GERRIT H. VAN BRUGGEN, GARY L. LILIEN, and MANISH KACKER*

Organizational research frequently involves seeking judgmental response data from informants within organizations. This article discusses why using multiple informants improves the quality of response data and thereby the validity of research findings. The authors show that when there are multiple informants who disagree, responses aggregated with confidence- or competence-based weights outperform those with response data–based weights, which in turn provide significant gains in estimation accuracy over simply averaging informant reports. The proposed methods are effective, inexpensive, and easy to use in organizational marketing research.

Informants in Organizational Marketing Research: Why Use Multiple Informants and How to Aggregate Responses

In many marketing studies, researchers relate independent variables to a dependent variable to investigate causal or associational relationships. In organizational marketing research, the level of theoretical interest is often at the firm or other organizational levels. Observations for unit- or firm-level variables often can be obtained from existing sources such as archives, accounting reports, and so on. If valid empirical information is available, it clearly should be used (Larréché and Moinpour 1983). However, if data are not available from existing sources or are not accessible (Kumar, Stern, and Anderson 1993), as is often the case with historical or confidential data, researchers must obtain proxy, retrospective, judgmental response data from informants.[1]

Researchers collecting information about organizational variables through the responses of informants face two important procedural issues: (1) how to determine the number of informants (i.e., a single or multiple informants) and (2) how to develop a way to aggregate response data if they are collected from multiple informants.[2] Although it is more convenient to rely on a single informant, several researchers have found that a multiple informant-based approach yields response data of far superior quality (Hill 1982; Hogarth 1978; Seidler 1974). Consequently, researchers often recommend relying on multiple informants for the study of both intraorganizational (e.g., Silk and Kalwani 1982) and interorganizational (e.g., John and Reve 1982; Philips 1981) phenomena.

In this article, we address whether using multiple informants should be expected to improve the quality of response data, as well as the understanding of the structural relationships they are used to investigate. Researchers do not always recognize the impact of measurement error on empirical results (Cote and Buckley 1988). Because (measurement) validity is a necessary condition for theory development and testing (Peter 1981), measurement error hampers theory development and testing in organizational marketing. In addressing the multiple informant question, we perform an integrative review of the existing literature on measurement theory (see, e.g., Cote and Buckley 1988) to underscore the need for multiple informants. Next, we develop and test several methods for the aggregation of response data once those data have been obtained from multiple informants, which is the main focus of this research. Our aggregation methods differ from previously developed methods in that they call for less effort from informants and researchers than do behavioral aggregation (e.g., Kumar, Stern, and Anderson

[1]Informants report their perceptions and judgments about particular organizational properties. They differ from respondents, who give information about themselves as individuals (Anderson 1987).

[2]Several labels have been used to refer to this issue, including aggregation, synthesizing, opinion pooling, merging, compromising, and consensus building (Lipscomb, Parmigiani, and Hasselblad 1998). We use the term *aggregation* exclusively here.

*Gerrit H. Van Bruggen is Professor of Marketing, Erasmus University, Rotterdam (e-mail: gbruggen@fac.fbk.eur.nl). Gary L. Lilien is Distinguished Research Professor of Management Science, The Pennsylvania State University (e-mail: g51@psu.edu). Manish Kacker is Assistant Professor of Marketing, Tulane University (e-mail: mkacker@tulane.edu). The authors acknowledge the support of Penn State's Institute for the Study of Business Markets. They thank Arvind Rangaswamy, John Rossiter, and Ale Smidts for their insightful comments on a previous draft of this article. All errors remain the authors' responsibility.

Journal of Marketing Research
Vol. XXXIX (November 2002), 469–478

1993; Libby and Blashfield 1978) or Bayesian (Morris 1977; Winkler 1981) methods, and they are computationally simpler than the latter. Our results show that applying our methods significantly improves the accuracy of organizational response data rather than averaging informants' responses, the most common practice in empirical organizational research.

COLLECTING RESPONSE DATA ON ORGANIZATIONAL VARIABLES: SINGLE OR MULTIPLE INFORMANTS?

Organizational variables such as sales and profits (which we call "pure" objective measures) and variables such as power and dependence (which we call quasi-objective measures, because though they appear to be subjective constructs, they are usually measured objectively using such items as "what percentage of your sales occur through dealer X;" for some recent examples, see Chandy and Tellis [1998]; Homburg and Pflesser [2000]; Kumar, Scheer, and Steenkamp [1998]) are measured at the firm or organizational unit level and have the property that a right or "true" score for the measure exists. The true score of such organizational variables will often differ from their measured value (e.g., in our case, the value of the informant's response) because of the presence of measurement error,

$$\text{Measured value} = \text{true score} + \text{error},$$

where

$$\text{Error} = \text{systematic error} + \text{random error}.$$

Organizational research also involves informants' reports of their idiosyncratic judgments about organizational variables. In such cases, no true score exists, and therefore no form of aggregation is appropriate.

The random error cited may result because people who are asked to assume the role of a (key) informant and make complex judgments find it difficult to make those judgments accurately (Philips 1981). The expected value of the random error can be assumed to equal zero. Systematic error is the degree to which expectations of judgments do not equal the true value (Einhorn, Hogarth, and Klempner 1977). The systematic error in an informant's response can result from both individual sources (i.e., because of the informant's individual characteristics and/or biases) and organizational sources (i.e., because of the informant's hierarchical or functional position within the organization). Furthermore, the methodology that is employed can be a source of systematic error (i.e., common method error). The size of the error component can be substantial; Philips (1981) notes that informant reports often exhibit less than 50% of the variance attributable to the trait factor under investigation, and random error and informant biases account for the rest of the variance. In a review of studies in marketing, psychology, sociology, other business areas, and education, Cote and Buckley (1987) find that, on average, 41.7% of the variance in a measured variable is due to the trait and that systematic and random error account for 26.3% and 32.0% of the variance. For the marketing studies they survey, these figures are 68.4% for the trait, 15.8% for the systematic error, and 15.8% for the random error.

