MULTIPLE RATERS IN SURVEY-BASED OPERATIONS MANAGEMENT RESEARCH: A REVIEW AND TUTORIAL

Kenneth K. Boyer

Department of Marketing and Supply Chain Management, *Eli Broad Graduate School of Management*, *Michigan State University*, *East Lansing*, *Michigan* 48824

Rohit Verma

Department of Management, *Kellstadt Graduate School of Business*, *DePaul University*, *Chicago*, *Illinois* 60604-2287, *USA*

Research in the area of operations strategy has made significant progress during the past decade in terms of quantity of articles published, as well as the quality of these articles. Recent studies have examined the published literature base and determined that, in general, the field has progressed beyond an exploratory stage to a point where there is a core set of basic terminology and models. Concurrent with the formation and solidification of a core terminology, there is an increasing emphasis on developing and employing a set of reliable, valid, and reproducible methods for conducting research on operations strategy.

We provide a review of common methods for assessing the degree of reliability and agreement of the responses provided by multiple raters within a given organization to a set of qualitative questions. In particular, we examine four methods of determining whether there is evidence of disagreement or bias between multiple raters within a single organization in a mail survey.

Introduction

Research in the area of operations strategy has made significant progress during the past decade in terms of the quantity and quality of articles published. In response to calls for more empirical research (Adam and Swamidass 1989; Flynn et al. 1990; Leong, Snyder, and Ward 1990; Swamidass 1991), authors have responded with a variety of approaches and subject areas intended to build new theory and test existing theories. Several recent studies have examined the published literature base and determined that, in general, the field has progressed beyond an exploratory stage to a point where there is a core set of basic terminology and models (Speier and Swink 1995; Swink and Way 1995). Concurrent with the formation and solidification of a core terminology, there is an increasing emphasis on developing and

employing a set of reliable, valid, and reproducible methods for conducting research on operations strategy (Vickery, Droge, and Markland 1994; Speier and Swink 1995; Verma and Goodale 1995; Hensley 1999).

Two recent reviews examine relative strengths and weaknesses in the use of empirical methods to study operations issues. Speier and Swink (1995) provide a retrospective examination of the specific research methods that have been used over the past 5–10 years. Their findings indicate that the state of operations strategy research is mixed. While there are several examples of research conducted using rigorous, valid methodologies, they find that the majority of articles are exploratory and theory building. On a positive note, they find that there is growth in the field, namely that more researchers are conducting theory testing research and using rigorous methodologies of the type described in earlier work by Meredith, Raturi, Amoako-Gyampah, and Kaplan (1989) and Flynn et al. (1990). In summarizing their research, Speier and Swink suggest that more attention needs to be given to research methods. Their specific recommendations include an increased emphasis on construct development and validation, the use of more objective data, an increased variety of analytic techniques, the use of longitudinal studies, and the increased use of multiple data sources within an organization. Similarly, a recent review of 25 survey-based research papers by Malhotra and Grover (1998) also indicated that few operations studies employ multiple raters to provide a form of triangulation.

The current research is motivated by a specific methodological shortcoming highlighted by both Speier and Swink (1995), and Malhotra and Grover (1998): the failure of many researchers to use multiple data sources within an organization. There are numerous pitfalls involved in using a single source of data (an individual respondent to a survey, for example) to represent a larger business unit, notably the possibility of subjective bias due to an individual's unique perspective and limited access to information (Jick 1979; Snow and Hambrick 1980). Research based on a single respondent to a mail survey is vulnerable to what can be described as single rater bias—the possibility that a given respondent provides a skewed or warped perspective on the larger business unit being analyzed. For example, suppose two independent raters in similar positions provide vastly different assessments of a manufacturing plant's performance. The disparate ratings weaken the validity of the overall study. This particular problem can occur quite easily due to the frequent use of subjective measures of performance as a means of easily obtaining data.

This paper provides a review of common methods for assessing the degree of reliability of the responses provided by multiple raters within a given organization to a set of qualitative questions. Our intent is to illustrate the value of including multiple raters and multiple methods when conducting survey-based research. We also seek to provide a tutorial or guide to researchers interested in incorporating such methods in their research. We also seek to clarify the differences between inter-rater reliability and agreement, since these terms are often used interchangeably in the literature despite substantive differences in meaning.

Jick (1979) describes the use of multiple raters for a single construct as a "within-methods" strategy for testing reliability. This strategy is one of many types of triangulation, which Jick (1979) groups on a continuum ranging from simple to complex research designs. The use of multiple raters is characterized as being toward the simple end of the continuum and suffering from the limitations involved in using only a single method. Jick (1979) recommends a more holistic approach toward triangulation, one that involves the use of multiple research methods to ensure greater accuracy and capture more complex relationships.

