3.00
Sign test	.94***			Crossovers test	.98***					
Telephones:										
Average	1.11	.61	.73	Average		.73	1.86	-.81	-1.81	-1.58
				Double	23	1.79	.84	-2.32	-2.45	1.55
Positive	75	68	69	Single	20	2.11	5.67	-1.11	-2.53	-.84
Zero	13	7	11	Neutral	45	1.55	7.00	-.65	-1.84	-3.30
Negative	13	24	20	R-single	9	-1.28	.00	2.33	.78	-2.00
				R-double	2	-2.00	-2.00	1.00	1.00	-4.00
Sign test	.78***			Crossovers test	.79***					
TV remote controls:										
Average	-1.05	.11	.27	Average		.27	.16	-.60	1.72	2.04
				Double	16	1.53	.53	-1.47	-1.20	2.93
Positive	17	43	53	Single	24	1.35	1.83	-1.87	1.43	3.13
Zero	11	14	8	Neutral	34	-.21	-1.67	-.75	2.31	2.38
Negative	73	43	39	R-single	17	-2.83	-8.00	.94	2.94	.56
				R-double	9	-3.17	-7.00	1.89	3.00	-.78
Sign test	.58 ^{NS}			Crossovers test	.60 ^{NS}					
Pooled data:										
Average	-.05	.13	.98	Average		.98	1.27	-.93	-.57	1.13
				Double	39	2.17	1.76	-2.05	-2.22	2.65
Positive	40	45	70	Single	17	1.61	3.29	-1.21	-.19	1.77
Zero	13	13	8	Neutral	31	.67	2.78	-.40	.03	-.45
Negative	47	42	22	R-single	9	-2.21	-5.00	1.35	2.19	-.31
				R-double	4	-2.79	-5.42	1.92	2.83	-1.00
Sign test	.76***			Crossovers test	.81***					

^aPercentage of observations with a positive effect.
^bPercentage of observations with an effect exactly equal to zero.
^cPercentage of observations with a negative effect.
*** $p < .0001$.
^{NS} $p > .05$.

Participants used a 19-point scale (-9 to +9) to rate the visual attractiveness of the stimuli. Each set of four product designs was presented on the same page, and the presentation order of sets was varied across subjects according to a Latin Square design. To simplify subsequent discussions, we will use the term “observation” to mean the response vector, $R_{kt} = (R_{11kt}, R_{12kt}, R_{21kt}, R_{22kt})$, of judgments given by the individual, k , to the members of a stimulus set, t .

The unity effect implies a positive value of β_{rc} , and 70 percent of all observations exhibited such an effect ($\beta_{rc} = 0$ for 8 percent and $\beta_{rc} < 0$ for 22 percent; β_{rc} was positive for 87 percent, 69 percent, and 53 percent of observations for dressers, telephones, and TV remote controls, respectively; see Table 2). However, as discussed earlier, if monotonic transformations of the data are allowed, then noncrossover interactions—even at the individual level—can be observed when there is no true unity

effect. Consequently, more convincing evidence of unity comes from crossover interactions.⁶

The overall mean ratings of visual attractiveness (across all three stimulus sets) exhibited a clear double crossover interaction that was consistent with a unity effect and was statistically significant in a traditional mixed effects ANOVA ($F(1, 83) = 146.0, p < .0001$): $R_{11} = 1.27, R_{12} = -.93, R_{21} = -.57, R_{22} = 1.13$, and $\beta_{rc} = .98$. However, the unity effect and the main effects of part modification significantly interacted with product. In such cases, it is typical for researchers to conduct separate ANOVAs

⁶This could result from nonlinearity in the output functions; however, Veryzer and Hutchinson (1998) wanted to rule out the possibility of interactions resulting from nonlinear (but monotonic) familiarity effects without genuine unity effects. Their analysis of an individual level measure, U' , is conceptually similar to the tests described here.

for each product. When this was done, the unity effect for all three products was found to be in the predicted direction, at least a single crossover (i.e., $|\beta_{rc}| > \min\{|\beta_r|, |\beta_c|\}$, see Table 2), and statistically significant ($F(1, 83) = 209.8$, $F(1, 83) = 201.6$, and $F(1, 83) = 27.4$, for dressers, telephones, and TV remote controls, respectively). It is tempting to end the analysis here and conclude that unity effects occur regardless of product type. As our subsequent analyses reveal, this exemplifies the none-some-all problem as well as the ignorance-is-bliss problem.

To probe more deeply whether a crossover interaction in the aggregate is mirrored by most individual respondents, one can examine each individual's set of the four ratings for each stimulus set and count the number of crossover interactions. This approach is consistent with the general advice of statisticians like John Tukey (1977) to examine the distribution of the raw data in addition to various summary statistics.

Nonparametric Tests

Given the hypothesis of a positive crossover interaction (i.e., a unity effect), the strongest support for this hypothesis is provided by sets of ratings that are ordered such that both diagonal cells (i.e., R_{11} and R_{22}) are larger than both off-diagonal cells (i.e., R_{12} and R_{21}). These orderings provide unambiguous support for a double crossover interaction at the individual level. There are four such orderings:

$$\begin{aligned} R_{11} &> R_{22} > R_{12} > R_{21}, \\ R_{11} &> R_{22} > R_{21} > R_{12}, \\ R_{22} &> R_{11} > R_{12} > R_{21}, \\ R_{22} &> R_{11} > R_{21} > R_{12}. \end{aligned}$$

Overall, there are $4! = 24$ unique orderings of four values. Each can be classified according to its support for or against the hypothesized interaction. Four orderings support a single crossover interaction:

$$\begin{aligned} R_{11} &> R_{12} > R_{22} > R_{21}, \\ R_{11} &> R_{21} > R_{22} > R_{12}, \\ R_{22} &> R_{12} > R_{11} > R_{21}, \\ R_{22} &> R_{21} > R_{11} > R_{12}. \end{aligned}$$

Eight orderings are neutral insofar as they support either main effects only or noncrossover interactions, and eight orderings support a single or double crossover interaction in the opposite direction.⁷ The remaining orderings support main effects or noncrossover interactions. The ordinal anal-

ysis section of Table 2 presents the relative frequencies with which each type ordering was observed in the data.

The purpose of these classifications is primarily descriptive. Table 2 shows that across the three products 56 percent (i.e., 39 percent + 17 percent) of the observed orderings were consistent with the crossover interaction predicted by Veryzer and Hutchinson (1998; i.e., a unity effect), 31 percent were neutral, and 13 percent represented reverse crossover interactions. Thus, their hypothesis about unity is supported for most, but not all, observations.

Before discussing statistical tests for the presence of such interactions, it is instructive to observe the biasing effects of averaging revealed in Table 2. As mentioned earlier, the overall means exhibit a positive double crossover interaction that is large relative to the main effects. In actuality, only 39 percent of the observations exhibited a positive double crossover, and most of these occurred for the dresser stimuli. The main effects are small because they differ in sign across products and, therefore, cancel in the pooled data. Thus, assuming that the overall means are representative of all or even most individuals is incorrect and illustrates the risk of misinterpretation (i.e., the none-some-all problem) when only overall means are reported (as is commonly done in our field). A complete theory for these data should explain not only the observed means but also the determinants of the heterogeneity that is evident in Table 2. Some of the variation is due to interactions with product category; however, heterogeneity is evident even within each stimulus set. For example, the percentage of subjects exhibiting the same type of interaction as the aggregate means ranges from 77 percent to 20 percent across the three stimulus sets. Perhaps, this variation is merely the result of random disturbances in psychological impressions or of simple measurement error. Alternatively, there might be systematic sources of heterogeneity. We propose two approaches to separating these potential explanations: nonparametric statistics and latent class modeling.