Researchers are interested in the relationships between the true scores of the variables of interest (the traits). However, they only observe relationships between measured values of those variables, which include error. Therefore, empirically assessed relationships between variables depend not only on the correlation between the true scores of these variables but also on the correlation between the systematic errors of the variables of interest, the magnitudes of the systematic errors, and the magnitude of the random errors. Cote and Buckley (1988) find that the stronger the true correlation between constructs, the more the observed or empirical correlation underestimates this true relationship. Conversely, the weaker the true correlation, the more the observed correlation overestimates this true relationship between variables. To ensure that the observed relationship between variables accurately reflects the relationship between the true scores of these variables, both the magnitude of the error components and the levels of the correlations between the systematic error components should be minimized.

The expected value of the correlations between systematic errors increases if the response data for certain variables come from the same source. The source of the systematic error (e.g., organizational perspective, personal characteristics, halo effects) affects the measurements of the different variables in a similar way. A simple way to reduce the effect of correlations between systematic errors is to use multiple sources (i.e., informants) for different variables. If the source of the systematic error is the informant, using different informants for different variables will be appropriate. If the systematic error stems from, say, the informant's organizational position, the informants that provide information about the different variables must vary with respect to this organizational position (i.e., the observations of the different informants should be independent and have only the trait under investigation in common). Using multiple sources will reduce the expected value of the correlation between systematic sources and thereby decrease the difference between the observed and true correlations. Furthermore, selecting the most knowledgeable informant per construct will most likely decrease error, because no single informant is likely the most reliable informant on all issues (Philips 1981), especially in large organizations (Seidler 1974).

If a researcher is interested in identifying the substantive impact of the source of the systematic error, multiple informants are needed to provide response data on the different variables in the research model. These multiple informants should show variation with respect to the presence of the systematic error source. Structural equation modeling can be used to analyze these data and separate the effects of structural factors from, for example, organizational perspectives (Anderson 1985, 1987). The systematic error source then becomes a variable in the research model.

In addition to minimizing the correlations between systematic error components, the systematic error and the measurement error should be minimized because error in a measurement attenuates the observed relationship between variables. Using multiple informants and aggregating their responses into a single composite score helps minimize error. Increasing the number of informants reduces random error through the averaging process, so larger samples will increase reliability. The optimum number of informants depends on the costs of obtaining additional independent judgments and of error in the final group judgment (Ferrell 1985). In a forecasting application, Ashton and Ashton (1985) find that combining between two and five forecasts is

effective, whereas Libby and Blashfield (1978) report that most of the gain from aggregating multiple judges can be obtained with three judges.

When there is systematic error in informants' responses, aggregation by averaging n individual judgments will give a group judgment with a variance smaller than that of the individual estimates, but it will not eliminate systematic error (Ferrell 1985; Rowe 1992). In such circumstances, it is valuable to identify the systematic error sources and find the informant with the smallest error. If the most accurate response can be identified with certainty, that response should be used. If the response accuracy of group members cannot be determined with certainty, a weighted average of the responses from members that assigns higher weights to those more likely to be accurate gives results whose accuracy falls between that of the equally weighted average and the best-member approach. The result will be closer to the best-member approach if informants with more accurate responses can be identified reliably.

In summary, in research contexts such as organizational research, obtaining reports from multiple informants is preferable to a single informant report, because such use reduces the correlation between systematic error components, averages out random error in individual responses, provides the opportunity to analyze the impact of error sources, and provides a method to correct for systematic error in informants' responses. To correct for systematic error, it is important to assess an informant's response accuracy. The question then becomes: How can we determine this response accuracy and use this information in the development of aggregated individual opinions into a group value? We address this key research question in the next section.

AGGREGATING MULTIPLE INFORMANTS' RESPONSES

Because response data collected from multiple informants often reveal a lack of agreement and because informants differ in their response accuracy, we are interested in how to aggregate the responses of the various members of a group into a single group composite value. Both behavioral and mathematical methods for aggregating individual informants' reports have been devised.

In behavioral aggregation, informants discuss the matter, work out their differences, and agree on a (group) value (Ferrell 1985). This approach solves the aggregation problem directly; however, behavioral aggregation requires considerable effort and (potentially impractical) coordination among informants in the collection of the response data. In addition, informant requirements for anonymity and confidentiality may make the approach difficult to apply. Furthermore, the consensus reached may be a poor indicator of perceptual agreement, because group properties and processes, such as power–dependence relations among informants, coalition formation, conformity pressures, and groupthink (Schwab and Heneman 1986), may affect it. Finally, the most influential group members may not be the most accurate ones (Larréché and Moinpour 1983). All of these factors could have a negative impact on the accuracy of the group value that results from the behavioral approach.

The effort and coordination required by the behavioral aggregation approach prompted Kumar, Stern, and Anderson (1993) to propose a hybrid consensus-averaging approach. They average responses to arrive at composite measures when there are only minor differences among informant reports. If there is a major disagreement among knowledgeable informants, they suggest the consensual approach. Kumar, Stern, and Anderson assess the performance of this approach using multiple informant response data (sales managers and fleet managers in a major vehicle rental company) and find significant differences between the initial individual reports of the two informant positions. The subsequent consensual responses were more highly correlated with responses of the hierarchically superior position (sales managers) than with the inferior position (fleet managers). This result supports the contention that the process used to arrive at these consensual responses may reflect underlying power–dependence relations and conformity pressures faced by underlings, which are the reasons Schwab and Heneman (1986) do not favor consensual approaches. An alternative technique, the Delphi method, does not suffer from these problems. However, as with most group decision schemes, it is costly, because it requires multiple informants and multiple, time-consuming iterations (Libby and Blashfield 1978).