Multiple Rater Reliability and Agreement Measures

It is our position that research that employs multiple raters, while more difficult than using a single rater, provides a greater degree of methodological rigor, thus lending a greater degree of confidence in the findings. We note that it is important that researchers first clearly define their unit of analysis (i.e. plant or strategic business unit, etc.). Once the unit of analysis is clearly delineated, it is important that respondents that can reasonably be expected to have good working knowledge of the constructs being assessed be chosen. Otherwise it is likely that some bias will be introduced due to inappropriate selection of multiple raters within an organization.

This section examines four different methods of assessing inter-rater reliability or agreement. We also distinguish between inter-rater reliability and agreement. These two terms have been used loosely in the literature, in many cases without a clear understanding of the substantive differences between the two. These methods are drawn from the literature and are labeled Correlation, Ratio, Percentage, and Interclass Correlation (ICC). Each measure will be discussed in turn.

Correlation Method

Examples of operations strategy studies that use multiple raters to measure constructs are rare. The earliest that we can find is provided by Nemetz (1989), who compares ratings for four strategic competitive priorities provided by two independent raters within the same firm. The method used to assess inter-rater reliability consists of computing the correlation between the two raters for a given item/construct. The correlations for the four items range from 0.48 to 0.84 and are statistically significant despite a low sample size ($p = 24$), thus indicating a reasonable level of inter-rater reliability. Nemetz (1989) notes that the inter-rater reliability for the quality construct is the lowest of the four constructs and goes on to suggest that this may be due to "attenuation due to restriction of range." Because the variability in the quality construct was low relative to the other constructs in the study, Nemetz suggests that the effect of low variability is to reduce correlations (Allen and Yen 1979; Kozlowski and Hattrup 1992). Thus, despite a relatively low correlation (0.48) there is still good support for a high degree of inter-rater reliability. This phenomenon of attenuation is one of which researchers should be aware. As a simple example of attenuation, consider the following sets of data from two paired raters: Sample 1, ([1,1.5], [2,2.5], [2.5,2]) and Sample 2, ([2,2.5], [4,4.5], [6,5.5]). The

correlation between raters for Sample 1 is 0.65, while the correlation for Sample 2 is 0.98. The lower correlation for Sample 1 is due to attenuation; there is less variability for this sample. Due to the prevalence of Likert type scales, which commonly use a seven-point or five-point range (less than that used by Nemetz), these scales are more prone to attenuation.

One of the challenges associated with the correlation method of assessing inter-rater reliability involves the issue of sample size. What indicates acceptable inter-rater reliability, a correlation above a certain level (i.e., >0.30) or a statistically significant correlation (i.e., $p < 0.05$)? Both standards have problems. If an absolute cutoff is used, the risk is that low correlations will be considered to be unacceptable, despite a general tendency toward low correlations in empirical research due to the inability to control or measure more than a very small number of variables. If a significance level is used, the inherent problem is that our assessment of inter-rater reliability is now directly tied to sample size. Nemetz (1989) had relatively high values on both fronts, in part due to her use of face-to-face interview techniques rather than a mail survey. The greater interaction involved in a face-to-face interview generally reduces problems due to respondent misinterpretation of the questions. Unlike an anonymous mail survey, respondents in a face-to-face interview can ask questions to clarify the researcher's intent. Thus, despite a very small sample size (n = 24), Nemetz (1989) indicated significant correlations for each of her measures.

The correlation method has also been used in two more recent studies. Ward, Leong, and Boyer (1994) report significance values for their four scales of $p < 0.01$ with a sample size of 65. Similarly, Vickery, Droge, and Markland (1994) report both the correlations and the significance values for 10 measures of strategic priority importance and 10 measures of strategic priority performance. The results here are mixed, with 13 of the 20 ratings significant at $p < 0.10$. This suggests a high degree of inter-rater reliability (particularly given that the sample size of 21 is relatively small) but raises a question: what should be done with constructs that do not exhibit high inter-rater reliability? Clearly one of the major difficulties is the sample size, which is very small (the first rater sample size is 65, but 2 responses are available for less than one-third of this sample). A similar challenge was noted by Boyer (1994), who reports a comparable difference in sample size (202 first raters, but second responses from only a subsample of 72) and nonsignificant or low correlations for a minority (but nonzero group of constructs). Boyer (1994) addresses this problem by computing a 95% confidence interval for the difference between the first and second rater. This addresses the issue of attenuation due to restriction of range originally raised by Nemetz (1989). If the confidence interval contains zero, then there is further support for the inter-rater agreement of the constructs. The 95% confidence interval is a test of inter-rater agreement, as opposed to reliability. A confidence interval that contains zero indicates that there is no statistical difference in the values provided by the two raters. The difference between inter-rater reliability and agreement is explored in more detail in the next section.