To provide a nonparametric test of the prevalence of crossover interactions across individuals, we adopt a traditional statistical approach.⁸ If there is no true interaction and the observed individual-level interactions are due to error from some symmetric distribution, then approximately equal numbers of observations should exhibit positive and negative crossover interactions. Thus, one can compute the number of orderings in Table 2 supporting a positive crossover interaction as a proportion of the orderings supporting either a positive or a negative crossover (i.e., nonneutral orderings). Then the standard z -test for proportions can be used to test the hypothesis that the true proportion is 50 percent. We call this the crossovers test. The crossovers test is the ordinal analog of the traditional sign test applied to

⁷The reversed double crossovers are $R_{12} > R_{21} > R_{11} > R_{22}$, $R_{12} > R_{21} > R_{22} > R_{11}$, $R_{21} > R_{12} > R_{11} > R_{22}$, and $R_{21} > R_{12} > R_{22} > R_{11}$, and the reversed single crossovers are $R_{12} > R_{11} > R_{21} > R_{22}$, $R_{12} > R_{22} > R_{21} > R_{11}$, $R_{21} > R_{11} > R_{12} > R_{22}$, and $R_{21} > R_{22} > R_{12} > R_{11}$.

⁸Nonparametric tests are robust and easily adapted to individual level hypotheses. However, they are less statistically powerful than tests that are based on specific assumptions about error distributions. Therefore, small but reliable deviations from homogeneity may not be detected by nonparametric tests. Subsequently, we discuss parametric methods using latent class and random coefficients modeling.

β_{rc} (see Guilford and Fruchter 1973). This sign test is also reported in Table 2.

Results of these tests for the aesthetic response data are given in the ordinal analysis section of Table 2. For example, the crossover test value for the pooled data is 81 percent and is computed as (39 percent + 17 percent)/(39 percent + 17 percent + 9 percent + 4 percent). When computed separately, the crossovers test is significant for dressers and telephones, but not for TV remote controls. Recall that the ANOVA found the unity effect interaction to be significant for TV remote controls. The sign test, the crossovers test, and the relatively large proportion of reverse crossovers (17 percent + 9 percent = 26 percent) indicate that the risk of making a none-some-all error is high for the TV remote controls data if only aggregate statistics are examined.

Overall, these results clearly show that a positive aesthetic response to unity is very common. However, this does not imply that the mean ratings are representative of all or even most participants. Recall that only 39 percent of observations exhibited a positive, double crossover interaction. In the next section, we show how these nonparametric tests can be followed up with more powerful parametric tests when the former indicate heterogeneity may be a problem.

Latent Class Modeling of Crossover Interactions

As noted earlier, if the primary source of heterogeneity is some easily observed variable, like gender or product category, it can be used as a blocking variable to remove aggregation biases. However, in most cases one is confronted with the needle-in-a-haystack problem of searching for the right blocking variables, making it likely that heterogeneity will be unobserved. Latent class modeling provides a solution to the needle-in-a-haystack problem because explicitly measured characteristics of the subjects are not required. For our purposes, the main objectives of the analysis are to test the homogeneity assumed by traditional ANOVA, and, if homogeneity is rejected, provide exploratory information about the sources of heterogeneity.

We achieve these objectives by applying a finite mixture of linear regressions (DeSarbo and Cron 1988), also known as latent class regressions. Instead of fitting a single linear model to the data from all subjects as in ANOVA, we assume that there might be S groups of subjects, with the members of each group sharing a common linear response function. Therefore, if a particular subject k belongs to a group s , then responses, R_{ijk} , from the 2×2 design discussed earlier would be modeled as follows:

$$R_{11k} = \beta_{0s} + \beta_{rs} + \beta_{cs} + \beta_{rcs} + \epsilon_{11ks}, \tag{6}$$

$$R_{12k} = \beta_{0s} + \beta_{rs} - \beta_{cs} - \beta_{rcs} + \epsilon_{12ks}, \tag{7}$$

$$R_{21k} = \beta_{0s} - \beta_{rs} + \beta_{cs} - \beta_{rcs} + \epsilon_{21ks}, \tag{8}$$

$$R_{22k} = \beta_{0s} - \beta_{rs} - \beta_{cs} + \beta_{rcs} + \epsilon_{22ks}. \tag{9}$$

Because the group-specific intercepts are not of substantive interest, we simplify the analyses reported here through the mean-centering of the dependent variable within each subject for each stimulus set (therefore, β_{0s} is not estimated). Also, although nonlinear output functions could be estimated, this adds model parameters and is not substantively necessary insofar as we can test for crossover interactions within each group. Thus, for these type of data the assumption that the model is linear is not particularly restrictive.

The probability that subject k with response vector $R_k = (R_{11k}, R_{12k}, R_{21k}, R_{22k})$ belongs to group s is given by

$$P[k \in s | y_k] = \frac{\pi_s L[R_k | \beta_{rs}, \beta_{cs}, \beta_{rcs}]}{\sum_{t=1}^S \pi_t L[R_t | \beta_{rt}, \beta_{ct}, \beta_{rct}]}, \tag{10}$$

where π_s is the relative size of group s , and

$$\begin{aligned} &L[R_k | \beta_{rs}, \beta_{cs}, \beta_{rcs}] \\ &= \phi^* \left(\frac{\epsilon_{11ks}}{\sigma_s} \right) \phi^* \left(\frac{\epsilon_{12ks}}{\sigma_s} \right) \phi^* \left(\frac{\epsilon_{21ks}}{\sigma_s} \right) \phi^* \left(\frac{\epsilon_{22ks}}{\sigma_s} \right) \end{aligned} \tag{11}$$

is the conditional likelihood of the observed data from subject k given the response function for group s , $\phi^* (\cdot)$ is the standardized normal density function, ϵ_{ijk} is the residual error for the observed value R_{ijk} (i.e., $R_{ijk} - R_{ijs}$), and σ_s^2 is the error variance within group s . In essence, a separate ANOVA is conducted for each group, but the groups themselves are formed such that they maximize the likelihood of observing the data. Thus, the needle-in-a-haystack problem is solved “internally” in the sense that no explicitly measured background variables are used to form groups. The Appendix provides a more detailed discussion of the relationship between latent class and ANOVA models for within-subjects and between-subjects experimental designs.

One of the major difficulties in applying the latent class approach is determining the “correct” number of classes. Typically, this decision is based on information criteria such as the Bayesian Information Criterion (BIC) or the Consistent Akaike Information Criterion (CAIC). In the analyses described next, these traditional criteria suggested at least two classes in all analyses, and we report results for the models chosen by these criteria. The exact number of latent classes is not critical, however. Our primary goal is to test the hypothesis that there is only one class (i.e., homogeneity). We also hope to gain insight into the nature of the unobserved heterogeneity whenever this test of homogeneity fails.

Details about the estimation of the parameters and about available software can be found in Wedel and Kamakura (1997). Our goal here is to illustrate the impact of unobserved heterogeneity on tests of crossover interactions. For this illustration, we estimated a separate latent class model for each of the three product types using the same aesthetic

TABLE 3
AGGREGATE AND CLASS-LEVEL ESTIMATES OF EFFECTS (STANDARD ERROR) FOR VISUAL ATTRACTIVENESS RATINGS FOR THREE SETS OF PRODUCT DESIGNS