Mathematical aggregation can be an attractive alternative to behavioral aggregation (Ferrell 1985). A widely used example of mathematical aggregation is the simple averaging of the judgments of separate informants. However, when informants exhibit substantial disagreement, such aggregation rarely produces the most accurate values (James 1982). The more individual judgments are biased, the less is the improvement in accuracy. In general, the group judgment process is not an averaging process (Sniezek and Henry 1989). Hill's (1982) review shows that group performance is often better than the performance of the average informant; however, group performance is often inferior to the potential suggested in a statistical pooling model.

Bayesian models have been proposed to combine informants' opinions (Morris 1977), especially with respect to probability assessments (Agnew 1985). These models provide a natural and flexible way to incorporate dependencies among informants, while acknowledging that the informants may disagree for a reason (Lipscomb, Parmigiani, and Hasselblad 1998). Although the Bayesian approach is theoretically elegant, it is challenging to apply in practice (Larréché and Moinpour 1983) because it requires the decision maker to assess complicated multivariate likelihood functions (Clemen and Winkler 1993). These methods thus require substantial effort from both the informants and the researcher.

The cost and effort of both behavioral aggregation methods and the more sophisticated Bayesian methods have most likely limited their use. Indeed, the methods sections of most organizational research articles cite time and cost constraints as reasons for choosing a single, key informant. Extant aggregation methods that clear this hurdle (e.g., the simple averaging of response data across informants) often yield inaccurate aggregate estimates. Therefore, we seek methods that are simple and inexpensive to implement (enhancing the likelihood of use) and capable of yielding accurate aggregate estimates (increasing the usefulness of the results).

If informants consistently over- or underestimate variables, aggregation by averaging n individual judgments will yield a group judgment with a variance smaller than that of

the individual estimates but will not eliminate any consistent bias (Ferrell 1985; Rowe 1992). Simple averaging thus is effective only if no systematic error is present in the responses of individual informants, a condition unlikely to hold in most organizational research settings.

Therefore, a need exists for a process beyond unweighted averaging if there are reasons to suspect biases in individual estimates (Sniezek and Henry 1989). As noted previously, we should use the response of the most accurate respondent if that respondent can be identified unambiguously. If not, we should use a weighting scheme, in which higher weights are assigned to the reports of the respondents more likely to be accurate. The question then becomes: How can we determine an informant's accuracy and include this information in aggregation procedures?

An individual informant's accuracy can be identified by using other group members' responses or assessments of personal or others' likely response accuracy. When using other group members' responses to assess a person's accuracy, we can compare the response of the person with the responses of the group as a whole. Using a "majority rules" guideline, we can define an informant's response inaccuracy as its deviation from the group's mean response. The larger the group size, the more accurate the (unweighted) group mean will be, and consequently, the more likely it is that the deviation of a person's response from this group mean will reflect the response (in)accuracy of this person.

Self-assessment of expertise, knowledge, or confidence provides an alternative approach to determine informants' accuracy. If informants are biased about their ability, this approach can lead to over- or underconfidence (Mahajan 1992). However, Rowe (1992) indicates that self-rated confidence may be an appropriate measure of expertise when subjects can actually evaluate their confidence in a specific problem area to which they are regularly exposed. Alternatively, either historical assessments of respondents' response accuracy or assessments of such accuracy in related tasks can be used, options that we investigate here. We present formal operationalizations of these ideas in the following section and apply them to two empirical studies in the subsequent sections.

THREE APPROACHES FOR RESPONSE DATA AGGREGATION

We describe and apply three approaches to aggregate the scores of informants in our empirical studies: (1) an unweighted group mean (our reference value), (2) a value for which weights are derived from the response data (i.e., using group information), and (3) a value for which the weights are derived from self-reported confidence scores. The two weighting procedures include information aimed at identifying and correcting for systematic errors in individual informants' responses in the development of an aggregated group value.

Unweighted Group Mean

Our first (and benchmark) aggregation method entails computing the arithmetic mean of the individual responses of group members. This is the simplest form of the aggregation approach (Kumar, Stern, and Anderson 1993). The value of the unweighted mean for variable X, $UNWMEAN_{xi}$, of group i can be computed as follows:

$$(1) \qquad UNWMEAN_{xi} = \frac{\left(\sum_{j=1}^{n_i} X_{ij}\right)}{n_i},$$

where

X_{ij} = the response for the value of variable X by informant j in group i, and
n_i = number of informants in group i.

Response Data–Based Weighted Mean

The second aggregation method derives from the view that the degree of agreement among informants' responses contains information that should be incorporated into the aggregate measure. For example, when two informants in a three-informant group provide similar responses and the third informant provides a substantially different value, the responses provided by the two agreeing informants might be weighted more heavily than that of the third. This approach assumes that the true value is closer to the responses provided by agreeing informants than to that of the deviating informant(s) and that the response of the deviating informant contains a larger systematic error component. Developing such aggregated values addresses James's (1992) call to demonstrate perceptual agreement among informants before aggregating measurements.

