Data Aggregation in Multilevel Analysis: 
A Review of Conceptual and Statistical Issues

Marlene A. Dixon  
Department of Kinesiology and Health Education  
The University of Texas at Austin

George B. Cunningham  
Department of Health and Kinesiology  
Texas A&M University at College Station

Understanding that the behavior of people takes place within a context, over the past 20 years research in education and the sport sciences has witnessed an increasing development of multilevel frameworks that are both conceptually and methodologically sound. Despite these advances, the use of multilevel models and research designs in education still could be improved. As noted by recent editorial commentaries and reviews, confusion often exists about appropriate levels of analysis both at the conceptual and statistical level. The purpose of this article is threefold. First, we explain how multilevel conceptualizations can aid our understanding of behavior, especially within a given context. Second, we review three aggregation methods for analyzing multilevel data: within group agreement, intraclass correlations, and within and between analysis. Third, we provide a conceptual and statistical example to demonstrate how these aggregation techniques could be utilized to answer questions in sport and physical activity.

Key words: multilevel analysis, aggregation, organizational behavior, sport, physical activity

Understanding that organizations are multilevel systems, over the past 20 years research in education, organizational behavior, and physical activity has witnessed
an increasing development of multilevel frameworks that are both conceptually and methodologically sound (Kozlowski & Klein, 2000; Richter et al., 2000). Dansereau, Alutto, and Yammarino’s (1984) Varietal Paradigm marked the beginning of a blossoming collection of research that has produced and refined methods for testing theoretical assertions dealing not only with relations between variables, but also with relations among organizational entities (e.g., individuals, teams, departments, industries, etc.). In education, researchers have long recognized the need to examine effects related to individual students, classrooms, schools, and entire school districts (Ostroff, 1992). In sport and exercise science, Zhu (1997b) demonstrated the viability of Hierarchical Linear Modeling (HLM) for demonstrating treatments at the group level, an analysis strategy that has since been reviewed and utilized in a number of studies (e.g., Duncan, Duncan, Strycker, & Chaumeton, 2002, 2004; O’Connell & McCoach, 2004; Zhu, 1997a).

Despite these advances, the use of multilevel models and research designs in education and social science still could be improved (Kozlowski & Klein, 2000). As noted by Silverman and Solmon (1998), confusion often exists about appropriate levels of analysis both at the conceptual and statistical level. Their review and a follow-up by Silverman (2004) addressed the issue of units of analysis in research conceptualization and design in an in-depth and instructive manner. Others have addressed the derivations of the statistical background and application of specific analytic methods (see e.g., Dansereau et al., 1984, on within and between analysis; see Bartko, 1976; Bliese, 2000; or McGraw & Wong, 1996, for intraclass correlations; see Raudenbush & Bryk, 2001, or Hofmann, 1997, for hierarchical linear modeling; and see James, Demaree, & Wolf, 1984, for within rater agreement). Though these works are invaluable for conducting the analysis, questions may exist regarding which design or statistical analysis is appropriate for a given study, or how to actually compute and interpret the statistics. This is especially true for designs where a specific level may not be prespecified or where one wishes to aggregate to a higher level, and assess the reliability of doing so. For example, Silverman (2004) suggested that treatments conducted on a work-group should be analyzed at the group level (because that is the appropriate level of analysis), though data will probably be collected at the individual level. Aggregation techniques such as those reviewed here assist researchers in assessing the confidence with which they can assert that their group mean is a reliable measure of a group property.

Given the need for multilevel research in sport and physical activity, the purpose of this article is to provide a comparison of the various methods of data aggregation to aid researchers in the appropriate design, analysis, and interpretation of some multilevel models. To this end, we first explain how multilevel conceptualizations can aid our understanding of both structure and behavior. Second, we review three statistical aggregation techniques that are used to determine the most appropriate level of analysis. Third, we provide a conceptual and statistical exam-
ple to demonstrate how these aggregation techniques could be utilized to answer questions in a research application.

OVERCOMING BARRIERS: THE USEFULNESS OF MULTILEVEL ANALYSIS

Proponents stress the need for multilevel analysis for at least two reasons. First, multilevel analysis allows the researcher to address the complexity of relations between variables. As Kozlowski and Klein (2000) noted, “Organizations are multilevel systems” (p. 3). Indeed, this principle is espoused in some of the earliest contributions to behavioral theory, including Lewin’s (1951) field theory, Thompson’s (1967) theory of organizational rationality, and Katz and Khan’s (1966) social organizational theory. In recognition of the multilevel orientation of organizations, we see that activities that take place in one part of the organization can impact outcomes at other (or the same) levels of the organization. For example, the task conflict present within teams can impact individuals’ affective reactions (i.e., satisfaction and stay intentions) and the performance of the overall group (Jehn, 1995). Within general education and physical education settings, these dynamics have been noted at the student–student, student–teacher, and student–classroom levels (e.g., Arnett & Lutz, 2003; Eitle & Eitle, 2003; Ostroff, 1992). Recent examples in health and physical activity applications call for the examination of individuals or groups within their larger social context such as family, neighborhood, or region (Duncan et al., 2002, 2004; Fisher, Li, Michael, & Cleveland, 2004; Maes & Lievens, 2003). Therefore, multilevel theories extend over various levels of social settings, describing the individual, dyad, team, division, organization, or industry (Klein, Tosi, & Cannella, 1999). For example, Xiang, McBride, and Guan’s (2004) investigation of individual factors on children’s motivation to participate in physical education follows the individual paradigm, whereas Knoppers’s (1992) work concerning the gendered nature of sport organizations and coaching more closely follows the broader organizational-level tradition.