Parameter	Stimulus set												
	A. Dressers with trim modifications				B. Telephones with shape modifications				C. TV remote controls with shape modifications				
	Aggregate	Class 1	Class 2	Class 3	Aggregate	Class 1	Class 2	Class 3	Aggregate	Class 1	Class 2	Class 3	Class 4
β_1 (row effect)	-.22* (.09)	-.58* (.23)	-.02 (.11)	-.17 (.18)	1.11*** (.12)	1.65*** (.16)	.11 (.08)	.11 (.08)	-1.05*** (.11)	-1.14*** (.13)	-1.95*** (.08)	-.16 (.14)	.59* (.29)
β_2 (column effect)	-.34*** (.09)	-.64** (.24)	-.22* (.11)	-.28* (.09)	.61*** (.12)	1.20*** (.15)	-.47 (.30)	-.47 (.30)	.11 (.11)	-.90*** (.13)	.79*** (.14)	2.11*** (.09)	-.81** (.27)
β_{12} (interaction effect)	1.93*** (.09)	.38 (.28)	3.30*** (.12)	1.70*** (.09)	.73*** (.12)	.38* (.16)	1.35*** (.31)	1.35*** (.31)	.27* (.11)	-.62*** (.14)	.95*** (.16)	-1.29*** (.27)	1.34*** (.15)
σ (standard deviation)	1.76	2.00	1.21	.90	2.33	1.77	2.39	2.39	2.24	1.29	1.51	1.60	1.62
R^2	.56	.18	.88	.79	.28	.58	.26	.26	.19	.60	.70	.71	.52
π (group size; %)	100.0	23.7	33.8	42.5	100.0	64.6	35.4	35.4	100.0	32.2	39.4	10.8	17.6
Type of interaction	D	N	D	D	S	N	D	D	S	N	S	R-S	D
Preference order:													
Most preferred item	S_{22}	S_{22}	S_{22}	S_{22}	S_{11}	S_{11}	S_{22}	S_{22}	S_{22}	S_{22}	S_{22}	S_{21}	S_{22}
	S_{11}	S_{11}	S_{11}	S_{11}	S_{12}	S_{12}	S_{11}	S_{11}	S_{21}	S_{21}	S_{21}	S_{11}	S_{11}
	S_{12}	S_{12}	S_{12}	S_{12}	S_{22}	S_{21}	S_{12}	S_{12}	S_{11}	S_{12}	S_{11}	S_{12}	S_{12}
Least preferred item	S_{21}	S_{21}	S_{21}	S_{21}	S_{21}	S_{22}	S_{21}	S_{21}	S_{12}	S_{11}	S_{12}	S_{22}	S_{21}

NOTE.—Maximum likelihood estimates of model parameters. D = double, S = single, N = neutral, R-S = reverse single, and R-D = reverse double.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

response data as was used in the nonparametric tests. To anticipate our results, all three latent class analyses confirmed that a majority of subjects exhibited a unity effect (i.e., a statistically significant positive interaction). However, in only one case—dressers—were the overall means (or equivalently the parameters of the one class solution) reasonably interpreted as representative of the population as a whole.

The latent class and aggregate results for the dressers stimulus set are shown in Table 3, part A. In this case, the conclusions drawn from the aggregate analysis are generally confirmed at the latent class level. All three classes show negative row and column effects, indicating that subjects prefer dressers with trim added to the drawers (although the row effect is weak for two of three classes). Classes 2 and 3, representing 76 percent of the sample, both show a significant preference for unity, and this effect is a double crossover (i.e., $\beta_{rc} > \max\{|\beta_r|, |\beta_c|\}$). Also, the most preferred dresser for all three classes was stimulus S_{22} . The three latent classes differ mostly on the magnitude of the effects. All effects have the same sign across classes, and these are the same as for the aggregate analysis.

When applied to the telephone/shape data (illustrated in Fig. 2), the latent class regression model produces a different portrait of aesthetic preferences within groups than does the aggregate analysis (see Table 3, part B). At the aggregate level, subjects appear to dislike changes in the handset or the base ($\beta_r > 0$ and $\beta_c > 0$) and to prefer unity ($\beta_{rc} > 0$). The unity effect is a single crossover and the prototype, S_{11} , has the highest aggregate mean. However, each of the two latent classes exhibits a very different pattern of effects. Members of class 1 (65 percent of subjects) strongly dislike the shape modifications and exhibit a small unity effect that is not a crossover interaction. Their most preferred stimulus is the prototype, S_{11} . Members of class 2 exhibit no significant main effects of modification but strongly prefer unity in these two aspects. The unity effect for this class is a double crossover interaction ($\beta_{rc} = 1.35 > .47 = \max\{|\beta_r|, |\beta_c|\}$). The most preferred stimulus is S_{22} , the least prototypical design.

In this telephones application, it is evident that the significant effects at the aggregate level can be interpreted as a mixture of two segments that exhibit either main effects only or only an interaction. When the data are pooled, it appears as though all effects are significant and that the unity effect is a single crossover interaction. However, under a strict interpretation of the latent class model, none of the subjects exhibits this general pattern and, more specifically, none of the subjects exhibits a single crossover unity effect.

Our latent-class analysis of the TV remote control data produces the most dramatic differences across groups, as shown in Table 3, part C. A standard aggregate analysis would lead to the conclusion that people prefer a rectangular front to a rounded front (i.e., a main effect of row), but don't care about the back end (i.e., no main effect of column) as long as it matches the front (i.e., a significant single

crossover interaction or unity effect). None of the four latent classes exhibits this pattern of results.

Class 2 (39 percent of subjects) is most similar to the aggregate analysis. It differs only in that the main effect of column is significant. The unity effect is a single crossover, and the rank order of preferences is the same as the aggregate model. The other three classes (61 percent of subjects) do not match the aggregate analysis in the pattern of significant effects (or even the directions of the effects), the type of interaction, or the preference order. Thus, the aggregate means are highly misleading as a representation of most subjects. Even the unity effect, which is positive for all classes in the dresser and telephone analyses, reverses in two latent classes (43 percent of subjects). These reversals are consistent with the fact that the nonparametric crossovers test failed only for this data set; moreover, the relative proportion of reverse interactions was high in both the interval-scale and ordinal-scale analyses in Table 2. Finally, the non-significant main effect of column is clearly the result of cancellation across significant class-level effects (see our earlier discussions of Table 1 and Fig. 1).

These three applications of latent class modeling demonstrate that traditional aggregate analyses of variance might not reflect the actual individual behavior under investigation. The first example (dressers) demonstrated that the aggregate analysis might reflect the general nature of the phenomenon under study but not portray the true intensity of the segment-level effects being investigated. The second (telephones) showed that effects that are significant in an aggregate analysis may exist separately, but not in combination, at the segment level. Finally, the application to the TV remote controls data demonstrated that when the underlying classes are very different, effects that are significant, but opposite in direction, at the segment level can cancel in the aggregate analysis. The result is that the aggregate means do not accurately represent any of the underlying segments. Having noted these problems with aggregate analyses, we should add that, in each case, most subjects belonged to a class with a significant positive interaction (i.e., a unity effect). Thus, the main hypothesis of Veryzer and Hutchinson (1998) is supported. However, the overall means were incomplete and, in some cases, misleading as indicators of size and stability of the unity effect. Veryzer and Hutchinson noted the variation in effect size across stimulus sets. These analyses show that similar heterogeneity exists even for responses to the same stimulus sets. Future theories of aesthetic response, therefore, need to account for this type of heterogeneity.

REVERSALS OF RELATIVE CHOICE PROPORTIONS

Simple attraction models of individual choice probabilities and population choice shares predict that the relative choice proportions for two items remain constant across variations in the composition of the choice set (e.g., Luce

1959; McFadden 1974).⁹ This property is often called independence from irrelevant alternatives (abbreviated IIA, Tversky [1972]). In contrast to this prediction, the composition of the set of alternatives offered in choice problems has been shown to result in reversals of choice proportions for alternatives that are common to all choice sets. The most well known reversal effects of this type are similarity effects (Tversky 1972), decoy effects (also called attraction or asymmetric dominance effects; Heath and Chatterjee [1996]; Huber, Payne, and Puto [1982]), and compromise effects (Simonson 1989; Simonson and Tversky 1992; Tversky and Simonson 1993).

Let $R_{xy} = P(x, \{x, y\})/P(y, \{x, y\})$ and $R_{xy|z} = P(x, \{x, y, z\})/P(y, \{x, y, z\})$. Property IIA predicts that $R_{xy} = R_{xy|z}$. When z is very similar to y , it is common to observe a similarity effect, $R_{xy|z} > R_{xy}$; that is, the addition of z to the choice set hurts y more than x . For example, in an election in the United States the entry of a conservative Independent candidate generally steals more votes from the Republican candidate than from the Democratic candidate. Similarity effects have been long interpreted as both population and individual-level phenomena and are consistent with most general models of choice probabilities (Allenby et al. 1998; Chintagunta, Jain, and Vilcassim 1991; Elrod and Keane 1995; Hutchinson 1986; Kamakura and Russell 1989; McFadden 1986; Tversky 1972). In particular, many models predict similarity effects because preferences are heterogeneous even though each individual satisfies IIA (e.g., only Republican supporters switch to the Independent).