Reliability Versus Agreement

It is critical to distinguish between inter-rater reliability and agreement. According to Kozlowski and Hattrup (1992), "reliability is referred to as an index of consistency; it references proportional consistency of variance among raters and is correlational in nature". In contrast, "agreement references the interchangeability among raters; it addresses the extent to which raters make essentially the same ratings" (Kozlowski and Hattrup 1992). Despite the common use of the term inter-rater reliability in many published papers, many authors actually measure the degree of agreement (see Shortell and Zajac 1990; Dean and Snell 1991; Snell and Dean 1992; Kotha, Dunbar, and Bird 1995; Kotha and Vadlamani 1995). This mislabeling of measures can lead to paradoxes, since it is possible to have ratings that are different yet proportional; thus, reliability will be high while agreement is low. Likewise, it is theoretically possible to have high agreement with low reliability, although this situation is extremely unlikely to occur in practice. Table 1 illustrates a situation in which there is a high degree of reliability coupled with a low degree of agreement. Researchers should be aware of potential conflicts between measures of reliability and agreement and should therefore choose an appropriate measure for the objectives of a given study. Since agreement implies reliability, but reliability does not imply agreement, it is generally more rigorous to employ measures of agreement than to rely on the correlation method of assessing reliability. Toward this end, we present the correlation technique of Nemetz (1989) as a measure of reliability, but examine three measures of inter-rater agreement in the following sections.

	Rater	
Case	1	2
1	3	5
2	4	6
3	2	5
4	1	3
5	2	3
6	6	7
7	5	7
8	4	6
9	3	6
10	3	5
Mean	3.3	5.3
Correlation = 0.90		
Ratio = 0.48		
Percentage = 20		

Table 1 An example of high reliability with low agreement

Ratio Method

Our first method of assessing agreement between multiple raters was developed by James, Demaree, and Wolf (1984). This method estimates the proportion of true variance relative to true variance plus error variance. Equation (1) is used to compute an index of inter-rater agreement with a maximum value of 1, representing perfect agreement. ${S_{xj}}^2$ is the observed variance on X_j, while ${\sigma_{EU}}^2$ is the variance on X_j that would be expected if all judgements were due exlusively to random measurement error. Thus, $r_{wg(1)}$ "gives the proportion of nonerror variance in the ratings, a reliability coefficient" (Finn 1970). The variance that is expected when judgements are theoretically due exclusively to random measurement errors can be calculated using the equation for the variance of a uniform distribution (Mood, Graybill, and Boes 1978), or ${\sigma_{EU}}^2 = (A^2 - 1)/12$, where A represents the number of alternatives in the response scale. It is important to note that it is possible to have negative values of $r_{wg(1)}$, and when this occurs James, Demaree, and Wolf (1984) recommend setting $r_{wg(1)}$ equal to 0.00.

$$r_{wg(1)} = 1 - \frac{{S_{xj}}^2}{{\sigma_{EU}}^2} \qquad (1)$$

James, Demaree, and Wolf (1984) provide examples to illustrate the calculation of $r_{wg(1)}$ for an item with 10 different raters and a response scale of 5, 7, or 9. This illustrates that $r_{wg(1)}$was originally intended for use in situations with a large number of raters for a single item on a single case. In contrast, the bulk of empirical research in operations strategy employs at most two raters for each item, but does have a large number of cases (different companies/plants). Hence, the methodology used to calculate $r_{wg(1)}$must be modified to accomodate this situation. Table 2 shows an example calculation of $r_{wg(1)}$for sample data representing 2 raters and 10 cases or companies. Here, $r_{wg(1)}$is calculated for each case independently and $R_{wg(1)}$is calculated as the average of the 10 cases. This is the method used by Dean and Snell (1991) and Snell and Dean (1992). The agreement coefficient for this case is 0.86, which indicates a level of agreement consistent with previous research. The correlation of the data for rater 1 and rater 2 is also shown to be 0.60. The method of assessing agreement that employs (1) will be labeled as Ratio throughout the remainder of this paper. Table 1 also contains the Ratio for that data, without accompanying calculations [note that calculated $r_{wg(1)}$for cases 3 and 9 were negative and therefore were updated to 0 while calculating the overall ratio].

Table 3 presents $r_{wg(1)}$ calculations for two raters for a 7-point scale assuming different levels of agreement among the two ratings, ranging from perfect agreement (both ratings 5 7; difference between ratings 5 0) to perfect disagreement (ratings of 1 and 7; difference in ratings 5 6). Table 3 shows that $r_{wg(1)}$ranges from 0 to 1.0 as agreement among the two ratings increases. This table can be used as a "rough guideline" when analyzing $r_{wg(1)}$or $R_{wg(1)}$for a sample data set. For example, a $r_{wg(1)}$of 0.875 (calculation shown in table 2) is approximately equivalent to a difference of 1 between the two ratings.