Though both perspectives have some explanatory value, neither single-level perspective can adequately explain behavior in organizations. As Kozlowski and Klein (2000) argued, a global organization or industry-level perspective neglects the manner by which individual’s attitudes, perceptions, and affect can give rise to higher level phenomena. Conversely, the individual perspective ignores the manner in which contextual factors can “constrain the effects of individual differences that lead to collective responses, which ultimately constitute macro phenomena” (Kozlowski & Klein, 2000, p. 7). Thus, we see that reliance on a single level of analysis can constrain the predictive validity of our research. Rather, it is through
the integration of these perspectives and the formulation of multilevel perspectives and theories that we are able to gain a richer understanding of behavior.

Second, multilevel analysis offers tools for appropriate measurement and analysis of different entities at different levels of analysis. These unit of analysis problems have arisen in a number of areas. The first problematic area is the design stage, where a treatment is applied at one level, but analyzed at another. As Silverman (2004) argued, if a treatment is applied to groups, then the group, not the individuals, becomes the unit of analysis, though often the construct is measured at the individual level.

Though clear individual (e.g., age, skill level, gender) and global (e.g., win–loss record for a team, neighborhood size, number of recreation centers) measures are rather straightforward, shared and configural properties are more tenuous and often call for special treatment in both measurement and analysis (Kozlowski & Klein, 2000). Shared and configural properties emerge from the individual level and should be measured as such. Then, the data can be examined to determine if, in fact, the property is shared (see also Silverman & Solmon, 1998). For example, a physical educator may want to assess the knowledge level of his or her students as a class to see if their knowledge can be understood as a shared property of the class or only as an individual property (Eccles & Tenenbaum, 2004). In this example, student knowledge is measured at the individual level. To assess knowledge at the group level, the data must be aggregated. Then, the researcher must interpret the data to assess whether the property is best conceptualized as an individual property or a shared one, and if it is shared, how to present the group data in a meaningful way.

In past applications, many researchers have simply utilized the group mean to represent group knowledge; one cannot simply assume, just because it is a convenient statistic, that the group mean is an appropriate estimate of group knowledge. There are at least two reasons that the mean may not be appropriate. First, group properties may emerge in various configurations or patterns (Eccles & Tenenbaum, 2004). For example, within a given classroom, activity levels may be very high for some students but very low for others (a bimodal distribution). These patterns would be lost by simply reporting a global measure or a “group” average (Kozlowski & Klein, 2000). Second, the students might demonstrate such a broad range of knowledge (i.e., little agreement) that the researcher must conclude that it is not a shared property; it is best conceptualized (and interpreted) at the individual and not the group level.

As shown in these examples, the utilization of multilevel techniques can clearly aid researchers in understanding behavior. To derive the most benefit from this type of analysis, however, appropriate conceptual and statistical tools must be utilized. The following sections outline examples of multilevel models, followed by a review of three potential aggregation techniques for testing them.
MULTILEVEL MODELS:
CONCEPTUAL AND DESIGN ISSUES

Though single-level analysis is certainly not without merit, and is called for in a variety of research designs and applications, research concerning the behavior of people within a context can be greater enhanced through the adoption of multilevel theories, models, and analytical techniques. Given this need, it is necessary to understand the types of multilevel conceptual models that can be used to examine various organizational phenomena. Klein, Dansereau, and Hall (1994) provided a useful typology for these purposes. Later, we describe the models and give examples of research that follows the particular paradigm.

Klein et al. (1994) outlined four multiple-level models: cross-level, mixed-effects, mixed-determinants, and multilevel. The authors did not suggest a hierarchy in these models (i.e., that one was superior to the others); rather, they argued that they simply represent different multiple-level models scholars can follow in conceptualizing their research. Similar to Rousseau’s (1985) typology, Klein et al. described cross-level models as those which contain independent and dependent variables at different levels. Other authors have referred to such models as multilevel models (see Behling, 1978; Dansereau et al., 1984). As Klein et al. noted, the most common cross-level models describe the impact of situational factors on individual reactions or behaviors. For example, Inglis, Danylchuk, and Pastore (1996) highlighted organizational factors (i.e., Work Balance and Conditions, Recognition and Collegial Support, and Inclusivity) thought to impact coaches’ and administrators’ decisions to leave or remain in the organization. As another example, Maiga, Haddad, Fournier, and Gauvin (2003) demonstrated the impact of policy-level factors (public vs. private health organizations) on the cost of health care for individuals.

Second, Klein et al. (1994) described mixed-effects models; in these models a single variable or intervention has an effect on other variables at various levels within the social system. These designs are less frequently utilized. As an example, Jehn, Northcraft, and Neale (1999), in their study of work teams, examined the relation between group diversity and various outcomes, both at the group and individual level. These researchers found that value diversity was negatively related to (a) group performance and efficiency (group-level outcomes), and (b) satisfaction, stay intentions, and commitment of the participants (individual-level outcomes).

The third model in Klein et al.’s (1994) typology is a mixed-determinants model. This model postulates that predictors at a variety of levels influence a specific criterion. In the social sciences, the designs usually postulate the combined influence of individual and situational factors on individual outcomes. For example, Duncan and colleagues (2004) investigated the influence of sibling activity
levels, family factors (activity level, size), and neighborhood characteristics (facilities, perceived opportunity) on children’s activity levels. Maes and Lievens (2003) investigated the influence of school and personal factors on adolescent risk behaviors such as smoking and drug use. Fisher et al. (2004) examined neighborhood characteristics (density, availability of recreation centers, income) and individual factors (age, income, gender) on the physical activity levels of aging adults. Researchers have been able to partition the variance such that a certain amount of influence is ascribed to each level (e.g., individual, family, neighborhood, etc.).