In contrast, if z is a decoy in the sense that it is dominated by y but not x , then $R_{xy|z} < R_{xy}$; that is, the addition of z to the choice set helps y and hurts x even though no one chooses z . For example, assume x is a low in both price and quality and y is high in both. If z is equal in quality to y but offered at a slightly higher price, then a decoy effect is likely to be observed. In this case, unobserved heterogeneity (among IIA individuals) can be ruled out as an explanation because decoy effects imply violations of regularity. Regularity requires choice probabilities for every alternative in a choice set to decrease (or remain unchanged) when one or more new alternatives are added to the set. No model based on unobserved heterogeneity can predict violations of regularity unless at least some individuals also violate regularity.¹⁰ Thus, unobserved heterogeneity is clearly ruled out as an explanation of context effects that violate regularity. Because no one chooses z in a decoy effect experiment, the choice proportion for y must increase for the effect to be observed.

Compromise effects are similar to decoy effects except that the third item, z , is not dominated. Typically, it extends the efficient frontier of the binary choice set in a way that makes y seem less extreme (e.g., z is higher than y in both

quality and price) and x is hurt more than y by the addition of z (i.e., $R_{xy|z} < R_{xy}$). Although violations of regularity are sometimes observed in these experiments (i.e., $P(y, \{x, y, z\}) > P(y, \{x, y\})$), the difference is often too small to be statistically reliable, and certain experimental paradigms do not permit a test of regularity (Heath and Chatterjee 1995; Simonson and Tversky 1992). Moreover, the theoretical models developed to explain these effects do not require that regularity be always violated. Thus, it is important to be able to rule out unobserved heterogeneity as an alternative explanation.

One approach is to find some auxiliary property other than regularity that is plausibly assumed and also rules out unobserved heterogeneity as an explanation. For example, Tversky and Simonson (1993) show that if individuals are context-independent value maximizers and the distribution of preference orderings across the population satisfies a property they call the ranking condition, then choice proportion reversals should not be observed. The technical definition of the ranking condition is rather complex; however, loosely speaking it requires that most people have preferences that vary monotonically with each attribute (e.g., from low price and quality to high price and quality). While the a priori plausibility of the ranking condition may be open to debate, Tversky and Simonson (1993) also report that they empirically verified this condition for many of their stimuli in preliminary experiments. Given this verification, their observed reversals imply failures of context independence at the individual level.

This general approach is exemplary and provides one way to rule out unobserved heterogeneity as an alternative explanation. However, the ranking condition is specific to their stimulus design, and such conditions may be difficult to deduce for other choice experiments. Moreover, if the ranking condition fails for some stimuli, it does not follow that unobserved heterogeneity can account for the results. It just means that it cannot be ruled out. We propose a more general and direct approach to the problem.

Unobserved Heterogeneity as an Explanation for Compromise Effects

Within-subjects designs are rare in choice experiments because memory for previous choices can bias subsequent choices (inhibiting reversals). For between-subjects experiments, there is not enough information to reliably estimate likelihoods that individual subjects belong to a given latent class, as we did for the aesthetic response data. One can, however, ask the question of whether some simple latent class model can account for the violations of IIA in the aggregate data without such violations within any class. For the following examples, we estimate the standard latent class model for aggregate contingency tables (e.g., Goodman 1979; Grover and Srinivasan 1987). Given the small number of observations, the relatively large number of model parameters, and our simple goal of finding a plausible alter-

⁹Simple attraction models define the probability of choosing item x from set T as $P(x, T) = u(x)/\sum_{y \in T} u(y)$, where u is a nonnegative, real-valued function that represents the attractiveness of each item.

¹⁰This property is easily shown. If $P_s(x, \{x, y\}) \geq P_s(x, \{x, y, z\})$ for all segments $s = 1$ to S , then $\sum_s P_s(x, \{x, y\})/S \geq \sum_s P_s(x, \{x, y, z\})/S$.

TABLE 4
CHOICE PROPORTIONS FOR PORTABLE GRILLS ILLUSTRATING A COMPROMISE EFFECT

	Choice alternative				
	v 160 sq. in. 4 lbs.	w 220 sq. in. 7 lbs.	x 280 sq. in. 10 lbs.	y 340 sq. in. 13 lbs.	z 400 sq. in. 16 lbs.
Latent class model parameters:					
Segment 1 scale values (size = 51%)	5	38	5	47	5
Segment 2 scale values (size = 49%)	34	3	30	3	30
Choice set 1 {v, w, x}, N = 77 :					
Observed overall	.31	.40	.29		
Predicted overall	.30	.43	.27		
Predicted segment 1	.10	.80	.10		
Predicted segment 2	.50	.05	.45		
Choice set 2 {w, x, y}, N = 70 :					
Observed overall		.27	.44	.29	
Predicted overall		.26	.44	.31	
Predicted segment 1		.43	.05	.52	
Predicted segment 2		.09	.83	.09	
Choice set 3 {x, y, z}, N = 72 :					
Observed overall			.26	.47	.26
Predicted overall			.28	.44	.28
Predicted segment 1			.09	.83	.09
Predicted segment 2			.48	.05	.47

NOTE.—Aggregate choice proportions are from Simonson and Tversky (1992, p. 290). Parameters were estimated by nonlinear least squares. Because the data provide only 6 degrees of freedom, the latent class model was constrained. Only five free parameters were estimated: v_1 , w_1 , v_2 , x_2 , and the size of segment 1. The imposed constraint was $v_1 = x_1 = z_1, w_2 = y_2$, class parameters sum to 100, and segment sizes sum to one. These constraints were based on an initial unconstrained analysis with 9 degrees of freedom.

native explanation, a standard nonlinear least squares method was used.

Simonson and Tversky (1992) report the results of a large number of choice experiments that demonstrate context effects of various types. This provides a rich source of data for illustrative purposes because we can search across the different stimulus sets to find particularly instructive examples. As before, the analyses we report mainly reinforce the authors' conclusions and raise interesting problems for future research on heterogeneity in choice experiments. We do not conclude that the published research was flawed by unobserved heterogeneity.

Consider a typical compromise effect from Simonson and Tversky (1992). In several of their experiments, five two-attribute alternatives were defined such that increases in one attribute were always paired with decreases in the other attribute. Table 4 gives a set of such stimuli. For portable grills, more cooking area and less weight are desirable. These five grills are defined such that an increase of 40 square inches in cooking area is always accompanied by an increase of 3 pounds in weight. Three choice sets were used in the experiment: {v, w, x}, {w, x, y}, and {x, y, z}. Alternative w was chosen more frequently than x for the first set, but x was chosen more frequently than w for the second set. Similarly, x was chosen more frequently than y for second set, but y was chosen more frequently than x for the third set. In each set, the middle alternative (i.e., the compromise position) was the modal choice. This is a violation

of IIA. (Note that a test of regularity is not possible because each set has three alternatives.)

Table 4 also shows how the observed proportions could result from aggregating two segments, each of which satisfy IIA. Here, the latent class solution is implausible given the attribute structure of the stimuli. The estimated utility functions for the two segments consist of alternating high and low values as lightness is traded for cooking area. This can be interpreted as segments with multiple ideal points that are precisely located so that there is always an ideal point from one segment between two adjacent ideal points for the other segment. This allows choice shares to reverse as the choice set is changed because it will always be the case that (1) one segment has a single preferred alternative, (2) the other segment divides its share between the other two alternatives, and (3) these roles alternate as the choice set is moved along the efficient frontier of the attribute space. Clearly, this is a very unlikely preference structure, so unobserved heterogeneity does not appear to be a good alternative explanation for these results.¹¹

In contrast to compromise effects, the following example demonstrates that apparent choice context effects are sometimes plausibly explained by unobserved heterogeneity. Ta-

¹¹Wernerfelt (1995) proposes an explanation that assumes that there are unobserved segments; however, each segment exhibits context effects (i.e., violations of IIA). Thus, our latent class analysis does not provide a test of his explanation.