To develop such a response data–based measure, we must compute weights for the responses of each informant. We first compute $DIST_{xij}$, the absolute distance of informant j's response on variable X from the unweighted, arithmetic mean of group i (to which j belongs):

$$(2) \qquad DIST_{xij} = \left|X_{ij} - UNWMEAN_{xi}\right|.$$

The weight assigned to informant j's response should be inversely related to its absolute distance from the unweighted mean for group i, relative to the distances of the other group members, so we compute the weight for informant j's response on variable X ($WEIGHT_{xij}$) as follows:

$$(3) \qquad WEIGHT_{xij} = \left[\frac{\left(\sum_{j=1}^{n_i} DIST_{xij}\right)}{DIST_{xij}}\right]^{\alpha}.$$

In Equation 3, we introduce a parameter α with reference value of 1. When the value of that parameter increases, the weights of observations close to the arithmetic mean increase relative to the weights for observations that are further away; as α approaches 0, the weights will approach those associated with the unweighted mean. Parameter α corrects for the impact of the systematic error in informants' responses. The higher the value of the optimal α, the smaller is the weight attached to responses from informants whose information is expected to contain substantial systematic error (i.e., those that are farther away from the unweighted mean). If the optimal value is 0, informant responses do not contain systematical error.

Finally, we compute the weighted mean $WDMEAN_{xi}$ of variable X for each group i, for which the responses for each

group member are weighted according to their distance from the unweighted group mean:

$$(4)\qquad WDMEAN_{xi} = \sum_{j=1}^{n_i}\left[\frac{WEIGHT_{xij}}{\left(\sum_{j=1}^{n_i} WEIGHT_{xij}\right)} \times X_{ij}\right].$$

Confidence-Based Weighted Mean

Our third aggregation approach uses weights based on informants' self-assessed confidence in the accuracy of each response estimate they give. Here, we weight response estimates provided by more confident informants more heavily than we do those from less confident informants. $WCMEAN_{xi}$, the value of variable X for group i in which informant j's response is weighted by his or her confidence $CONF_{xij}$ in the accuracy of that response, is given as follows:

$$(5)\qquad WCMEAN_{xi} = \sum_{j=1}^{n_i}\left[\frac{CONF_{xij}^{\alpha}}{\left(\sum_{j=1}^{n_i} CONF_{xij}^{\alpha}\right)} \times X_{ij}\right].$$

Again, we introduce a parameter α (with a reference value of 1) that makes it possible to manipulate the weight assigned to responses from more confident informants (i.e., those that are expected to show smaller systematic errors). As previously, when α approaches 0, the response estimate reduces to the arithmetic mean. Although there are many other possible approaches, these three models represent a range of possible aggregation procedures.

STUDY 1: AGGREGATING RECALL RESPONSE DATA

Methodology

To assess the accuracy of these aggregation methods and measure the benefits of having different numbers of informants, we needed to collect response data in a realistic organizational setting in which we could compare informants' response estimates with objective, true values. We used the environment of MARKSTRAT (Larréché and Gatignon 1990), a computer-based, marketing strategy simulation that has been widely used by researchers to study decision making (Glazer, Steckel, and Winer 1992). In the simulation, groups of participants play over several periods and make strategic and tactical marketing decisions for different, competing firms.

The informants in this study were 67 marketing students participating in a capstone marketing strategy course at a large Midwestern U.S. university. The students formed 20 groups of 2, 3, and 4 people to make decisions for one of five companies operating in one of four MARKSTRAT industries. The students made each decision after analyzing results from previous periods and reviewing market research studies. All groups had the same amount of time to make decisions, and all groups made decisions simultaneously.

After the groups played the game for a few periods, we asked each informant to complete a questionnaire individually. Among other questions, we asked them to recall the values of eight variables (the levels of marketing mix variables, such as advertising, price, and sales effort) for decisions they had just made, as well as the size of the marketing budget they had available to spend for the next set of decisions. We also asked them to record their confidence in their responses to these questions on a nine-point scale, where 1 indicated "not certain at all" and 9 reflected "completely certain" about the accuracy of their response. All informants filled out the questionnaire on paper at the same time during a classroom meeting, so there was no variation, across informants, in the amount of time between submitting the MARKSTRAT decisions and filling out the questionnaire. Simultaneity of response is important because variation in the time interval could influence the relative accuracy of the informants' responses. To ensure involvement, stimulate accuracy, and discourage cooperation among group members, we awarded prizes to participants who provided the most accurate response estimates.

Two elements of the research context deserve mention. First, all group members were students who were not assigned any specific hierarchical positions or functional responsibilities. Consequently, they shared the same (homogeneous) viewpoint, with limited scope for hierarchical or functional bias (i.e., systematic error stemming from organizational sources). Second, the MARKSTRAT program provided the actual values of all variables, so we were able to assess the accuracy of the informants' reports explicitly.

Table 1
MAPE AND AIC OF THREE AGGREGATION PROCEDURES
(STUDY 1: RECALL RESPONSE DATA)

Aggregation Procedure	*MAPE*	*AIC*
Unweighted group mean	16.30 (10.08)	5.89
Response data–based weighted mean	14.20 (10.22)	5.71
Confidence-based weighted mean	12.81 (8.37)	5.44

Notes: Standard deviations are in parentheses.