Case	Rater 1	Rater 2	S_{xj}^2	$r_{wg(1)}$
1	4	3	0.50	0.88
2	5	5	0.00	1.00
3	4	4	0.00	1.00
4	2	3	0.50	0.88
5	3	5	2.00	0.50
6	3	3	0.00	1.00
7	5	6	0.50	0.88
8	4	5	0.50	0.88
9	5	4	0.50	0.88
10	3	2	0.50	0.88
Mean	3.8	4.0		$R_{wg(1)} = 0.88$
$A = 7$				Correlation = 0.60
$\sigma_{EU}^2 = 4.0$				Percentage = 90

Table 2 An example calculation of ratio method

When *J* items ($J > 1$) are used to represent a single construct, James et al. (1984) suggest modifying (1) as

$$r_{wg(J)} = \frac{J\left(1 - \frac{S_j{}^2}{\sigma^2}\right)}{J\left(1 - \frac{S_j{}^2}{\sigma^2}\right) + \frac{S_j{}^2}{\sigma^2}} \qquad (2)$$

The $r_{WG(j)}$ in (2) represents the inter-rater agreement for responses based on J "essentially parallel" indicators of same construct. S_j is the mean of the observed variances on the J items. Similar to the single item case, $r_{WG(j)}$ can be calculated for respondents from each company and their average [$R_{WG(j)}$] can be used as a measure of the aggregate inter-rater agreement based on multiple items.

Table 4 demonstrates the calculation of $R_{WG(j)}$ for a construct represented by three items answered by two individuals from 10 different organizations. The average inter-rater agreement [$R_{WG(j)}$] for this example was found to be 0.876. This value of $R_{WG(j)}$ is approximately equivalent to a difference in rating of 1 between the two raters (see Table 3). A close examination of Table 3 shows that the majority of ratings are indeed one unit apart for the two raters.

Percentage Method

Another method of assessing inter-rater agreement is provided by Shortell and Zajac (1990). This method computes the percentage of paired respones that are either identical or within one category of each other. Shortell and Zajac (1990) report agreement levels ranging from 69 to 82%. Similar agreement levels are reported by Kotha, Dunbar, and Bird (1995) and Kotha and Vadlamani (1995). Applying this measure to our example data shown in Table 2 yields an agreement level of 30% for identical responses, and 90% of the measures are within

one category. This method of assessing agreement will be labeled as Percentage throughout the remainder of this paper. Table 1 shows another illustration of the use of the Percentage method.

	Rater				
Case	1	2	S_{xj}^2	$r_{wg(1)}$	
1	1	7	18.00	−3.5 (=0)	Perfect Disagreement
2	2	7	12.50	−2.1 (=0)	
3	3	7	8.00	−1.0 (=0)	
4	4	7	4.50	−0.1 (=0)	
5	5	7	2.00	0.5	
6	6	7	0.50	0.88	
7	7	7	0.00	1.00	Perfect Agreement

Table 3 Possible Agreement Values

	Rater 1			Rater 2				
Case	Item 1	Item 2	Item 3	Item 1	Item 2	Item 3	Average S_{xj}^2	$r_{wg(j)}$
1	4	5	4	3	4	5	0.500	0.955
2	5	7	6	5	5	4	1.333	0.857
3	4	3	4	4	3	4	0.000	1.000
4	2	3	4	3	4	5	0.500	0.955
5	3	4	2	5	7	5	3.667	0.214
6	3	4	5	3	4	4	0.167	0.986
7	5	6	4	6	5	4	0.333	0.971
8	4	5	5	5	7	6	1.000	0.900
9	5	4	5	4	3	5	0.333	0.971
10	3	4	5	2	3	4	0.500	0.955
Mean	3.8	4.5	4.4	4.0	4.5	4.6		$R_{wg(1)} = 0.876$

Table 4 Agreement for Multiple Item Measures

ICC Method

The final method for assessing inter-rater agreement presented in this research note is known as Interclass Correlation (ICC). ICC represents a family of reliability indices developed in the psychology literature. The exact equation for ICC varies according to the experimental design used. For further information about the differences among ICC equations please refer to Ebel (1951), Tinsley and Weiss (1975), Shrout and Fleiss (1979), Lahey, Downey, and Saal (1983), and Futrell (1995). Each of the various ICC indices is based on a partitioning of variance among two or more components. We focus on a single ICC index that is most applicable to operations strategy researchers and is common to the studies cited above.