Multilevel models represent the final type of model in Klein et al.’s (1994) typology. This type is similar to Rousseau’s (1985) conceptualization of multilevel models and Dansereau et al.’s (1984) notion of cross-level models. With multilevel models, patterns of relations are replicated across various levels of analysis. Within this model, “individuals, groups, and organizations each are conceptualized as homogeneous and independent of higher level units” (Klein et al., 1994, p. 223). After each level is examined independently, effects between levels can be analyzed. For example, within national sport governing bodies, the associations between (a) president effort and organizational level performance, (b) team effort and team performance, and (c) individual effort and individual performance may all be related within the organization. In multilevel analysis, each level would be tested separately (often nesting individuals within teams, and teams within organizations), then the analysis would include between-level comparisons to examine the patterns of relations.

Given that theory development and research questions or hypotheses asked at various levels of analysis call for different analytical procedures, it is critical for researchers to choose appropriate data analytic methods to complement their theory and design. Though more comprehensive multilevel analysis techniques (e.g., HLM) and cross-level operator (CLOP) analysis are available (see Castro, 2002; Hofmann & Gavin, 1998; and James & Williams, 2000, for thorough reviews of these methods), the focus of this article is on the common denominator of most multilevel analysis, which is data aggregation.

Though some would argue that the need for stand-alone data aggregation techniques is becoming more limited with the advent and demonstrated usefulness of more comprehensive multilevel analysis, data aggregation techniques are still quite valuable for at least two reasons. First, they are valuable for designs where the levels have not been specified a priori or where researchers would like to aggregate to the group level, but must assess the appropriateness of doing so (Castro, 2002; Dansereau et al., 1984; Dansereau & Yammarino, 2000). Second, the evidence provided by these analyses can be utilized to justify subsequent use of HLM or other nesting-variance apportioning analysis (Castro, 2002; Hofmann, 1997).

The following sections highlight the importance of data aggregation techniques for these types of applications. For example, if we hypothesized that more satisfied teachers within the school will result in higher school performance (e.g., Ostroff,
1992), but we collected data at the individual level (from the teachers), then we would need to choose the appropriate statistical tools to investigate whether individual satisfaction scores reliably aggregate to the school level before proceeding with additional relationship-type analysis (e.g., HLM, multivariate analysis of variance; Hofmann, 1997). Because methods such as HLM and CLOP represent entire analytical systems that deal not only with aggregation but with multiple relations with multiple variables at multiple levels of analysis, they are beyond the scope of this article. This does not, however, discount their importance and the reader is advised to consult texts such as Raudenbush and Bryk (2001) or Hofmann (1997) for a thorough explanation of the use of these techniques once aggregation decisions are made.

STATISTICAL TECHNIQUES

Three statistical techniques that can be utilized to examine multilevel data are reviewed in this section: within-group interrater agreement (rwg), intraclass correlations (ICC), and within and between analysis (WABA). The focus of these three techniques is solely on the decision of aggregation, which concerns questions of whether individual level variables constitute sufficient agreement to warrant aggregation or if they are more accurately represented as individuals. Once these decisions are reached, researchers can employ additional multilevel techniques if warranted.

Within-Group Interrater Agreement

Within-group agreement refers to the degree to which ratings from different persons within a group are interchangeable (Bliese, 2000; James, Demaree, & Wolf, 1993; Kozlowski & Hattrup, 1992). Thus, the agreement reflects the extent to which all members essentially provided the same ratings. The most frequently used measure of within-group interrater agreement is rwg, which is a measure calculated by comparing the observed group variance to the expected random variance (James, Demaree, & Wolf, 1984; James et al., 1993). One strength of the technique is that, unlike other measures such as ICC, rwg is not based on between-group variance and provides a measure of agreement for each individual group on each individual item (Bliese, 2000). The latter point is particularly important if the researcher wishes to identify which specific items show agreement within their own group (Bliese, 2000).

To fully understand rwg, let us consider an example of aging women rating the effectiveness of a new exercise education program using a single-item measure anchored by a 5-point Likert-type scale from 1 (strongly disagree) to 5 (strongly
$S_x^2$ is the observed rater variance from the rating of variable $x$, or in our case, the rating of program effectiveness ($S_x^2 = \text{standard deviation of the raters’ scores squared}$). If raters were in complete agreement, $S_x^2 = 0$; however, because raters do not agree and due to measurement errors, $S_x^2$ is likely to be greater than 0. Because $S_x^2$ values greater than 0 occur due to errors, this is referred to as error variance and is essentially the sum of actual variance and measurement error. To develop a measurement of agreement among raters, it is necessary to develop a benchmark to which $S_x^2$ can be compared. Because $S_x^2 > 0$ represents deviation from agreement, this benchmark must reflect $S_x^2$ in a condition where departure from 0 is due to random measurement errors. This expected random variance is referred to as $\sigma_E^2$, and represents “a theoretical benchmark for responses attributed totally to random error measurement errors” (James et al., 1993, p. 307).

The next step is to consider the extent to which the actual ratings represent a reduction in error variance from the theoretical benchmark. “The term reduction in error variance refers to the degree to which the observed error (between-rater) variance ($S_x^2$) indicates a decrease in the variation of judgments relative to $\sigma_E^2$ and is estimated by the difference between $\sigma_E^2$ and $S_x^2$” (James et al., 1993, p. 307, emphasis original). Therefore, if the women’s assessment of program effectiveness was completely due to error, then $\sigma_E^2 \equiv S_x^2$. Conversely, if there was no measurement error in the women’s ratings, then $S_x^2 \equiv 0$. James et al. (1984, 1993) then converted the measure to account for reduction in error variance. This conversion “involves dividing the actual reduction in error variance by the maximum possible reduction in error variance” (James et al., 1993, p. 307), resulting in the following formula:

$$r_{wg} = (\sigma_E^2 - S_x^2) / \sigma_E^2 = 1 - \left( S_x^2 / \sigma_E^2 \right)$$

(1)

where $r_{wg}$ is defined “as the proportional reduction in error variance” (James et al., 1993, p. 307, emphasis original).