Empirical Results

We applied the aggregation procedures described in the previous section to the response data provided by our experimental groups. We use the mean absolute percentage error (MAPE), a dimensionless metric, as our index of relative performance. The MAPE of group i for a specific aggregation approach (averaged over the eight MARKSTRAT variables, k = 1, ..., 8) is computed as follows:

$$(6)\quad MAPE_i = \left(\sum_{k=1}^{8}\left|\frac{\text{Estimated value of } X_{ki} - \text{real value of } X_{ki}}{\text{Real value of } X_{ki}}\right| \times 100\%\right)/8.$$

In Table 1, we present the MAPE (along with Akaike's information criterion [AIC]) for the three aggregation approaches. Following Greene (1997), AIC is computed as follows:

$$(7)\qquad AIC_m = \ln\left(\frac{e_m' e_m}{n}\right) + \frac{2K_m}{n}.$$

In Equation 7, e_m is the error vector of aggregation approach m, K_m is the number of fitted parameters using

Table 2
MAPE AND AIC OF THE WEIGHTED AGGREGATION PROCEDURES FOR DIFFERENT VALUES OF α
(STUDY 1: RECALL RESPONSE DATA)

Aggregation Procedure	$\alpha = 1$	*Uniform α (optimized across MARKSTRAT variables)*	*Variable Specific α (optimized per MARKSTRAT variable)*
Response data–based weighted mean	14.20 (10.22)	12.45 (10.76) (α = 25.70)	12.35 (10.71) (α range = 2.88–77.00)
AIC	5.71	5.68	6.37
Confidence-based weighted mean			
Using item-specific confidence scores	12.81 (8.37)	7.90 (6.24) (α = 12.87)	7.53 (6.07) (α range = 5.46–25.80)
AIC	5.44	4.70	5.32
Using single, average confidence scores	14.69 (8.77)	8.79 (7.16) (α = 13.18)	8.34 (6.79) (α range = 5.05–165)
AIC	5.67	4.94	5.53

Notes: Standard deviations are in parentheses.

aggregation approach m, and n is the number of observations.

The results in Table 1 show that weighting improves accuracy ($F = 15.45$, $p = .001$) and thus decreases MAPE and improves the value of AIC. Furthermore, confidence-based weighting performs better than does response data–based weighting; compared with the unweighted mean, the confidence-based weighted mean improves accuracy by more than 20% ($F = 20.88$, $p < .001$). For the response data–based and confidence-based weighting procedures results in Table 1, we used the reference α value of 1. Next, we investigated whether accuracy could be improved by allowing the value of α to differ from its reference value.

For the response data–based weighting procedure, we calculated the value of α in Equation 2 that minimized MAPE (in Equation 7) using the Solver module in Microsoft Excel. We computed an optimal value of α for each of the eight MARKSTRAT variables (variable-specific α). We also computed a single, optimal α that was restricted to the same value for all eight MARKSTRAT variables (uniform α). The results in Table 2 show that the accuracy of the response data–based weighting procedure can be improved by approximately 15% through this procedure ($F = 9.110$, $p = .007$). The optimal uniform value of α was 25.70, whereas optimal variable-specific values of α ranged from 2.88 to 77.00. The difference between the MAPE of the uniform α approach and the MAPE of the variable-specific α approach is quite small (12.45/14.20 = 87.7% for a 12.3% improvement in MAPE versus 12.35/14.20, which yields a 13.0% improvement). Apparently, most of the gain in MAPE derives from weighting the agreeing informants most heavily (α substantially greater than 1), though MAPE is relatively insensitive to the actual value of the higher weight. The AIC values show that if the fit improvement for the loss of degrees of freedom is discounted, the approach using the uniform α performs better than does the approach using the variable specific α; that is, it does not provide a sufficient (statistical) return on the investment needed to extract individual α values.

We optimized α in the confidence-based weighting approach in a similar manner. We calculated the value of α in Equation 5 that minimized the MAPE (in Equation 6) and computed both variable-specific values of α and a uniform, optimal α. Again, we found that increasing the weights of the more confident informants leads to substantially more accurate aggregation results: The MAPE for the unweighted group mean is 16.30, setting α equal to 1 gives a MAPE of 12.81, a single optimal α brings the MAPE down to 7.90, and using item-specific values for α yields a MAPE of 7.53. Thus, with confidence-based weights, MAPE can be reduced from 16.30 (for the unweighted group mean) to 7.53, yielding a reduction of more than 50% ($F = 19.876$, $p < .001$). As with the response data–based weights, most of this gain comes from increasing the value of α to well above 1, with only incremental improvements arising from item-specific modifications. Again, the values of AIC show that making item-specific α adjustments does not pay off if fit improvement is discounted for the loss of degrees of freedom.

On the basis of the results of Tables 1 and 2, we conclude that (1) applying a weighting procedure leads to considerably more accurate aggregation results than does using the arithmetic mean and (2) confidence-based weights perform better than response data–based weights. To determine how robust the latter finding is, we investigated how well the confidence-based approach would perform if we used a single (overall) confidence value as opposed to the item-specific values. Table 2 also gives these results, for which we used the informants' overall confidence score, obtained by averaging specific confidence levels indicated for each variable. We find that MAPE increases by only 11% for both the uniform α (from 7.90 to 8.79) and variable-specific values of α (from 7.53 to 8.34). Thus, for judgments on many items, our results suggest it is reasonable to seek only a single overall confidence judgment; the loss in accuracy is minimal, and the reduced cognitive burden on informants is likely to enhance the quality and quantity of responses.

Our results were similar when we used other measures of central tendency (median) or alternative functional forms involving more than one parameter for calculating weights for the response data–based and confidence-based aggregation approaches.