TABLE 5
CHOICE PROPORTIONS FOR DENTAL INSURANCE
ILLUSTRATING AN INVERSE COMPROMISE EFFECT

	Choice alternative				
	v	w	x	y	z
	50% ^a \$110 ^b	60%	70%	80%	90%
Latent class model parameters:					
Segment 1 scale values (size = 36%)	59	28	9	2	2
Segment 2 scale values (size = 64%)	2	2	9	28	59
Choice set 1 {v, w, x}, N = 81 :					
Observed overall	.32	.20	.48		
Predicted overall	.34	.18	.48		
Predicted segment 1	.62	.29	.09		
Predicted segment 2	.18	.12	.70		
Choice set 2 {w, x, y}, N = 77 :					
Observed overall		.29	.22	.49	
Predicted overall		.29	.24	.48	
Predicted segment 1		.72	.24	.04	
Predicted segment 2		.04	.24	.72	
Choice set 3 {x, y, z}, N = 85 :					
Observed overall			.29	.23	.47
Predicted overall			.31	.23	.46
Predicted segment 1			.70	.12	.18
Predicted segment 2			.09	.29	.62
Choice set 4 {x, y}, N = 71 :					
Observed overall			.49	.51	
Predicted overall			.46	.54	
Predicted segment 1			.85	.15	
Predicted segment 2			.25	.75	

NOTE.—Aggregate choice proportions from Tversky and Simonson (1993, p. 295). Parameters were estimated by nonlinear least squares. The latent class model was constrained. Only five free parameters were estimated: v_1 , w_1 , x_1 , y_1 , and the size of segment 1. The imposed constraint was $z_2 = v_1$, $y_2 = w_1$, $x_2 = x_1$, $w_2 = y_1$ (and without loss of generality $z_1 = v_2 = 1$, and segment sizes sum to one). To aid meaningful comparison, the values in this table were rescaled to sum to 100 for each segment.

^aCoverage.

^bAnnual premium.

ble 5 illustrates an “inverse-compromise effect” in which the relative choice proportions for a given alternative are smaller (rather than larger) when it is in the middle position than when it is at one of the extremes (from app. B of Simonson and Tversky [1992]).¹² Again, a two-segment latent class model captures this effect and fits the data well. Moreover, the pattern of scale values is plausible. To see how such patterns arise, consider linear models of preference. The stimuli in this experiment fall on a straight line in the attribute space. Individuals with linear preference functions will find increasingly expensive alternatives more attractive if the slope of the stimulus line is less than the

ratio of the importance weight for quality to that of price. They will exhibit the opposite pattern if the slope is greater than this ratio. Thus, “straight line” stimulus sets are particularly prone to this type of heterogeneity, and we conclude that unobserved heterogeneity is a plausible explanation for these data. Whether the results are due to context effects or to unobserved heterogeneity can only be determined with additional within-subjects data, which would allow us to distinguish individual differences in preferences from individual-level departures from IIA.

GENERAL DISCUSSION

In empirical consumer research, aggregate data are often used to estimate population parameters, provide statistical tests on those parameters, and test theoretical hypotheses. These aggregate results are often assumed to represent individual consumers well. We have shown that this need not be the case. It is possible for aggregate results to exhibit properties not possessed by any specific individual and for individuals to exhibit properties not evident in aggregate analyses. We have referred to this fact as the none-some-all problem because aggregate results may hold for none, some, or all of the individuals. Even when statistical analyses appropriately reject the (null) hypothesis that none of the individuals possess a certain property, implying only that at least some individuals have it, we are often tempted to extend that inference from “some” to “most” or “all” as we discuss the implications of empirical results.

In this article, we have described methods for addressing the none-some-all problem. These methods are summarized in Figure 3. We propose that experimental tests of individual-level theories should use at least one of these methods to rule out unobserved heterogeneity as an alternative explanation. Although open to debate, we find most a priori assumptions suspect and prefer empirical tests for the presence of unobserved heterogeneity. Moreover, given the prevalence of individual differences in prior behavioral research, we believe a heterogeneous model that is simple at the individual level should be preferred to a homogeneous model that requires a complex account of individual behavior. If the data are consistent with both types of models, then our prior beliefs about the likelihoods of specific patterns of heterogeneity and homogeneity in human populations should drive our posterior beliefs about the theories being tested (Brinberg, Lynch, and Sawyer 1992).

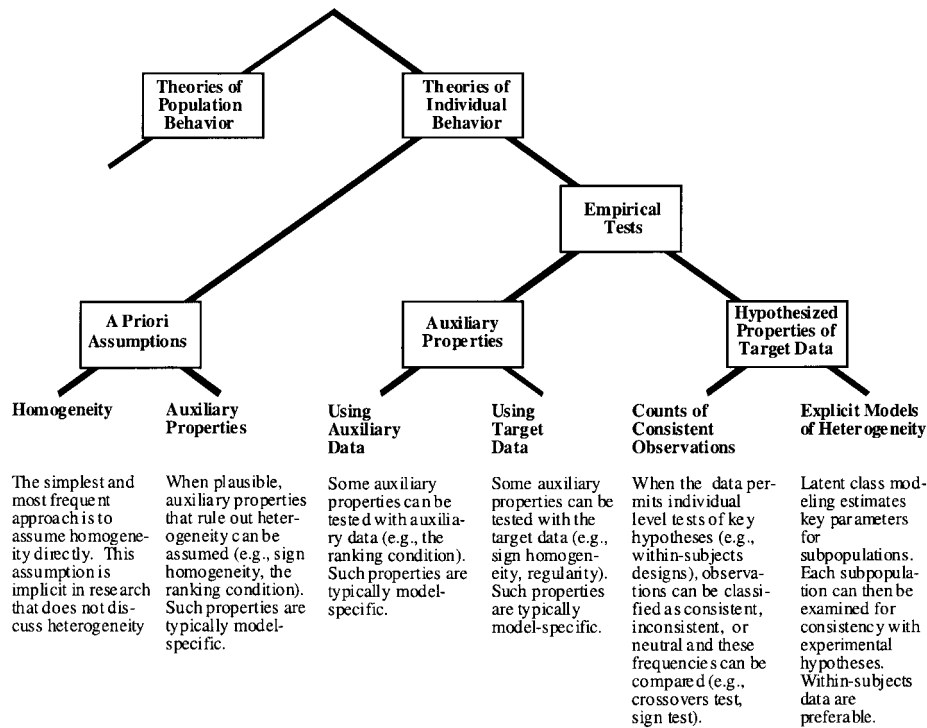
External Validity

At one level, our position simply restates the call of many researchers to evaluate the external validity of their findings—that is, the degree to which their findings are likely to generalize across subpopulations defined by different levels of some background factor or factors (Cook and Campbell 1979; Lynch 1982, 1983). Calder, Phillips, and Tybout (1982) legitimately criticized Lynch’s arguments as a counsel of despair because the number of background factors that could moderate the key findings is indefinitely

¹²In the original article, these data are included with several other sets and titled “Polarization: Additional Examples.” In a personal communication, Itamar Simonson reported that this was an error and these results exemplify neither polarization nor compromise as defined in his papers.

FIGURE 3

METHODS FOR RULING OUT UNOBSERVED HETEROGENEITY AS AN ALTERNATIVE EXPLANATION OF EXPERIMENTAL RESULTS



large—the needle-in-a-haystack problem. Moreover, they made a valid point in arguing that nonsignificant results in a test of a background factor × treatment interaction provide no evidence that other, untested background factors do not in fact interact. This is the problem of induction.

The diagnostic tests used here address both the needle-in-a-haystack problem and the problem of induction by freeing the researcher from the task of specifying which background variables are in fact interacting with the key treatment effects. Latent class analyses, in particular, are extremely helpful. We simply test whether some latent class model exists that leads to a substantively different interpretation than the overall means. If not, the claim of robustness is strengthened, and belief is strengthened that the aggregate results hold at the individual level. On the other hand, if a latent class solution is both plausible and leads to different conclusions than the overall means, the latent classes can be viewed as the levels of a latent qualitative (nominal) factor moderating the relationships observed in the overall means. The researcher now can attempt to find specifiable background variables that explain membership in the different latent classes and can take on the theoretical task of accounting for class differences in patterns of treatment effects.