LeBreton, James, and Lindell (2005), following James et al. (1984, 1993), also provided an equation for calculating $r_{wg(J)}$, or agreement index for multi-item scales. This equation accounts for differences in $r_{wg}$ weights between scales of a few versus many items. Note that it is not an indication of internal index reliability (such as Cronbach’s alpha or Spearman–Brown formulas). This calculation for multi-item scales is expressed in the following formula:

$$r_{wg(J)} = ((j)\bar{r}_{wg}) / ((j)\bar{r}_{wg}) + (1 - \bar{r}_{wg})$$

(2)

Where $\bar{r}_{wg} = \text{the average } r_{wg} \text{ across items}$, and $J = \text{the number of items}$.

Like the values of the Pearson product–moment coefficient of correlation, the values of $r_{wg}$ will theoretically range from 0 to 1, with values greater than .70 gen-
erally considered the minimum accepted value for aggregation. Note that sometimes values will fall outside of the theoretical range of 0 to 1, and can even be negative, if the obtained variance exceeds the expected variance (see James et al., 1993, and Klein et al., 2000, for further explanation).

To demonstrate how \( \text{rwg} \) is utilized, consider two groups. Group A has an observed error (rater) variance of \( S_x^2 = 0.0001 \) (very high observed agreement) and a random expected variance of \( \sigma_E^2 = 0.03 \). Therefore, the \( \text{rwg} = 1 - (0.0001/0.03) = 0.99 \). This value is above the recommended 0.70, and therefore one can appropriately aggregate the data to the group level (i.e., use the group mean to represent the ratings of the group members). This agreement (\( S_x^2 \) departure from 0) is due to actual departure and not expected random error (\( \sigma_E^2 \)). Group B, however, has an observed (rater) error variance of \( S_x^2 = 0.02 \) (smaller observed agreement) and a random expected variance of \( \sigma_E^2 = 0.03 \). Therefore, the \( \text{rwg} = 1 - (0.02/0.03) = 0.34 \), and the data should not be aggregated. This agreement is due more to random error than it is to actual variance. Note again that the \( \text{rwg} \) is a measure of agreement within a group compared to an expected error term, not agreement compared to other groups.

To extend this example, suppose a researcher was interested in studying group level ratings of effectiveness, such that he or she could compare one group of cyclists' performance to another group. This researcher has data from 30 groups of cyclists; however, only 25 of the groups meet the .70 suggested cutoff for aggregation. For group-level analysis, (a) should the researcher only include the 25 groups above 0.70 or should he or she include all 30 groups, and (b) how should the researcher report the \( \text{rwg} \) values used to justify his or her decision? Klein and colleagues (2000) explained as follows:

If one is using \( \text{rwg} \) to test the appropriateness of aggregation, should one remove variables and/or units with \( \text{rwg} \) values that fall below 0.70? The answer is open to debate. Many researchers who use \( \text{rwg} \) simply report the average \( \text{rwg} \) for their groups (for each measure they wish to aggregate to the group level), using \( \text{rwg} \) values near or above .70 to justify aggregation. Reporting the range of \( \text{rwg} \) values and the percentage of units with values greater than .70 seems reasonable. (p. 517)

Klein et al. (2000) also added that the decision to include or exclude groups must be based on additional factors to complement the \( \text{rwg} \) scores used to interpret a single group on a single variable. For example, if one has a large number of groups and only a very few are below .70, those groups could probably be removed without great implications to the overall findings. Or, if one group has very low agreement on multiple variables, that group may need to be excluded. Conversely, if a large percentage of groups fall below the .70 mark, the data probably cannot be justifiably aggregated to the group level. It must also be noted that although Klein and colleagues endorsed reporting the overall average \( \text{rwg} \), many researchers absolutely
reject this practice as a basis for aggregation, as it can be misleading and promote poor inferences (see Castro, 2002, for a more complete discussion).

Therefore, when determining what to do with the aggregated group data using $r_{wg}$, the best practice seems to follow these four steps. First, understand the conceptual issues. For example, if agreement tends to be low, perhaps one is trying to aggregate a property that is better conceptualized at the individual level. Second, understand the data as a whole. For example, if some groups are high in agreement and some are low, one should probably attempt to understand factors related to why agreement is high or low, and not simply dismiss groups that are low. Third, report as much information as possible. If the number of groups is small, one could report the $r_{wg}$ values for each group. If not, the most informative practice is to report at least the range of $r_{wg}$ values, and the percentage of groups with $r_{wg}$ greater than the cutoff of .70. Fourth, utilize multiple techniques (e.g., ICC and WABA) and well-supported theory as additional evidence to support aggregation.

**Intraclass Correlation**

Intraclass correlations [ICC(1) and ICC(2)] are used to assess interrater reliability (consistency of responses among raters: Bliese, 2000; Castro, 2002). ICC(1) estimates the amount of variance in individual-level responses that can be attributed to group-level properties or the degree to which a measure varies between versus within groups. ICC(1) is not dependent on the number or size of groups (Castro, 2002). Though Raudenbush and Bryk (2001) defined ICC(1) as the proportion of total variance that can be explained by group membership, James (1982) viewed it as the reliability of a single assessment of the group mean, or the extent to which individuals are substitutable within a group. If one utilizes the James interpretation, it is clear why ICC(1) can be utilized as a criterion for aggregation.

ICC(1) can be considered a measure of reliability of a group mean based on a single assessment (or rating) of that mean. ICC(2), in contrast, is a measure of the reliability of the group mean based on all assessments within a group (Bartko, 1976; Bliese, 1998; James, 1982; McGraw & Wong, 1996; Shrout & Fleiss, 1979). James (1982) suggested that when ICC(1) is large, a single assessment (or a single individual) can provide a reliable estimate of a group mean, but when ICC(1) is small, multiple assessments are necessary and the use of ICC(2) is warranted.