STUDY 2: AGGREGATING FORECASTING RESPONSE DATA

The relatively strong performance of the confidence-based weighting procedure may seem surprising because these types of self-assessments have not always been found to be accurate (Larréché and Moinpour 1983). The relatively simple character of the task (i.e., straight recall) used in our

Table 3
MAPE AND AIC OF AGGREGATION PROCEDURES APPLIED TO FORECASTING RESPONSE DATA (STUDY 2)

Aggregation Procedure	*MAPE*	*AIC*
1. Unweighted group mean	26.88 (32.17)	7.43
2. Response data–based weighting		
$\alpha = 1$	25.57 (32.29)	7.39
Uniform optimized α (= 5.47)	23.82 (32.00)	7.48
Variable specific optimized α (ranges from 5.47 to 9.76)	23.80 (31.90)	7.78
3. Confidence-based weighted mean		
(a) Variable specific confidence		
$\alpha = 1$	22.96 (29.30)	7.18
Uniform optimized α (= 53.34)	16.40 (17.63)	6.47
Variable specific optimized α (ranges from 4.08 to 109.40)	16.21 (17.66)	6.77
(b) Average confidence		
$\alpha = 1$	21.59 (24.40)	6.92
Uniform optimized α (= 20.91)	15.75 (16.27)	6.35
Variable specific optimized α (ranges from 1.83 to 73.83)	15.60 (16.39)	6.66
(c) Overall confidence		
$\alpha = 1$	23.95 (29.07)	7.21
Uniform optimized α (= 6.04)	19.19 (16.44)	6.58
Variable specific optimized α (ranges from 3.93 to 7.86)	18.48 (17.23)	6.88
4. Competence-based weighted mean		
(a) Recall competence		
$\alpha = 1$	22.81 (19.08)	6.75
Uniform optimized α (= 2.82)	20.77 (18.26)	6.76
Variable specific optimized α (ranges from 1.99 to 12.40)	20.41 (18.38)	7.05
(b) Forecasting competence		
$\alpha = 1$	19.02 (16.38)	6.41
Uniform optimized α (= 5.64)	15.34 (16.80)	6.36
Variable specific optimized α (ranges from 3.77 to 7.57)	15.11 (16.93)	6.66

Notes: Standard deviations are in parentheses.

study could help explain our findings in the context of the extant literature. In a second study, we investigate the performance of the various aggregation methods for a more complex task: forecasting. In such a setting, the self-assessed, confidence-based weight might be expected to perform worse than it would for a simpler recall task.

The informants in Study 2 were 39 marketing students who formed 13 groups in a MARKSTRAT exercise, as in Study 1, and they generally followed Study 1's procedures. We asked informants to complete a questionnaire individually after their group had played for a few periods. This time, informants were asked for forecasts of the values of three variables: the two brand awareness levels of the two brands they were responsible for managing and the marketing budget they expected to have available for the next period. This budget was a function of the profit their company would generate on the basis of the decisions they were making. We asked them to record their confidence (in their responses to these questions) on a nine-point scale, where 1 indicated "not certain at all" and 9 reflected "completely certain" about the accuracy of the estimate. As in the previous study, we awarded prizes to informants who provided the most accurate values. As in Study 1, the MARKSTRAT program provided the actual values of the three forecasted variables, so we were able to assess the accuracy of the informant reports explicitly. Again, the time between submitting the MARKSTRAT decisions and filling out our questionnaire was the same for all informants.

Empirical Results

In Table 3, we present the results of applying the aggregation procedures to the forecasting data. Overall, these results are consistent with those of Study 1—weighting improves accuracy, especially for confidence-based weights. For both the variable-specific confidence scores and the average confidence scores (averaged for the two brand awareness forecasts and the budget forecast; sections 3a and b of Table 3), the values of the optimal α are higher in Study 2 than in Study 1. Because a higher value of the optimal α means that the opinions of more confident informants are weighted more heavily, we conclude that for the more complex forecasting task, the self-stated confidence scores are more informative than for the recall task.

In addition to replicating the analyses from Study 1, we also applied three other types of weights:

1. A single overall confidence score that expressed the informants' stated global confidence in all the forecasts they provided (Table 3, section 3c);
2. A recall competence score that reflected the informants' accuracy in recalling variables from the previous MARKSTRAT period (i.e., the same eight variables used in Study 1; Table 3, section 4a); and
3. A forecasting competence score that reflected the informants' accuracy in providing forecasts on two other MARKSTRAT variables (i.e., the sales for Brand 1 and Brand 2; Table 3, section 4b).

To calculate the competency-based weights, we use the formulation in Equations 2–4, with "actual value" replacing UNMEAN in those equations.

The results in Table 3 show that using the overall confidence score as a weight produces less accurate aggregated values than does using variable-specific confidence scores. Considering the relatively low optimal α for this type of weight (compared with the average or individual item

confidence-based weights), we conclude that the overall confidence score is less informative than other confidence scores. Evidently, informants knew that they were less accurate on some variables than on others and expressed this knowledge in the item-specific confidence scores.

We find that using the forecasting competency scores as weights leads to results that are as accurate as using the confidence scores ($F = .05$, $p = .83$), whereas using recall competency leads to less accurate results, though this difference is not significant. The lack of significance is probably due to the small number of observations in Study 2. Apparently, performance on a specific task (i.e., forecasting) is a good predictor of accuracy on a similar task (i.e., forecasting other variables), whereas accuracy scores on a different task (i.e., recalling variables) produce less useful information.

Overall, Study 2 shows that confidence-based weighting improves the accuracy of aggregated variables. Competence-based weighting can perform as well as confidence-based weights if that competency is measured on a task similar to the one under study.

DISCUSSION

When there is error in informants' responses, using multiple versus a single informant improves the quality of response data and thereby the validity of reported relationships in organizational marketing research. Our focus in this research is how responses of multiple informants should be aggregated. Drawing on analyses of response data collected from informants in the MARKSTRAT simulation, we show that aggregating multiple informants' responses significantly enhances the quality of objective, recall, and forecast response data. The quality of our results varies depending on the aggregation method. In our research setting, aggregation through the computation of a simple unweighted mean added accuracy to individual response data through a reduction in the random error component of the individual-level response data. Unweighted aggregation improves accuracy by averaging individual-level errors and biases that are random (Rousseau 1985). However, arithmetic means of informant reports were far less accurate than were techniques that weighted informant reports using self-assessed confidence, measured competence, or response data–based distance weights. The latter methods are more effective in incorporating information about systematic errors in informant responses.