Discrete versus Continuous Heterogeneity

The latent class model implies a discrete form of consumer heterogeneity. It assumes that consumers can be grouped into classes that are relatively homogeneous in their response to experimental treatments. One might also assume the main effects (β_{rk} , β_{ck}) and interactions (β_{rck}) to be continuously distributed in the population, say with a multivariate normal distribution, leading to a random-coefficient model that can be estimated either by generalized least squares (Hausman 1978; Swamy 1971) or using a hierarchical Bayes formulation (Zeger and Karim 1991). One must be aware, however, that in this case, the variance and covariance of the random coefficients capture unobserved heterogeneity. Interpretation and testing of the effect means may produce misleading conclusions not unlike those drawn from aggregate models. In other words, the mean of the population distribution of the random coefficients might show a crossover interaction, when, depending on the variance of the random coefficients, a significant proportion of the population might exhibit a smaller, noncrossover interaction, no interaction, or an interaction of the opposite sign. Random coefficient models also require the a priori specification of the continuous distribution of the response coefficients (β_{rk} , β_{ck} , and β_{rck}) in the population of consumers.

The latent class model (in fact, a finite mixture of regressions) on the other hand, requires no a priori assumption about the functional form for the distribution of response coefficients. Instead, it requires that the heterogeneity in coefficients can be represented by distinct segments of consumers who are relatively homogeneous in their response to the experimental treatments. Because we mainly want to test homogeneity, and when that test fails provide exploratory information about heterogeneity, we use latent class modeling. For developing theoretical explanations of heterogeneity effects, both types of models should be considered.

Within-Subjects Designs

Because this article is aimed primarily at an audience of experimental consumer researchers who have not given great attention to problems of aggregation biases, we briefly consider the implications of our arguments for experimental design. The main implication is that researchers should make greater use of within-subjects designs than they do currently. (Recall that most choice experiments have been conducted completely between subjects and that only five of 45 recent articles in the *Journal of Consumer Research* and the *Journal of Marketing Research* used within-subjects experimental designs to test predicted interactions.) If designs are exclusively between subjects the researcher's options are limited. One can use aggregate latent class analysis to establish the possibility that the observed results were produced by individuals who do not behave according to the proposed theory. However, with between-subjects data, one cannot conduct nonparametric tests or identify latent class models properly. Thus, it is difficult to provide definitive evidence for or against unobserved heterogeneity as an alternative explanation. Moreover, one cannot then follow up with analyses that would help understand the background factors that explain class membership.

Researchers have often preferred between-subjects designs to within-subjects designs on the grounds that the latter are more reactive (i.e., subject to sensitization and demand characteristics; Greenwald [1976]; Shepanski, Tubbs, and Grimlund [1992]). Moreover, there are various kinds of carryover effects (Keppel 1991) and self-generated validity effects (Feldman and Lynch 1988) that can bias results in within-subjects designs. These are all valid concerns, and the indiscriminate use of within-subjects designs might produce a cure for potential aggregation biases that is worse than the disease.

However, in many cases these traditional problems of within-subjects designs can be avoided. For example, it is sometimes possible to develop experimental procedures that control or eliminate the reactivity and carryover problems (e.g., distractor tasks, delay, etc.). Additionally, one can counterbalance ordinal position with treatment effects in a within-subjects design. This permits statistical tests of position by treatment interactions. Significant results signal the presence of differential carryover. In the extreme, all but the first observation could be deleted from subsequent anal-

yses, reducing the data to a simple between-subjects design. Clearly this comes at a cost in terms of data collection. If carryover effects are absent, all the data can be used, and the diagnostics for unobserved heterogeneity can be applied without fear of countervailing disadvantage of within-subjects designs. If the analyses for order effects revealed main effects of ordinal position of the treatment, those effects should be removed prior to latent class analysis (see Keppel [1991, pp. 361–364] for details).

Conclusion

This article is a cautionary tale. Although we have focused on spurious reversal effects (i.e., crossover interactions and changes in relative choice proportions), the astute reader will have realized that the problem is more pervasive than that. It applies to main effects as well as interactions and to "canceled" real effects as well as apparent effects that are invalid. In its essence our claim is simple. Aggregate means from within-subject or between-subjects designs are not always representative of individual behavior. Few would argue with this. However, most research observes a large number of data points collected in a variety of experimental conditions. Thus, it is very tempting to summarize the results with averages. Our goal here has been to show that the risks are great when averages are taken at face value. When researchers couple within-subjects designs with explicit analyses for heterogeneity, they can make significant progress in understanding the external validity of patterns of treatment means and in refining theories of individual behavior.

APPENDIX

It is useful to compare a latent class approach to modeling heterogeneity to that familiar to researchers working within an ANOVA framework. Consider a 2×2 Row (R) \times Column (C), and contrast four analyses of the same 400 numbers that differ in model assumptions as follows:

1. A completely between-subjects ANOVA, $R \times C \times S/RC$ (i.e., a Row by Column factorial design in which 400 subjects are nested within rows and columns).
2. A completely within-subjects ANOVA, $R \times C \times S$ (i.e., a Row by Column factorial design in which 100 subjects each judged all four combinations of row and column).
3. The same analysis as in analysis 2, but with an explicitly measured blocking variable on Subjects (e.g., gender, with 50 males and 50 females, yielding a $R \times C \times G \times S/G$ ANOVA).
4. A latent class analysis with two classes as described earlier.

Consider first a completely between-subjects design, analysis 1. Variance among the 400 numbers is decomposed into four sources, where r is the number of levels in for the row effect, c is the number of levels in for the column effect, and n is the number of subjects in each between-subjects condition (see Table A1).

Contrast this to the completely within-subjects design of

TABLE A1

ANALYSIS 1: A COMPLETELY BETWEEN-SUBJECTS DESIGN

Source	<i>df</i>	Error term
1. Intercept	1	
2. Rows (<i>R</i>)	$(r - 1) = 1$	<i>S/RC</i>
3. Columns (<i>C</i>)	$(c - 1) = 1$	<i>S/RC</i>
4. <i>R</i> × <i>C</i>	$(r - 1)(c - 1) = 1$	<i>S/RC</i>
5. <i>S/RC</i>	$rc(n - 1) = 396$	
Total	$rcn = 400$	

TABLE A2

ANALYSIS 2: A COMPLETELY WITHIN-SUBJECTS DESIGN

Source	<i>df</i>	Error term
1. Intercept	1	
2. Rows (<i>R</i>)	$(r - 1) = 1$	<i>R</i> × <i>S</i>
3. Columns (<i>C</i>)	$(c - 1) = 1$	<i>C</i> × <i>S</i>
4. <i>R</i> × <i>C</i>	$(r - 1)(c - 1) = 1$	<i>R</i> × <i>C</i> × <i>S</i>
5. Subjects (<i>S</i>)	$(n - 1) = 99$	
6. <i>R</i> × <i>S</i>	$(r - 1)(n - 1) = 99$	
7. <i>C</i> × <i>S</i>	$(c - 1)(n - 1) = 99$	
8. <i>R</i> × <i>C</i> × <i>S</i>	$(r - 1)(c - 1)(n - 1) = 99$	
Total	$rcn = 400$	

analysis 2. A repeated measures ANOVA decomposes the variance into the following sources (see Table A2).

Note that in line 5 of Table A2 we devoted $(n - 1) = 99$ degrees of freedom to model heterogeneity among subjects as a “main effect” (i.e., 99 parameters are computed), in line 6 we devoted $(r - 1)(n - 1) = 99$ degrees of freedom to estimate the mean square error (i.e., heterogeneity among subjects in the Row main effect) for testing the main effect of Row, in line 7 we devoted $(r - 1)(n - 1) = 99$ degrees of freedom to estimate the mean square error (i.e., heterogeneity among subjects in the Column main effect) for testing the main effect of Column, and in line 8 we devoted $(r - 1)(c - 1)(n - 1) = 99$ degrees of freedom to estimate the mean square error (i.e., heterogeneity among subjects in the Row by Column interaction) for testing the Row by Column interaction. These sources of variation were pooled in the between-subjects analysis (i.e., Table A1) as Subjects nested within Row and Column, which was the proper error term for the Row, Column, and Row by Column effects.