To calculate either form of ICC, one utilizes a one-way independent fixed effects analysis of variance (ANOVA) where the variable of interest is the dependent variable and the group membership is the independent variable (Bliese, 2000). For example, if one were measuring whether membership in a particular physical education class impacted fitness ratings, individual fitness ratings would be the dependent variable, and group (school) membership would be the independent variable. ICC(1), following Bartko (1976), can be calculated by the following:
\[ ICC(1) = \frac{(MSB - MSW)}{[MSB + (k - 1)MSW]} \]  

(3)

where \( MSB \) is the between-group mean square, \( MSW \) is the within-group mean square, and \( k \) is the number of individuals in each group. Usually the average number of individuals can be utilized for \( k \) if group sizes differ, but Bliese (2000) explained a formula for calculating \( k \) if group sizes differ dramatically. ICC(2) is generally calculated by the following formula (Bliese, 2000):

\[ ICC(2) = \frac{(MSB - MSW)}{MSB} \]  

(4)

If group size is known, and especially if group sizes are large (Castro, 2002; James, 1982), ICC(2) can also be calculated using the Spearman–Brown formula as demonstrated by Shrout and Fleiss (1979):

\[ ICC(2) = \frac{k(ICC(1))}{1 + (k - 1) ICC(1)} \]  

(5)

Note that this use of the ICC assumes a one-way ANOVA, and considers \( k \) to be the number of individuals within a group who report a one-time assessment. This type of ICC differs somewhat from a repeated measures design used in many activity settings where \( k \) represents the number of trials for \( n \) number of individuals. In a repeated measures type of design, a two-way ANOVA is utilized so that interaction effects can be accounted for dependent on various assumptions concerning the criterion score. For a detailed explanation on a repeated measures application of ICC as a reliability measure, see Baumgartner, Jackson, Mahar, and Rowe (2003).

When utilizing ICC(1), if the \( F \) test for between groups from the ANOVA is significant, then aggregation of participants within each group is considered justified (Klein et al., 2000) because variance between the groups is not due to measurement error. As ICC(2) is a measure of the reliability of the group means on an aggregate variable, it can be interpreted in a similar fashion to other reliability measures; that is, “Common practice suggests that values of 0.70 and higher are acceptable, values between 0.50 and 0.70 are marginal, and values lower than 0.50 are poor” (Klein et al., 2000, p. 518). That is, usually only ICC(2) values of at least 0.70 are considered acceptable for aggregation. Baumgartner et al. (2003), in the context of physical performance tests and repeated measures ANOVA, provided more stringent guidelines for acceptable reliability. They suggested that .70 to .79 is below-average acceptable, .80 to .89 is average acceptable, and .90 to 1.0 is above-average acceptable. Additionally, one must note that small between-unit differences [evidenced by small ICC(1) values], can often lead to reliable mean group differences. ICC(1) and ICC(2) can and should often be used together to make an aggregation decision (Klein et al., 2000). Furthermore, Baumgartner et al. and Klein et al. (2000), in providing aggregation guidelines, commented that there are varying standards used to determine whether data aggregation is acceptable,
and one must also take into consideration the context, previous standards utilized, and the implications of the decision to aggregate.

Within and Between Analysis

WABA is utilized to examine single level, multiple level, multiple variable, and multiple relation issues (Dansereau et al., 1984; Dansereau & Yammarino, 2000). Though we provide a general overview of these issues, the complete rationale and scope of WABA is well beyond the confines of this tutorial. Readers are referred to Dansereau et al. and Dansereau and Yammarino for a more complete mathematical and conceptual explanation of the uses of the technique, especially those that extend beyond data aggregation.

For aggregation, WABA is utilized to determine whether individuals nested within groups should best be conceptualized as whole groups (known as the homogeneous or wholes condition), or as relative individuals who are complementary but not similar (known as the heterogeneous, parts, or frog-pond condition), or as unique individuals (known as the individual or equivocal condition; Dansereau et al., 1984; Firebaugh, 1980; Klein et al., 1994).

For example, consider an athletic department with 1 athletic director and 10 coaches. A wholes condition would suggest that the athletic director treated all the coaches (the whole group) as highly similar individuals, or that the whole group of coaches held highly similar views on a particular variable such as loyalty, or leader characteristics. In essence, they tended to behave as a whole unit. A parts condition would suggest that though the individuals work together, their attitudes and views reflect their relative position within the group; they demonstrate patterned responses and behave more like parts, than like a whole unit. For example, the athletic director may give different responsibilities to each coach according to his or her ability or tenure, thereby ranking the office. Therefore, the higher ranked persons would tend to report higher satisfaction with the leader (reporting as a whole or average would miss the patterning). In the equivocal condition, the athletic director views each coach as an individual, and the coaches also view themselves as individuals—not a group or an interdependent team. Therefore, they would all demonstrate independent attitudes and behaviors.

With one independent variable (WABA I), the researcher utilizes a one-way independent effects ANOVA to examine the variance between and within the groups. In this way, WABA is similar to ICC in that it takes into account both within and between group variance. It is different, however, from $r_{wg}$, which only investigates within-group agreement.

In general, the following relations are utilized to determine aggregation alternatives:

- **Wholes**: $BV > WV$
- **Equivocal**: $BV = WV > 0$
where BV is between-group variance and WV is within-group variance. Similar to the ICC example, if there is more variation between groups than within them, the groups are seen as wholes. If there is equal variation within and between groups, then they are viewed as independent individuals. If there is more variation within groups than between them, then the groups are viewed as frog ponds or parts. If there is almost no variation within or between groups, the null condition is warranted, as is assumed to indicate error within groups (Dansereau & Yammarino, 2000).