The total variance in the empirical data collected from multiple respondents from

different organizations can be classified as one of two components: (1) Within group (MSW) variance and (2) Between group (MSB) variance. The MSW represents the average variation in responses from multiple respondents within the same organization. The MSB represents the average variation between the responses from different organizations. If a high level of agreement exists among the respondents from the same organization, then MSW would be small compared to MSB. ICC is a representation of the fraction of between group variation that does not contain within group variation. Therefore ICC is defined as:

$$ICC = \frac{MSB - MSW}{MSB} \qquad (3)$$

The ICC index has a maximum value of 1, with higher values representing more reliable data, or data in which the majority of variance is between groups or companies rather than within groups. The MSB and MSW can be calculated by a one-way ANOVA using organizations as groups. The ANOVA statistically tests if MSW is negligible compared to MSB. In other words, the ICC for a statistically significant ANOVA suggests a high degree of inter-rater agreement.

Table 5 presents ANOVA results and ICC calculations for the data presented earlier in Table 2. MSB (2.09) is much larger than MSW (0.50), and the resulting F-ratio (4.18) is statistically significant ($p < 0.05$). The corresponding ICC was found to be 0.76, which indicates that approximately 76% of between group variance is free from variation within groups or companies.

Although a number of articles have presented ICC measures for different situations, none have shown how to calculate ICC for constructs with multiple item measures (as shown in

Source	SS	d.f.	*MS*	*F*	*p*-value
Between Groups	18.80	9	2.10	4.18	0.02
Within Groups	5.00	10	0.50		
Total	23.80	19			

ICC = 2.10–0.50/2.10 = 0.76. Note: example is based on the data contained in Table 2.

Table 5 Example calculation of ICC Index

Table 4 for the Ratio method). For multiple item constructs, the total variation includes variations due to different items used. However, an average of the ICCs for individual items can be used as an aggregate measure of ICC for a multiple item construct. Table 6 presents the calculation of this aggregate ICC measure for the multiple item data shown earlier in Table 4. Applying this method involves calculating ICC for the individual items and then taking an average. The ICCs for items 1, 2, and 3 were found to be 0.76, 0.59, and 20.35. A negative ICC (such as 20.35 for Item 3) score means that within group variation is far greater than the between group variation. When this happens, we suggest that the ICC index for that item should

be updated to 0 (similar to ratio calculations) or the item should be eliminated from the scale. When a negative ICC occurs, this may be indicative of a situation where the measure is misapplied or is an inappropriate measure. Therefore, researchers should carefully consider their treatment of this item, with the elimination of this item from the scale serving as a conservative approach to maintaining reliability. Table 6 shows the aggregate ICC score for the resulting two item scale (items 1 and 2) as 0.674 and for three-item scale as 0.449. Another method of calculating an ICC for multiple item constructs would be to calculate the ICC based on the aggregate scale (sum/average of items 1, 2, and 3) scores rather than the individual item scores.

Source	*SS*	d.f.	*MS*	*F*	*P*-value
Item 1*					
Between Groups	18.80	9	2.08	4.18	0.01
Within Groups	5.00	10	0.50		
Total	23.80	19			
Item 2†					
Between Groups	24.00	9	2.67	2.42	0.09
Within Groups	11.00	10	1.10		
Total	35.00	19			
Item 3‡					
Between Groups	6.00	9	0.67	0.74	0.67
Within Groups	9.00	10	0.90		
Total	15.00	19			

Note: example is based on the data contained in Table 4. Aggregate ICC for 2-Item Scale (Items 1 and 2) = (0.76 + 0.59)/2 = 0.674. Aggregate ICC for 3-Item Scale = (0.76 + 0.59 + 0)/3 = 0.449.

* $ICC = (2.08–0.50)/2.08 = 0.76$.

† $ICC = (2.67–1.10)/2.67 = 0.59$.

‡ $ICC = (0.67–0.90)/0.67 = -0.35$ = updated to 0.00.

Table 6 Example ICC calculation for Multiple Item constructs

Source	Correlation Nemetz (1989)	Ratio James et al. (1984)	Percentage Shortell and Zajac (1990)	ICC Ebel (1951)
Advantages	• Subject to statistical tests	• Easy to interpret with clear range from 0.0 to 1.0 • Applicable for multiple raters (more than two)	• Easy to compute	• Subject to statistical tests • Easy to interpret as a percentage • Applicable for multiple raters (more than two) • Intuitively logical
Disadvantages	• Susceptible to attenuation of range • Designed for only two raters, but multiple cases • No standards for "acceptable" levels • Correlations tend to be low, even for raters with high agreement	• Developed for multiple raters (more than 2) and a single case • No standards for "acceptable" levels	• No standards for "acceptable" levels—does not account for standard deviation of responses • Designed for single item measures	• Not well established in operations strategy literature
Suggested Standard	• 0.20 or higher	• 0.80 or higher	• 85% within one response category • 50% identical	• 0.60 or higher

Table 7 Overview of methods of assessing reliability/agreement

A Comparison of Measures

The previous section examined four different methods of assessing inter-rater reliability/ agreement. In contrast to inter-item reliability, where Cronbach's coefficient alpha is the accepted standard measure (Flynn et al. 1990), there is no clear consensus regarding which measure of inter-rater reliability/agreement is best. Therefore, we present a discussion of the relative merits and deficiencies of each measure. Table 7 provides a summary of the advantages and disadvantages associated with each of the four measures of reliability/ agreement.