One can ask the question of whether the between-subject or within-subject design yields higher power for the Row by Column interaction (or for any specific effect). Although it is often assumed that within-subjects designs are uniformly more powerful than between-subjects designs, it is easy to show that the error term for the interaction in the between-subjects analysis in Table A1, *MS(S/RC)*, is a weighted average of *MS(S)*, *MS(R × S)*, *MS(C × S)*, and *MS(R × C × S)* in the within-subjects analysis (Table A2). It follows that the within-subjects analysis for the interaction will be more powerful only when *MS(R × C × S)* is less than the weighted average of *MS(S)*, *MS(R × S)*, *MS(C × S)*, and *MS(R × S)*. Often this is reasonable, as in when “yea-saying” and “nay-saying” is expected to vary widely across subjects. However, if the effect of subject is expected to be small relative to the interactions of subject with the other effects, then a between-subjects design would be more powerful. Note that in the Veryzer and Hutchinson (1998) study it was natural to assume that subjects might differ in the general preference for products (e.g., telephones vs. dressers, the main effect of subjects) and in their preference for different modifications (e.g., trim vs. no trim, the separate interactions of subject with Row and Column), but they believed that unity would be preferred regardless of these other preferences. Thus, based on their prior expectations, a

within-subjects design would be most powerful for testing the *R* × *C* interaction (i.e., unity), but not necessarily for testing the main effects.

Now consider a modification of the within-subjects design that includes a blocking variable, Gender, to account for some of the heterogeneity (i.e., in Table A3, $g = 2$ and $n = 50$).

Comparing this analysis to the purely within-subjects analysis in Table A2, the sum of squares and degrees of freedom for the Subjects main effect in (2) are decomposed in (3) into Gender and Subjects nested within Gender. *R* × *S* in (2) is decomposed in (3) into *R* × *G* and *R* × *S/G*. Similarly, *C* × *S* in (2) is decomposed in (3) into *C* × *G* and *C* × *S/G*, and *R* × *C* × *S* in (2) is decomposed in (3) into *R* × *C* × *G* and *R* × *C* × *S/G*. Therefore, if there are real Row × Gender, Column × Gender, or Row × Column × Gender interactions, the analysis in Table A3 is superior in two respects: it avoids aggregation biases from pooling across Gender, and it removes these three sources of variance from error terms for Row, Column, and Row × Column effects, respectively. However, the model still assumes implicitly that there are stable individual differences among subjects of the same Gender in how they respond to the Row manipulation, and that this Row × Subjects/Gender interaction is conceptually distinct from individual differences in response to column effects, Column × Subjects/Gender, and from individual differences in sensitivity to the interaction, Row × Column × Subjects/Gender.

In contrast, the latent class model in Table A4 is conceptually similar to estimating the model in Table A1 separately for each of *S* classes which are assumed to be homogeneous. The latent class analysis uses degrees of freedom to estimate the following parameters (for two latent classes in this example).¹³

The latent class analysis estimates a separate set of Row, Column and Row × Column parameters for each class. The

¹³The denotations of *r* and *c* are different in each column. They index parameters in the “Estimated Parameters” column, as in our earlier discussions of latent class, and they denote the number of levels in each factor in the *df* column, consistent with the traditional ANOVA notation used in this Appendix. Lower case *s* indexes classes and upper case *S* is the number of classes.

TABLE A3
ANALYSIS 3: A WITHIN-SUBJECTS DESIGN WITH A BLOCKING VARIABLE

Source	df	Error term
Between:		
1. Intercept	1	
2. Gender (<i>G</i>)	$(g - 1) = 1$	<i>S/G</i>
3. Subjects (<i>S/G</i>)	$g(n - 1) = 2(50 - 1) = 98$	
Within:		
4. Rows (<i>R</i>)	$(r - 1) = 1$	<i>R × S/G</i>
5. <i>R × G</i>	$(r - 1)(g - 1) = 1$	<i>R × S/G</i>
6. Columns (<i>C</i>)	$(c - 1) = 1$	<i>C × S/G</i>
7. <i>C × G</i>	$(c - 1)(g - 1) = 1$	<i>C × S/G</i>
8. <i>R × C</i>	$(r - 1)(c - 1) = 1$	<i>R × C × S/G</i>
9. <i>R × C × G</i>	$(r - 1)(c - 1)(g - 1) = 1$	<i>R × C × S/G</i>
10. <i>R × S/G</i>	$(r - 1)g(n - 1) = 98$	
11. <i>C × S/G</i>	$(i - 1)g(n - 1) = 98$	
12. <i>R × C × S/G</i>	$(r - 1)(c - 1)g(n - 1) = 98$	
Total	$rcgn = 2 × 2 × 2 × 50 = 400$	

error variance is also estimated for each class and used to estimate R^2 for each segment. Statistical tests, however, are based on standard errors estimated by the inverse of the information matrix. In our analyses, the data were zero-centered for each subject (i.e., the main effect of subject was removed and intercepts are therefore zero). The estimation procedure searches for values of these parameters that maximize the likelihood that the 400 data values would be observed. When there is one class, this is equivalent to the analysis in Table A1. Note that the homogeneity assumption renders the analyses in Tables A1 and A2 equivalent because it implies that the expected values of $MS(S)$, $MS(R × S)$, $MS(C × S)$, and $MS(R × C × S)$ used in Table A2 are identical to each other and to any weighted average of these values, including $MS(S/RC)$ which is used in Table A1. When there are two classes, the separate estimation of error variance for each class makes the latent class analysis in Table A4 like the blocking variable analysis in Table A3, insofar as it corrects for aggregation errors from interpreting Row, Column, and Row × Column effects pooled across classes, and in removing the interactions of Class with Row, Column, and Row × Column effects from error terms.

TABLE A4
ANALYSIS 1: A LATENT CLASS MODEL FOR A WITHIN-SUBJECTS DESIGN

Estimated parameters	df
1. Intercept	$S = 2$
2. Row effect (β_{rs})	$S(r - 1) = 2$
3. Column effect (β_{cs})	$S(c - 1) = 2$
4. $R × C(\beta_{rcs})$	$S(r - 1)(c - 1) = 2$
5. Class size (π_s)	$S - 1 = 1$
6. Error variance (σ_s^2)	$S = 2$
7. Residual	$400 - 11 = 389$
Total	400

However, latent class analysis provides a more parsimonious account of the heterogeneity. Any differences among respondents who belonged to the same latent class are assumed to be due to purely random error—not to some residual within-class Row × Subjects, Column × Subjects, or Row × Column × Subjects interactions. In this way, latent class analysis differs even from an ANOVA account in which the latent classes were known, and could be treated as a blocking factor. In addition, the blocking achieved by latent class is more optimal (for the observed data) than could be achieved by some a priori partitioning of subjects because the number of subjects in each class and the assignments of subjects to classes depends directly on the estimated parameters which are maximum likelihood.

We should note that latent class analysis removes one previously discussed disincentive to using within-subjects designs. In standard ANOVA, there are cases when between-subjects designs might be expected to be more powerful. For instance, if the primary interest in the analysis were in the Row × Column interaction effect but the researcher anticipated that $MS(R × C × S)$ might be large relative to $MS(S)$, $MS(R × S)$, and $MS(C × S)$, the intelligent user of ANOVA would normally be discouraged from using within-subjects designs. In principle, the latent class analysis would remove this disincentive in such a case. In latent class, one would find large differences between classes in the Row × Column interaction effects and, ideally, little within-class error, as revealed by a values of σ_s^2 substantially smaller than would have been observed for the error term $MS(S/RC)$ in the between-subjects analysis 1 or the $MS(R × C × S)$ that would be the error term in the within-subjects analysis in Table A2.

[Received January 1998. Revised March 2000. Robert E. Burnkrant served as editor, and Brian Ratchford served as associate editor for this article.]