E and F ratios (garnered from a one-way ANOVA) are utilized to test the practical and statistical significance of the variance, respectively. The E ratio is determined by calculating \( \eta \) (eta) values for between and within as follows (see Dansereau et al., 1984, for a complete explanation of eta, its relation to eta squared, and the linear tensor mathematical model on which the variant paradigm is based):

\[
\eta_{bx} = \sqrt{\frac{SS_b}{SS_T}} \\
\eta_{wx} = \sqrt{\frac{SS_w}{SS_T}} \\
E = \frac{\eta_{bx}}{\eta_{wx}}
\]

The E ratio determines whether variance lies mostly between groups or mostly within groups. A larger E indicates more between-group variance, whereas a smaller E indicates more within-group variance. After the E value is calculated, the following decision rules are employed utilizing geometric properties called the 15º rule (Dansereau et al., 1984):

- **Wholes** \( \infty \geq E \geq 1.30 \)  
- **Equivocal** \( 1.30 > E > 0.77 \)  
- **Parts** \( 0.77 \geq E \geq 0 \)

The F test for groups from the one-way ANOVA indicates statistical significance. It can also be calculated with the following formula:

\[
F = E^2 \left( \frac{N-J}{J-1} \right)
\]
where \( N \) is the number of scores, and \( J \) is the number of cells. If the \( F \) from the ANOVA is statistically significant, the following decision criteria are then utilized to determine the aggregation decision. Note that the \( F \) must first show statistical significance, then a decision is made (e.g., an \( F \) value above 1 is not considered a wholes condition unless it is also statistically significant):

- **Wholes**: \( F > 1 \)
- **Equivocal**: \( F \approx 1 \)
- **Parts**: \( F < 1 \)

The \( F \) ratio must demonstrate statistical significance and the E ratio practical significance for a decision to be reached to utilize group means as the unit of analysis. Incorporation of both E and \( F \) statistics is useful, especially when large sample sizes could artificially inflate the significance of the \( F \) test, and in a true WABA analysis the E test takes precedence over the \( F \) test (F. Dansereau, personal communication, November 18, 2004; Dansereau et al., 1984). Note that when a parts condition is suggested, the inverse \((F^{\wedge})\) of the \( F \) statistic is utilized to assess the significance of the within component (see Dansereau et al., 1984, for a full explanation and appropriate statistical tables).

Aggregation decisions in WABA are similar to those in ICC in that they are based on the statistical probabilities of within-group versus between-group variance. If within-group variance is significantly larger than between-group (equivocal), the group should be conceptualized as individuals. If however, within-group variance is significantly smaller than between-group (wholes), the individuals can be aggregated to the group level. The added advantage of WABA is the statistical ability to also identify the parts condition. If the within component is significant, and a parts condition is warranted, usually the data are described according to its patterning (e.g., bimodal, ranked) and the groups are not aggregated

**Differences Among Techniques**

After first addressing conceptual and design issues, and ensuring alignment with choice of statistical technique, several critical differences exist between the techniques that also help the researcher decide which tool to use and what inferences to draw from the aggregation technique (Klein et al., 2000). First, as pointed out earlier, \( r_{wg} \) examines agreement, whereas ICC(1) and (2) are assessments of reliability. Second, \( r_{wg} \) examines variability within each group compared to just that group, whereas ICC and WABA examine the variability within and between each group as compared to the entire sample. Thus, \( r_{wg} \) allows the researcher to investigate each group separately, to examine patterns of agreement. ICC and WABA examine the relative agreement in the whole sample, which allows one a sense of the
distribution, patterning, and agreement not just within each group, but across the entire sample.

Furthermore, \( r_wg \) is designed to examine only within-group assessments of variability, whereas WABA and ICC utilize both within and between assessments. This difference is important because it could lead to different aggregation decisions. For example, if agreement within each group is high, but there is little variation across groups, a researcher using \( r_wg \) would conclude that aggregation is warranted; however, when within and between variance are both accounted for, a researcher using ICC or WABA may argue that aggregation is not warranted due to insufficient variation between groups. The overriding decision to aggregate lies in the study’s design and purpose, particularly with respect to whether only within or both within and between variance is important.

Conversely, because ICC and WABA decisions are based on relative between and within variance across the whole sample, aggregation decisions may overlook individual groups with low agreement. That is, in ICC and WABA, a decision to aggregate is based on the across sample within variance being significantly less than the across sample between variance. Within the entire sample, however, there may be individual groups with low within agreement, and aggregating the data overlooks the nuances of those groups. In many designs, merging a small subset of “low agreement” groups into the larger aggregate is not a problem; however, in other instances, researchers may not want to aggregate unless the entire sample and every group within that sample shows high consistency of response.

A third difference in the techniques is that \( r_wg \) and ICCs allow for only two decisions (wholes or individuals), whereas WABA allows for three conditions (wholes, individuals, or parts). The possibility for only two decisions is not necessarily a weakness of \( r_wg \) and ICC, but as demonstrated by the example shown later in this tutorial, the possibility of a third decision is a strength of using WABA that can sometimes prove quite useful in interpreting seemingly contradictory data.

Given the differences between the analyses, Klein et al. (2000) argued that several indexes should be utilized before making an aggregation decision unless the theoretical framework clearly specifies otherwise. The use of multiple indexes adds to the robustness of data interpretation. For example, if researchers are interested in training performance results from only one program, or if they want to investigate multiple training programs and ensure they are all performing at a certain level, they should utilize \( r_wg \), because it assesses agreement within each group. If they are also interested in whether these groups differ significantly from those in the entire sample, they should utilize ICC or WABA because with these two methods, the variability both within and between groups is assessed. If the \( r_wg \) and ICC and WABA all pointed to aggregation, the findings would be very robust.

Klein and colleagues (2000) also offered the following advice for dealing with conflicting outcomes: “When the indices lead to differing conclusions regarding the merits of aggregation, researchers may rely on theory, prior research, and/or
belief in the superior merits of one of the indices in deciding whether or not to aggregate their measure(s)” (p. 520). That is, there is no absolutely superior method to determine if aggregation is warranted; researchers are encouraged to first examine their conceptual model to assess whether error should be defined relative to each group ($r_{wg}$) or relative to the entire sample (ICC and WABA). Then, they should assess the strengths and weaknesses of the statistical tools to give them the desired level of confidence that their error is properly assigned. Then, data should be analyzed with the appropriate tool for the design, even if that tool is statistically less beneficial. Making conceptual and design decisions first helps not only in explaining results, but also avoids simply choosing an index that essentially makes the data say what one wants them to.