We start with the simplest measure to calculate, Percentage. The primary advantage of this measure is its ease of use. This measure requires only a counting of values that are either identical, or within a single response category for two different raters. When the percentages that are identical or within a single response category are higher, the agreement between the two raters is higher.

While easy to calculate and interpret, there are several drawbacks to the Percentage method. First, there are no standard cutoff values for what represents "good" agreement and no method of performing a statistical test. The test does not account for either the number of response categories (wider scales such as a nine-point versus a seven-point scale, are less likely to have two raters differ by a single category, ceteris paribus) or the amount of standard deviation for an item. For example, when an item has a large standard deviation combined with a small difference between two raters, the percentage method is less likely to detect agreement differences than when an item has a small standard deviation combined with a large difference between two raters. In addition, the test is designed to be used for only two raters and for a single item. The test will not work in situations with more than two raters, or for a scale comprised of two or more individual questions, although modifications to this method could be made.

The Ratio method of assessing inter-rater agreement is more sophisticated than the Percentage method, providing a measure that ranges from 0.0 to 1.0 and is easy to interpret. This method is applicable for multiple raters and has been used successfully in a number of studies in the fields of psychology and strategy (James et al. 1984; Dean and Snell 1991; Snell and Dean 1992). In the absence of accepted standards regarding what indicates "good" agreement, we suggest that 0.80 be used as a rough standard. Based on a review of published results, this value appears to be one that 70–80% of reported measures exceed.

The Correlation method assesses reliability rather than agreement. Its major advantage is that statistical tests of significance can be applied to the results of this measure. This method has been used in several published studies of operations strategy. Unfortunately, this method is susceptible to attenuation of range, potentially with deceptively low values. In addition, the Correlation method is only applicable for two raters.

The ICC method possesses several advantages over the other three methods. Like the Correlation method, it can be tested for statistical significance. In contrast to the Correlation method, the ICC method is not burdened with the same disadvantages. The ICC method is more readily interpretable because it represents a percentage of variance that is free from within group variance. Thus, when the sample size is low, the ICC still provides easily interpretable data, even if the associated ANOVA test is insignificant due to the small sample size. Furthermore, the ICC method is applicable for multiple raters and is intuitively logical.

We have developed a list of several advantages and disadvantages for the four measures (see Table 7). Obviously, there are trade-offs between the measures, and each researcher must make their own decision regarding which method is most appropriate. Nevertheless, we believe that the ICC method offers the best overall assessment of inter-rater agreement/reliability. Although this method is not well established in the operations strategy literature, we feel that it should be the first method of choice for studies employing multiple raters. As further research is done in this area, an accepted standard for what is considered to be "good" inter-rater agreement should emerge. For now, we note that Futrell (1995) recommends 0.70 as an acceptable lower bound. This standard corresponds to a situation in which the between group variance is approximately three times as much as the within group variance. Given the absence of existing operations strategy articles that apply this measure, we suggest that an initial standard of 0.60 provides a more pragmatic standard that can be readily changed as research in this area evolves.

Summary and Future Research

This paper has examined various methods of assessing the reliability and agreement of multiple raters or respondents to a survey. Despite clear deficiencies in each of the existing measures, it is critical that research in operations strategy employ some combination of these measures. The majority of current research in operations strategy employs only a single respondent and is thus vulnerable to the possibility of single rater bias in which a biased respondent may provide skewed or inaccurate data. The collection of data from multiple respondents provides a means of assessing the reliability and agreement of respondents within the same organization, and thus provides an increased level of validity.

In addition to using measures of inter-rater reliability and agreement purely as a component of a broader research study, there are several potential research ideas that focus on a comparison of the viewpoints of people occupying different positions within an organization. We briefly discuss two potential areas in which an examination of inter-rater reliability/agreement would yield interesting insights into operations strategy.

RESEARCH IDEA 1. *In what situations would we expect to have different ratings for a given construct from independent raters within an organization*?

In certain situations, we might expect responses for a given construct to differ depending on the individual providing the rating and their position within an organization. For

example, several studies have assessed the degree of commitment to empowering workers (Ward et al. 1994), yet the responses to questions regarding this construct are likely to differ depending on whether the questions are asked of managers or lower level employees. Similarly, one of Deming's 14 principles of total quality management (TQM) holds that top management must be totally committed to providing quality products and must work every day to achieve this goal (Deming 1986). Yet, perceptions of managerial commitment to quality are quite likely to differ based on whether the question is asked of top managers or lower level employees. For example, Flynn, Schroeder, and Sakakibara (1994) develop a scale to measure quality leadership and pose this set of questions to a variety of respondents (plant manager, supervisors, and process engineers). While their goal to develop a reliable set of scales to measure TQM was achieved, there was no assessment of inter-rater reliability or agreement. Given the availability of multiple respondents in different positions for each plant, these data could have been used to address the question: Is quality leadership perceived differently at different levels? This type of analysis should yield interesting insights regarding the degree of uniformity of employee perceptions of quality leadership within a manufacturing plant.