REFERENCES

- Ailawadi, Kusum, Karen Gedenk, and Scott A. Neslin (1998), "Preference Heterogeneity and Purchase Event Feedback in Choice Models: An Empirical Analysis with Implications for Model Building," Working Paper No. 97-102, Amos Tuck School of Business, Dartmouth College, Hanover, NH 03755.
- Allenby, Greg M., Neeraj Arora, and James L. Ginter (1998), "On the Heterogeneity of Demand," *Journal of Marketing Research*, 35 (August), 384–389.
- Brinberg, David L., John G. Lynch, Jr., and Alan G. Sawyer (1992), "Hypothesized and Confounded Explanations in Theory Tests: A Bayesian Analysis," *Journal of Consumer Research*, 19 (September), 139–154.
- Calder, Bobby J., Lynn W. Phillips, and Alice M. Tybout (1982), "Beyond External Validity," *Journal of Consumer Research*, 9 (December), 240–244.
- Chintagunta, Pradeep K., Dipak Jain, and Naufel J. Vilcassim (1991), "Investigating Heterogeneity in Brand Preferences in Logit Models for Panel Data," *Journal of Marketing Research*, 28 (November), 417–428.
- Cook, Thomas and Donald Campbell (1979), *Quasi-Experimentation: Design and Analysis Issues in a Field Setting*, Chicago: Rand McNally.
- De Sarbo, Wayne and William L. Cron (1988), "A Maximum Likelihood Methodology for Clusterwise Linear Regression," *Journal of Classification*, 5 (2), 249–282.
- Elrod, Terry and Michael P. Keane (1995), "A Factor-Analytic Probit Model for Representing the Market Structure in Panel Data," *Journal of Marketing Research*, 32 (February), 1–16.
- Feldman, Jack M. and John G. Lynch, Jr. (1988), "Self-Generated Validity and Other Effects of Measurement on Belief, Attitude, Intention, and Behavior," *Journal of Applied Psychology*, 73 (August), 421–435.
- Gescheider, George A. (1988), "Psychophysical Scaling," *Annual Review of Psychology*, 39, 169–200.
- Goodman, Leo A. (1979), "On the Estimation of Parameters in Latent Structure Analysis," *Psychometrika*, 44 (March), 123–128.
- Greenwald, Anthony (1976) "Within-Subjects Designs: To Use or Not To Use," *Psychological Bulletin*, 83 (2), 216–229.
- Grover, Rajiv and V. Srinivasan (1987), "A Simultaneous Approach to Market Segmentation and Market Structuring," *Journal of Marketing Research*, 24 (May), 139–153.
- Guilford, J. P. and Benjamin Fruchter (1973), *Fundamental Statistics for Psychology and Education*, New York: McGraw-Hill.
- Hausman, Jerry A. (1978), "Specification Tests in Econometrics," *Econometrica*, 46 (November), 1251–1271.
- Heath, Timothy B. and Subimal Chatterjee (1995), "Asymmetric Decoy Effects on Lower Quality versus Higher Quality Brands: Meta-Analytic and Experimental Evidence," *Journal of Consumer Research*, 22 (December), 268–284.
- Huber, Joel, John W. Payne, and Christopher Puto (1982), "Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis," *Journal of Consumer Research*, 9 (June), 90–98.
- Hutchinson, J. Wesley (1986), "Discrete Attribute Models of Brand Switching," *Marketing Science*, 5 (Fall), 350–371.
- , Wagner Kamakura, and John Lynch (1999), "Unobserved Heterogeneity as an Alternative Explanation for 'Reversal' Effects in Behavioral Research," Working Paper, University of Pennsylvania, Philadelphia, PA 19104.
- Kamakura, Wagner A. and Gary J. Russell (1989), "A Probabilistic Choice Model for Market Segmentation and Elasticity Structure," *Journal of Marketing Research*, 26 (November), 379–390.
- Keppel, Geoffrey (1991), *Design and Analysis: A Researcher's Handbook*, Englewood Cliffs, NJ: Prentice-Hall.
- Krantz, David H. and Amos Tversky (1971), "Conjoint Measurement Analysis of Composition Rules in Psychology," *Psychological Review*, 78 (March), 151–169.
- Luce, R. Duncan (1959), *Individual Choice Behavior*, New York: Wiley.
- Lynch, John G., Jr. (1982), "On the External Validity of Experiments in Consumer Research," *Journal of Consumer Research*, 9 (December), 225–239. (Erratum in March 1983, p. 455.)
- (1983), "The Role of External Validity in Theoretical Research," *Journal of Consumer Research*, 10 (June), 109–111.
- (1985), "Uniqueness Issues in Decompositional Modeling of Multiattribute Overall Evaluations: An Information Integration Perspective," *Journal of Marketing Research*, 22 (February), 1–19.
- Martindale, Colin, Kathleen Moore, and Alan West (1988), "Relationship of Preference Judgments to Typicality, Novelty, and Mere Exposure," *Empirical Studies of the Arts*, 6 (1), 79–96.
- McFadden, Daniel (1974), "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, ed. P. Zarembka, New York: Academic Press, 105–142.
- (1986), "The Choice Theory Approach to Marketing Research," *Marketing Science*, 5 (4), 275–297.
- Meyers-Levy, Joan and Alice M. Tybout (1989), "Schema Congruity as a Basis for Product Evaluation," *Journal of Consumer Research*, 16 (June), 39–54.
- Nedungadi, Prakash and J. Wesley Hutchinson (1985), "The Prototypicality of Brands: Relationships with Brand Awareness, Preference, and Usage," in *Advances in Consumer Research*, Vol. 12, ed. Elizabeth C. Hirschman and Morris Holbrook, Provo, UT: Association for Consumer Research, 498–503.
- Petty, Richard E. and John T. Cacioppo (1979), "Issue Involvement Can Increase or Decrease Persuasion by Enhancing Message-Relevant Cognitive Responses," *Journal of Personality and Social Psychology*, 37 (10), 1915–1926.
- Rossi, Peter and Greg Allenby (1993), "A Bayesian Approach to Estimating Household Parameters," *Journal of Marketing Research*, 30 (May), 171–182.
- Shepanski, A., Richard M. Tubbs, and Richard A. Grimlund (1992), "Issues of Concern Regarding Within- and Between-Subjects Designs in Behavioral Accounting Research," *Journal of Accounting Literature*, 11, 121–150.
- Simonson, Itamar (1989), "Choice Based on Reasons: The Case of Attraction and Compromise Effects," *Journal of Consumer Research*, 16 (September), 158–174.
- and Amos Tversky (1992), "Choice in Context: Tradeoff Contrast and Extremeness Aversion," *Journal of Marketing Research*, 29 (August), 281–295.
- Simpson, E. H. (1951), "The Interpretation of Interaction in Contingency Tables," *Journal of the Royal Statistical Society, B* 54, 145–156.
- Sternthal, Brian, Alice M. Tybout, and Bobby J. Calder (1994), "Experimental Design: Generalization and Theoretical Ex-

- planation," in *Principles of Marketing Research*, ed. Richard P. Bagozzi, Cambridge MA: Blackwell, 195–223.
- Swamy, P. A. V. B. (1971), *Statistical Inference in Random Coefficient Regression Models*, New York: Springer-Verlag.
- Tukey, John (1977), *Exploratory Data Analysis*, New York: Wiley.
- Tversky, Amos (1972), "Elimination by Aspects: A Theory of Choice," *Psychological Review*, 79 (4), 281–299.
- and Itamar Simonson (1993), "Context Dependent Preferences," *Management Science*, 39 (10), 1179–1189.
- Tybout, Alice M. (1995), "The Value of Theory in Consumer Research," in *Advances in Consumer Research*, Vol. 20, ed. Frank R. Kardes and Mita Sujan, Provo, UT: Association of Consumer Research, 1–8.
- Veryzer, Robert W. and J. Wesley Hutchinson (1998), "The Influence of Unity and Prototypicality on Aesthetic Responses to New Product Designs," *Journal of Consumer Research*, 24 (March), 374–394.
- Wedel, Michel and Wagner A. Kamakura (1997), *Market Segmentation: Conceptual and Methodological Foundations*, Boston: Kluwer.
- Wernerfelt, Birger (1995), "A Rational Reconstruction of the Compromise Effect: Using Market Data to Infer Utilities," *Journal of Consumer Research*, 21 (March), 627–633.
- Zeger, Scott L. and M. Rezaul Karim (1991), "Generalized Linear Models with Random Effects: A Gibbs Sampling Approach," *Journal of the American Statistical Association*, (March), 79–86.