**EMPIRICAL EXAMPLE**

In this section, a sample data set is used to demonstrate aggregation decisions and conclusions drawn from each of the three statistical techniques. The data are drawn from a larger study regarding human resource management in athletic departments and its relation to individual, group, and organizational-level outcomes (Dixon, 2002). Individual-level variables (satisfaction and commitment) were garnered from the individual coaches because they represent the attitudes and behaviors of individuals (Kozlowski & Klein, 2000; Silverman, 2004; Silverman & Solman, 1998). Then, the three statistical techniques were employed to assess whether aggregation to the group level was warranted (to test if human resource management systems create group-level effects in addition to individual-level effects). The data set is instructive because it is not clearly straightforward in its interpretation.

A sample of coaches ($N = 681$) from 86 National Collegiate Athletics Association (NCAA) Division III athletic departments were asked to respond to a questionnaire containing items concerning their Satisfaction and Commitment. Informed consent was obtained from all participants. Satisfaction was measured with a five-item subset of the Brayfield and Rothe (1951) Job Satisfaction Index (a Likert-type scale with values ranging from 1–7). Individual Organizational Commitment was assessed through the eight items of Allen and Meyer’s (1990) Affective Commitment Scale (a Likert-type scale with values ranging from 1–7).

**Summary of Data**

Table 1 contains the means, range, and standard deviations for each of the 13 items and scale summations contained in the individual-level data (individual items from the scales are included as they are computed separately for $r_{wg}$). Groups ranged in size from 3 to 17, with an average membership of 7.9 coaches. (Note that 7.9 is
substituted for $k$ in ICC[1]). Though satisfaction and commitment were negatively skewed, the individual items showed no restriction in range.

Tables 2 and 3 display the analysis of variance scores for Satisfaction and Commitment subscale scores, respectively. A general overview of the analysis indicates that overall variance lies mainly within groups ($SS_w > SS_b$), but that when degrees of freedom are taken into account, variance lies mainly between groups ($MS_B > MS_V$). Table 4 contains the aggregation statistics used for ICC and WABA for each of the summed scales (Satisfaction, Commitment). These figures were calculated using equations 3 to 8, as outlined earlier in the article.

Concerning $r_{wg}$ aggregation, for sake of simplicity, rather than reporting all 86 $r_{wg}$ scores for all items in each scale, the following summary reports the $r_{wg}$ values

### TABLE 1
Range, Mean, and Standard Deviations for Individual-Level Satisfaction, Commitment, and Performance Variables and Scales

<table>
<thead>
<tr>
<th>Variable</th>
<th>Range</th>
<th>$M$</th>
<th>$SD$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satisfied with present job</td>
<td>1–7</td>
<td>5.88</td>
<td>1.16</td>
</tr>
<tr>
<td>Enthusiastic about work</td>
<td>1–7</td>
<td>6.16</td>
<td>0.88</td>
</tr>
<tr>
<td>Each day seems never-ending (reverse)</td>
<td>1–7</td>
<td>5.77</td>
<td>1.42</td>
</tr>
<tr>
<td>Consider job unpleasant (reverse)</td>
<td>1–7</td>
<td>6.41</td>
<td>0.93</td>
</tr>
<tr>
<td>Find enjoyment in work</td>
<td>1–7</td>
<td>6.25</td>
<td>0.88</td>
</tr>
<tr>
<td>Satisfaction scale summary ($\alpha = 0.78$)</td>
<td>1–7</td>
<td>6.09</td>
<td>0.79</td>
</tr>
<tr>
<td>Happy to spend career in department</td>
<td>1–7</td>
<td>5.28</td>
<td>1.64</td>
</tr>
<tr>
<td>Enjoy discussing department outside</td>
<td>1–7</td>
<td>5.3</td>
<td>1.47</td>
</tr>
<tr>
<td>Department’s problems are my own</td>
<td>1–7</td>
<td>4.28</td>
<td>1.64</td>
</tr>
<tr>
<td>Could easily attach to another (reverse)</td>
<td>1–7</td>
<td>3.83</td>
<td>1.74</td>
</tr>
<tr>
<td>Do not feel part of department (reverse)</td>
<td>1–7</td>
<td>5.74</td>
<td>1.47</td>
</tr>
<tr>
<td>Not emotionally attached (reverse)</td>
<td>1–7</td>
<td>5.28</td>
<td>1.66</td>
</tr>
<tr>
<td>Department has personal meaning</td>
<td>1–7</td>
<td>5.24</td>
<td>1.49</td>
</tr>
<tr>
<td>Do not feel belonging (reverse)</td>
<td>1–7</td>
<td>5.41</td>
<td>1.59</td>
</tr>
<tr>
<td>Commitment scale summary ($\alpha = 0.83$)</td>
<td>1–7</td>
<td>5.04</td>
<td>1.09</td>
</tr>
</tbody>
</table>