Although some previous research has suggested that individual response weighting is not needed (e.g., Armstrong 1986; Einhorn and Hogarth 1975), our results indicate that individual weighting is effective in enhancing the quality of recall or forecast measures based on multiple informant reports. These differences between results may be due to the different types of weights in the studies. In our approach, we weighted reports provided by more confident and competent informants more heavily than we did those from less confident and less competent informants. In using competence-based weights, the calibration of weights should be based on similar tasks, because the ability to differentiate the expertise of informants is likely to be task specific (Ashton and Ashton 1985). Furthermore, though informants may have difficulty assessing their own expertise and though these assessments may be systematically biased (Alba and Hutchinson 2000), they still prove useful as weights. Our methods are both simple and quite effective. In that sense, they confirm Clemen's (1989) finding that simple combination methods often work reasonably well compared with more complex combinations.

Our recommendations for the use of confidence-based weights draw from the work of Alba and Hutchinson (2000), who show that when informants are very confident (which might be the case for more concrete and objective items), they are likely to be overconfident, whereas when they rely on intuition or think they are guessing, they are actually more accurate than they realize. Alba and Hutchinson also show that overconfidence increases when informants have more expertise or have performed exhaustive analyses. A degree of systematic bias will almost always be the norm, leading to either over- or underconfidence. However, as long as some relationship between accuracy and confidence exists, the relationship can be exploited to provide information for the design of effective aggregation procedures. Therefore, using confidence weights to aggregate reports for a "simple" recall task does not necessarily lead to better results than those obtained when using a similar aggregation approach for a more subjective judgment task. Whereas the second task may suffer from underconfidence, the first may suffer from overconfidence; both will lead to some miscalibration and could have a similar negative influence on the effectiveness of the weighting procedure. As long as this tendency toward over- or underconfidence is likely to be similar across informants and the relative weights reflect the relative accuracy of the informants, our procedures will be effective.

To ensure that reports are from key informants, many researchers include questions in their survey instruments that assess informants' competency. The most effective technique for doing so entails using specific measures that assess the informant's knowledge of each major issue in the study (Kumar, Stern, and Anderson 1993). Our results show how using such competency measures as weights for informant reports can yield composite measures that are of superior quality to unweighted group means.

Our proposed weighting schemes are relatively simple to apply, especially compared with more advanced hierarchical modeling (e.g., Lipscomb, Parmigiani, and Hasselblad 1998) or behavioral aggregation approaches (Ferrell 1985). To apply our measures, researchers need only to (1) obtain measures of reporting confidence for each group of responses and/or (2) include one or more questions for some responses in the study that are closely related to those in question for which the answer is known (to assess informants' competence). Using the latter information, researchers can compute the optimal value of α, which can subsequently be used to develop aggregated values for those variables for which the objective true value is not known to the researchers. Our aggregation methods provide more accurate estimates than do prevailing methods (simple mean), even when α is removed (i.e., $\alpha = 1$). The inclusion of α (and the subsequent calculation of an optimal α) only magnifies and enhances the improvements in accuracy that result from the use of our proposed aggregation methods. Thus, researchers who are unable to identify pure objective variables (with known true scores) for the calculation of the optimal α still can benefit from using our aggregation methods.

As was noted previously, our results are directly applicable for the major set of organizational research studies

involving either objective or quasi-objective variables. Although the approach is not appropriate for truly subjective research in this domain, we argue that no form of aggregation is appropriate there, as no true value of the focal variable exists.

Our study has several limitations. Our response data were collected in a simulated setting. As with all such research, replications in other settings, in both the laboratory and the field, will be needed to understand the realm of applicability of our findings better. It might be that only our procedure (and not our specific empirical findings) has more general applicability. All our informants reported on the same set of variables and were homogeneous, in that there were no sources of functional or hierarchical bias. Although this design improves the internal validity of our empirical analysis, further research should investigate how these results generalize. In addition, our informants provided retrospective reports on observable phenomena, as well as forecasts of these types of variables. Both tasks are much easier than making complex social judgments (John and Reve 1982; Philips 1981). It is likely that the nature and magnitude of the perceptual agreement problem would become even more significant in settings in which informants are required to provide complex, subjective responses. However, our approach should be appropriate as long as a true value of the variable exists.

Overall, we are encouraged to have developed what appears to be both an easy-to-apply and relatively robust response data aggregation procedure. This procedure justifies the collection of response data from multiple informants in an organizational unit.

REFERENCES

Agnew, Garson E. (1985), "Multiple Probability Assessments by Dependent Experts," *Journal of the American Statistical Association*, 80 (390), 343–47.

Alba, Joseph W. and J. Wesley Hutchinson (2000), "Knowledge Calibration: What Consumers Know and What They Think They Know," *Journal of Consumer Research*, 27 (September), 123–56.

Anderson, James C. (1985), "A Measurement Model to Assess Measure-Specific Factors in Multiple-Informant Research," *Journal of Marketing Research*, 22 (February), 86–92.

——— (1987), "An Approach for Confirmatory Measurement and Structural Equation Modeling of Organizational Properties," *Management Science*, 33 (April), 525–41.

Armstrong, J. Scott, (1986), "The Ombudsman: Research on Forecasting: A Quarter-Century Review, 1960–1984," *Interfaces*, 16 (1), 89–109.

Ashton, Alison Hubbard and Robert H. Ashton (1985), "Aggregating Subjective Forecasts: Some Empirical Results," *Management Science*, 31 (12), 1499–1508.

Chandy, Rajesh K. and Gerard J. Tellis (1998), "Organizing for Radical Product Innovation: The Overlooked Role of Willingness to Cannibalize," *Journal of Marketing Research*, 35 (November), 474–87.