> RESEARCH IDEA 2. *How much agreement between multiple, independent raters for a given organization represents a coherent, consistent operations strategy? What are the relationships between the degree of agreement and the performance of the organization*?

Operations strategy has been defined by Hayes and Wheelwright (1984, p. 30) as a "pattern of decisions actually made." The implication is that even a strategy that is perfectly tailored for a given organization will not prove effective unless there is organization wide understanding and comprehension so that decision makers at all levels of the organization are able to work in tandem. Several researchers have pointed out that there is a strong need to study the process of developing an operations strategy, in addition to the content (Adam and Swamidass 1989; Anderson, Cleveland, and Schroeder 1989; Leong, Ward, and Snyder 1990). While there has been an increased focus on process over the last several years, there have not been studies that measured the consistency of different raters' views regarding operations strategy. The inter-rater reliability/agreement methods examined in this paper should be combined with more in-depth research on one or a few organizations in order to assess how strategic views differ among independent raters within a company. Based on the literature, our expectation would be that companies with a greater degree of dissemination and consistency in their strategy would be more likely to be successful than companies that have discordance and widely varying perceptions regarding their operations strategy.

Conclusion

In response to several recent assessments of the state of research on operations strategy, we have provided a review of methods for assessing the reliability and agreement of multiple raters in survey-based studies. Survey-based research that relies on a single respondent may be biased because that respondent may potentially present a skewed or

inaccurate view of the organization as a whole. To address this difficulty, we have presented several methods drawn from the literature for assessing the degree of reliability and agreement of multiple respondents within an organization. One method of assessing inter-rater reliability and three methods of assessing agreement have been presented. Reliability and agreement are not identical, despite a tendency in the literature to use these terms interchangeably. Since it is possible to have high reliability with low agreement, it seems prudent that researchers should focus on the use of agreement rather than reliability measures. Finally, we also present suggested standards for each measure. These standards are based on our evaluation of previous research and are designed to encourage greater standardization of the use of inter-rater agreement measures.

We believe that it is important that research in operations strategy move toward the more holistic approach described by Jick (1979). Such an approach involves the use of multiple respondents or the triangulation of data. This type of research provides more accurate information than relying on a single respondent, who may or may not provide an unbiased representation of the organization as a whole. As the field of operations strategy evolves, it is essential that research methods evolve also. Our examination of methods for assessing the reliability and agreement of data provided by multiple respondents is intended to serve as a rough guide and an initial exploration of the issue for researchers in this area.

References

ADAM, E. E. AND P. M. SWAMIDASS (1989), "Assessing Operations Management from a Strategic Perspective," *Journal of Management*, 15, 2, 181–203.

ALLEN, M. J. AND W. M. YEN (1979), *Introduction to Measurement Theory*, Brooks/Cole Publishing Co., Monterey, CA.

ANDERSON, J. C., G. CLEVELAND, AND R. G. SCHROEDER (1989), "Operations Strategy: A Literature Review," *Journal of Operations Management*, 8, 2, 133–158.

BOYER, K. K. (1994), *Patterns of Advanced Manufacturing Technology Implementation—Technology and Infrastructure*, Unpublished doctoral dissertation, The Ohio State University, Columbus, OH.

DEAN, J. W. AND S. A. SNELL (1991), "Integrating Manufacturing and Job Design: Moderating Effects of Organizational Inertia," *Academy of Management Journal*, 34, 4, 776–804.

DEMING, W. E. (1986), *Out of the Crisis*, Massachusetts Institute of Technology Center for Advanced Engineering Study, Cambridge, MA.

EBEL, R. L. (1951), "Estimation of the Reliability of Ratings," *Psychometrika*, 16, 4, December, 407–424.

FINN, R. H. (1970), "A Note on Estimating the Reliability of Categorical Data", *Educational and Psychological Measurement*, 30, 71–76.

FLYNN, B. B., S. SAKAKIBARA, R. G. SCHROEDER, K. A. BATES, AND E. J. FLYNN (1990), "Empirical Research Methods in Operations Management," *Journal of Operations Management*, 9, 2, April, 250–284.

———, R. G. SCHROEDER, AND S. SAKAKIBARA (1994), "A Framework for Quality Management Research and an Associated Measurement Instrument," *Journal of Operations Management*, 11, 4, 339–366.