**Note.** $N = 681$. $\alpha =$ coefficient alpha, internal consistency of scales.

### TABLE 2
One-Way Independent Fixed Effects Analysis of Variance of Coaches’ Satisfaction by Group Membership

<table>
<thead>
<tr>
<th></th>
<th>$SS$</th>
<th>$df$</th>
<th>$MS$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>91.95</td>
<td>85</td>
<td>1.082</td>
<td>1.90*</td>
</tr>
<tr>
<td>Within groups</td>
<td>337.745</td>
<td>595</td>
<td>0.568</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>429.695</td>
<td>680</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.** $N = 681$. *$p < .001$.**
for the scales (using equation 2 to calculate), including their range, number of values outside the expected 0 to 1 range, and percentage of groups meeting the 0.70 cut-off. The $r_{wg}$ values for the satisfaction scale for all 86 groups ranged from –3.21 to 0.99, with 80 of 86, or 93% of the groups meeting the 0.70 cutoff. If one group is deleted from this analysis because its $r_{wg}$ value (–3.21) is out of range, we obtain a final range of 0.50 to 0.99, with 80 of 85 groups, or 94%, being above the 0.70 cutoff. The values of $r_{wg}$ for the commitment scale for the entire sample ranged from –18.10 to 1.55, with 6 groups being out of range and 71 of 86, or 83%, being above the 0.70 cutoff. After removing 6 groups because they were out of range or negative, the revised $r_{wg}$ values for commitment ranged from 0.02 to 0.97 with 69 of 80, or 86% being above the 0.70 cutoff.

### Aggregation Decisions

According to the data presented earlier and in Tables 2 to 4, we would reach different aggregation decisions depending on which technique was utilized. Regarding Commitment, $r_{wg}$ values for individual groups showed rather high agreement (83%
above the 0.70 cutoff). If one was interested in only the individual to group level, one would aggregate the 69 groups that are above the 0.70 cutoff. As noted in the \( r_{wg} \) section, however, what should be done with the other 11 groups that are either out of range or are below 0.70 (i.e., should they also be aggregated and utilized for subsequent group analysis, or should they be eliminated)? Following Klein et al. (2000), in this case, one might recommend keeping the groups because of the following reasons: (a) the same groups that are out of range or low in agreement on the Commitment scale are not the same ones with low agreement on the Satisfaction scale, (b) there are relatively few groups that are below the cutoff, and (c) the other indicators (i.e., ICC) also point to aggregation. ICC(1) would also suggest aggregation as the between-groups \( F \) statistic from the ANOVA is significant; however, the value of ICC(1) is quite low, suggesting the need for group evaluation using ICC(2) according to James (1982). The ICC(2) decision to aggregate is tenuous, as a value of 0.53 is generally interpreted as medium reliability (Klein et al., 2000). Finally, WABA, which has the most stringent requirements for aggregation, suggests an equivocal (or nonaggregation) condition. Though from an ANOVA perspective, the results would infer wholes because MSB > MSW, in WABA, the test of practical significance, or E test, is not statistically significant. The E ratio (derived from SS) falls between zero and 0.77, which points to a parts condition. However, the inverse \( F \) ratio, which tests the statistical significance of the within component, is not statistically significant; therefore, the data infer an equivocal condition and should not be aggregated (Dansereau et al., 1984; F. Dansereau, personal communication, November 18, 2004).

Satisfaction results are somewhat similar, as \( r_{wg} \) values support aggregation (values from 0.50 to 0.99 and 94% of groups meeting the 0.70 cutoff). ICC(1) would support aggregation, but ICC(2) would suggest nonaggregation as it falls below 0.50. WABA would suggest an equivocal condition for the same reasons as Commitment.

The \( r_{wg} \) values indicate that compared to each other, individuals within each department generally hold the same perceptions of Satisfaction and Commitment. Thus, one could probably reliably report a group mean for each of the 86 departments on these variables if one’s research design was to analyze them without reference to the entire sample. WABA and ICC(1) values, however, indicate that compared to the variance in the entire sample (coaches both within and outside of their departments), the individuals within each department generally do not hold the same perceptions of Satisfaction and Commitment, and reporting a group mean would not be a reliable estimate of these variables. The coaches, according to their Satisfaction and Commitment levels, should be viewed as individuals, not as groups. Following Ostroff and Bowen’s (2000) model, we would suggest that the human resource management system is not strong enough to create group-level effects, and it does not create a strong organizational climate toward Commitment.
These data demonstrate the value of a variety of statistical tools. In other words, each statistical tool carries different information. The conclusion from $r_{wg}$ suggests that the groups have very high agreement when only within-group variance is considered. In this case, the high agreement within groups could be due to a very high mean satisfaction for the entire sample (6.09 on a 7-point scale with a $SD$ of only 0.79). ICC(1) and ICC(2), taking into account the within and between variance, however, suggest that though the bulk of variance lies between groups, it would be tenuous to infer a group condition. In other words, though agreement may be high, reliability (or the ability to substitute one respondent’s response for another) is not. Yet the $F$ values for ICC are significant, and some would consider this sufficient for aggregation (Klein et al., 2000). WABA, however, helps further clarify the issue by introducing the test of practical significance. In essence, a WABA researcher would suggest that though the statistical significance of the ANOVA points to a wholes condition, the vast proportion of the variance lies within groups ($\eta_w$; Dansereau et al., 1984). Therefore, a wholes condition cannot be inferred. Though one could aggregate on the basis of the $r_{wg}$ or ICC results (if solely based on the $F$ test), the reliability of such aggregation would be potentially questionable. The most conservative approach for this data set would be to interpret the data at the individual level and proceed accordingly.

**CONCLUSION**

Though research in sport and physical activity has seen a rapid increase in multilevel designs, level of analysis issues, both conceptual and statistical, still exist (Silverman, 2004). This work contributed by investigating three commonly utilized aggregation techniques for inferring levels of analysis. Though we encourage researchers to investigate additional multilevel techniques, our review has provided a starting point for understanding the need for multilevel analysis, the conceptual basis for such theory, and the differences in different data analytic techniques. For many of the problems we need to address in sport and exercise, it is critical for us to not only correctly specify multilevel theories, but also follow appropriate methods in testing them.

**REFERENCES**


LeBreton, J. M., James, L. R., & Lindell, M. K. (2005). Recent issues regarding $r_{wg}$, $r^*_{wg}$, $r_{wg(j)}$ and $r^*_{wg(j)}$. *Organizational Research Methods, 8*, 123–138.