Clemen, Robert T. (1989), "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting*, 5 (4), 559–83.

——— and Robert L. Winkler (1993), "Aggregating Point Estimates: A Flexible Modeling Approach," *Management Science*, 39 (4), 501–15.

Cote, Joseph A. and M. Ronald Buckley (1987), "Estimating Trait, Method, and Error Variance: Generalizing Across 70 Construct Validation Studies," *Journal of Marketing Research*, 24 (August), 315–18.

——— and ——— (1988), "Measurement Error and Theory Testing in Consumer Research: An Illustration of the Importance of Construct Validation," *Journal of Consumer Research*, 14 (March), 579–82.

Einhorn, Hillel J. and Robin M. Hogarth (1975), "Unit Weighting Schemes for Decision Making," *Organizational Behavior and Human Performance*, 13 (2), 171–92.

———, ———, and Eric Klempner (1977), "Quality of Group Judgment," *Psychological Bulletin*, 84 (1), 158–72.

Ferrell, William R. (1985), "Combining Individual Judgments," in *Behavioural Decision Making*, G. Wright, ed. New York: Plenum Press, 111–45.

Glazer, Rashi, Joel Steckel, and Russell Winer (1992), "Locally Rational Decision Making: The Distracting Effect of Information on Managerial Performance," *Management Science*, 38 (2), 212–26.

Greene, William H. (1997), *Econometric Analysis*, 3d ed. Upper Saddle River, NJ: Prentice-Hall.

Hill, Gayle W. (1982), "Group Versus Individual Performance: Are N+1 Heads Better Than One?" *Psychological Bulletin*, 91 (3), 517–39.

Hogarth, Robin M. (1978), "A Note on Aggregating Opinions," *Organizational Behavior and Human Performance*, 21 (1), 40–46.

Homburg, Christian and Christian Pflesser (2000), "A Multiple-Layer Model of Market-Oriented Organizational Culture: Measurement Issues and Performance Outcomes," *Journal of Marketing Research*, 37 (November), 449–62.

James, Lawrence R. (1982), "Aggregation Bias in Estimates of Perceptual Agreement," *Journal of Applied Psychology*, 67 (2), 219–29.

John, George and Torger Reve (1982), "The Reliability and Validity of Key Informant Data from Dyadic Relationships in Marketing Channels," *Journal of Marketing Research*, 19 (November), 517–24.

Kumar, Nirmalya, Lisa L. Scheer, and Jan-Benedict E.M. Steenkamp (1998), "Interdependence, Punitive Capability, and the Reciprocation of Punitive Actions in Channel Relationships," *Journal of Marketing Research*, 35 (May), 225–35.

———, Louis W. Stern, and James C. Anderson (1993), "Conducting Interorganizational Research Using Key Informants," *Academy of Management Journal*, 36 (6), 1633–51.

Larréché, Jean Claude and Hubert Gatignon (1990), *MARKSTRAT2: A Marketing Simulation Game*. Palo Alto, CA: The Scientific Press.

——— and Reza Moinpour (1983), "Managerial Judgment in Marketing: The Concept of Expertise," *Journal of Marketing Research*, 20 (February), 110–21.

Libby, Robert and Roger K. Blashfield (1978), "Performance of a Composite as a Function of the Number of Judges," *Organizational Behavior and Human Performance*, 21 (2), 121–29.

Lipscomb, Joseph, Giovanni Parmigiani, and Vic Hasselblad (1998), "Combining Expert Judgment by Hierarchical Modeling: An Application to Physician Staffing," *Management Science*, 44 (February), 149–61.

Mahajan, Jayashree (1992), "The Overconfidence Effect in Marketing Management Predictions," *Journal of Marketing Research*, 29 (August), 329–42.

Morris, P.A. (1977), "Combining Expert Judgments: A Bayesian Approach," *Management Science*, 23 (March), 679–93.

Peter, J. Paul (1981), "Construct Validity: A Review of Basic Issues and Marketing Practice," *Journal of Marketing Research*, 18 (February), 1–10.

Philips, Lynn W. (1981), "Assessing Measurement Error in Key Informant Reports: A Methodological Note on Organizational Analysis in Marketing," *Journal of Marketing Research*, 18 (November), 395–415.

Rousseau, Denise M. (1985), "Issues of Level in Organizational Research: Multi-Level and Cross-Level Perspectives," in

Research in Organizational Behavior, Vol. 7, L.L. Cummings and B. Staw, eds. Greenwich, CT: JAI Press,1–37.

Rowe, Gene (1992), "Perspectives on Expertise in the Aggregation of Judgments," in *Expertise and Decision Support*, George Wright and Fergus Bolger, eds. New York: Plenum Press, 155–80.

Schwab, Donald P. and Herbert G. Heneman (1986), "Assessment of a Consensus-Based Multiple Information Source Job Evaluation System," *Journal of Applied Psychology*, 71 (2), 354–56.

Seidler, John (1974), "On Using Informants: A Technique for Collecting Quantitative Data and Controlling for Measurement Error in Organization Analysis," *American Sociological Review*, 39 (December), 816–31.

Silk, Alvin J. and Manohar U. Kalwani (1982), "Measuring Influence in Organizational Purchase Decisions," *Journal of Marketing Research*, 19 (May), 165–81.

Sniezek, Janet A. and Rebecca A. Henry (1989), "Accuracy and Confidence in Group Judgment," *Organizational Behavior and Human Decision Processes*, 43 (1), 1–28.

Winkler, Robert L. (1981), "Combining Probability Distributions from Dependent Information Sources," *Management Science*, 27 (April), 479–88.