FUTRELL, D. (1995), "When Quality is a Matter of Taste, Use Reliability Indexes," *Quality Progress*, 28, 5, May, 81–86.

HAYES, R. H. AND S. C. WHEELWRIGHT (1984), *Restoring Our Competitive Edge: Competing Through Manufacturing*, John Wiley & Sons, New York.

HENSLEY, R. (1999), "A Review of Operations Management Studies Using Scale Development Techniques," *Journal of Operations Management*, 17, 3, 343–358.

JAMES, L. R., R. G. DEMAREE, AND G. WOLF (1984), "Estimating Within-group Interrater Reliability With and Without Response Bias," *Journal of Applied Psychology*, 69, 85–98.

JICK, T. D. (1979), "Mixing Qualitative and Quantitative Methods: Triangulation in Action," *Administrative Science Quarterly*, 24, 602–611.

KOTHA, S., R. DUNBAR, AND A. BIRD (1995), "Strategic Action Generation: A Comparison of Emphasis Placed on Generic Competitive Methods by U.S. and Japanese Managers," *Strategic Management Journal*, 16, 192–220.

——— AND B. L. VADLAMANI (1995), "Assessing Generic Strategies: An Empirical Investigation of Two Competing Typologies in Discrete Manufacturing Industries," *Strategic Management Journal*, 16, 75–83.

KOZLOWSKI, W. J. AND K. HATTRUP (1992), "A Disagreement About Within-Group Agreement: Disentangling Issues of Consistency Versus Consensus," *Journal of Applied Psychology*, 77, 2, 161–167.

LAHEY, M. A., R. G. DOWNEY, AND F. E. SAAL (1983), "Interclass Correlations: There's More There Than Meets the Eye," *Psychological Bulletin*, 93, 3, 586–595.

LEONG, G. K., D. SNYDER, AND P. WARD (1990), "Research in the Process and Content of Manufacturing Strategy," *Omega*, 18, 109–122.

MEREDITH, J., A. RATURI, K. AMOAKO-GYAMPAH, AND B. KAPLAN (1989), "Alternative Research Paradigms in Operations," *Journal of Operations Management*, 8, 4, 297–326.

MALHOTRA, M. AND V. GROVER (1998), "An Assessment of Survey Research in POM: From Constructs to Theory," *Journal of Operations Management*, 16, 4, 407–425.

MOOD, A. M., F. A. GRAYBILL, AND D. C. BOES (1978), *Introduction to the Theory of Statistics*, McGraw-Hill, New York.

NEMETZ, P. L. (1989), *Flexible Manufacturing Strategies*, *Technologies*, *and Structures*: *A Contingency-Based Empirical Analysis*, Unpublished doctoral dissertation, University of Washington, Seattle, WA.

SHORTELL, S. M. AND E. J. ZAJAC (1990), "Perceptual and Archival Measures of Miles and Snow's Strategic Types: A Comprehensive Assessment of Reliability and Validity," *Academy of Management Journal*, 33, 4, 817–832.

SHROUT, P. E. AND J. L. FLEISS (1979), "Interclass Correlations: Uses in Assessing Rater Reliability," *Psychological Bulletin*, 86, 2, 420–428.

SNELL, S. A. AND J. W. DEAN (1992), "Integrated Manufacturing and Human Resource Management: A Human Capital Perspective," *Academy of Management Journal*, 35, 3, 467–504.

SNOW, C. C. AND D. C. HAMBRICK (1980), "Measuring Organizational Strategies: Some Theoretical and Methodological Problems," *Academy of Management Review*, 4, 4, 527–538.

SPEIER, C. AND M. SWINK (1995), "Manufacturing Strategy Research: An Examination of Research Methods and Analytical Techniques," Indiana University Working Paper.

SWAMIDASS, P. M. (1991), "Empirical Science: New Frontier in Operations Management Research," *Academy of Management Review*, 16, 4, 793–814.

SWINK, M. AND M. H. WAY (1995) "Manufacturing Strategy: Propoditions, Current Research, Renewed Directions," *International Journal of Operations and Production Management*, 15, 7, 4–26.

TINSLEY, H. E. A. AND D. J. WEISS (1975), "Interrater Reliability and Agreement of Subjective Judgement," *Journal of Counseling Psychology*, 22, 4, 358–376.

VERMA, R. AND J. C. GOODALE (1995), "Statistical Power in Operations Management Research," *Journal of Operations Management*, 13, 139–152.

VICKERY, S. K., C. DROGE, AND R. E. MARKLAND (1994), "Strategic Production Competence: Convergent, Discriminant, and Predictive Validity," *Production and Operations Management*, 3, 4, Fall, 308–318.

WARD, P., G. K. LEONG, AND K. K. BOYER (1994), "Manufacturing Proactiveness and Performance," *Decision Sciences*, 25, 3, May/June, 337–358.